非结构化数据处理技术及应用

陈 燕　李桃迎　张金松/编著

科学出版社

非结构化数据处理技术及应用

陈　燕　李桃迎　张金松　编著

本书由辽宁省交通厅项目（编号：201401）、国家自然科学基金项目（编号：71271034）、中央高校基本科研业务费（编号：3132016306）资助

科学出版社
北京

内 容 简 介

本书系统详细地阐述了非结构化数据的处理方法与技术。通过对非结构化数据特点的分析，从非结构化数据的基础知识和理论、开源工具及应用举例、数据预处理、预测模型研究、网页数据的采集、非关系型数据库存储、结构化大数据分析平台、电商个性化推荐系统的应用、网购评语情感挖掘、全文检索技术、基于主题的检索系统等不同角度给出了结构化与非结构化数据的分析、挖掘与应用内容。

本书可作为信息管理与信息系统、电子商务、计算机应用、软件工程等高年级本科相关专业的教科书；同时也可以作为管理科学与工程、计算机应用及软件工程、工业工程等相关学科研究生的教科书或参考资料。

图书在版编目（CIP）数据

非结构化数据处理技术及应用/陈燕，李桃迎，张金松编著. —北京：科学出版社，2017.12

ISBN 978-7-03-053188-9

Ⅰ. ①非… Ⅱ. ①陈… ②李… ③张… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2017）第 128176 号

责任编辑：李 莉 / 责任校对：贾伟娟

责任印制：吴兆东 / 封面设计：无极书装

科 学 出 版 社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京京华虎彩印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2017 年 12 月第 一 版 开本：720×1000 1/16

2017 年 12 月第一次印刷 印张：22

字数：450 000

定价：132.00 元

（如有印装质量问题，我社负责调换）

作 者 简 介

陈燕（Chen Yan），博士，大连海事大学交通运输管理学院教授、博士生导师，管理科学与工程学科一级学科带头人并为省重点学科负责人，担任辽宁省物流航运管理系统工程重点实验室主任、辽宁省创新团队负责人。曾撰写《数据挖掘技术与应用》、《数据仓库与数据挖掘》、《数据仓库技术及其应用》、《管理信息系统开发教程》、《信息经济学》及《信息系统集成技术与方法教程》等学术专著与教材 10 余部。主持并完成多项国家自然科学基金、国家科技计划项目及多项省部市级项目，获得省部级奖励 10 余项，发表相关学术论文 200 余篇。

前　言

在大数据时代，互联网中充斥着大量以文本、图形、图像、视频、声音等形式存在的非结构化数据，这些数据与传统结构化数据相比较具有鲜明的特征，因此，传统数据挖掘的方法已经难以对非结构化数据进行有效的处理，而需要在对其进行加工的基础上，才能完成相关分析与挖掘方法的研究。本书在此基础上，对非结构化数据的处理方法进行综述，并通过实验加以说明。

作者多年来通过各项科研项目的积累，在结构化数据的处理方法、数据挖掘方法等研究方面取得了丰硕的研究成果。在大数据背景下，非结构化数据的剧增，对数据分析提出了新的挑战。本书对非结构数据处理方法进行梳理，并加以实验佐证。

撰写本书的目的在于：为非结构化数据的处理方法提出有效的方案，对方案进行梳理，使读者能够对非结构化数据处理方法有一定的认识，通过实验佐证，帮助读者加深印象。

孙骏雄、李鹏辉、王任远、丁雯雯、韩红云、王琦、陈志珍、李欧、李盼、张鑫、徐慧颖、高鸽、李墨等同学参与并完成全书的校对工作。

本书旨在通过应用案例，从几个角度对非结构化数据处理的相关理论与技术进行说明。在编写过程中，作者查阅了国内外大量文献资料，谨向书中提到的和参考文献中列出的学者表示感谢。

同时，由于时间仓促和编者能力有限，书中难免存在一些不足之处，敬请广大读者批评指正。

作　者

2016 年 10 月

目　录

基础知识篇

结构化数据挖掘技术篇

非结构化数据挖掘方法及应用篇

基础知识篇

第 1 章　非结构化数据的基础知识

1.1　大数据的基本概念

1.1.1　大数据的定义及特点

“大数据”一词来自于英文“Big Data”，之前我们称之为海量数据。对于什么是大数据这个问题，迄今还没有一个权威的定义。大数据是一个抽象的概念，除了数据量庞大这一特征之外还具有其他的特征，研究学者、科技企业、数据分析师等由于各自的关注点和侧重点不同，分别从不同的角度给出了各自对大数据的定义和观点。通过以下定义，可以帮助我们更好地理解大数据在技术、经济和其他应用中的不同内涵（张引等，2013）。

最早提出大数据时代已经到来的麦肯锡咨询机构，对大数据给出的定义为，其是指大小超过常规数据库工具的具备获取、存储、管理和分析能力的“数据集”。该定义包括两方面的内涵：一是符合大数据标准的数据集的大小会随着时间的推移、技术的进步而增长（胡文静等，2015）；二是不同部门符合大数据标准的数据集的大小会存在差别。麦肯锡全球研究院（McKinsey Global Institute，MGI）报告指出，数据集的大小并不是评判大数据的唯一标准，数据规模的不断扩大和无法使用传统的数据管理工具满足数据处理需求也是大数据的特点。

IBM 则将大数据的特点总结为 4 个 V——数量（volume）、多样性（variety）、速度（velocity）和真实性（veracity）。IBM 认为，尽管前 3 个 V 涵盖了大数据本身的关键属性，但真实性是当前企业亟须考虑的重要维度（闫城榛和韩志国，2013），将促使企业利用数据融合（data fusion）和先进的数学方法进一步提升数据的质量，从而创造更高价值。

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（李鹏，2014）。从数据的类别上看，大数据是指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（吕本富和陈健，2014）。

关于大数据，另一个被各学术和应用领域广泛引用的定义是维基百科给出的，即大数据是指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（何海波，2014）。

此外，美国国家标准和技术研究院从学术的角度也对大数据做出了定义："大数据是指其数据量、采集速度或数据表示限制了使用传统关系型方法进行有效分析的能力，或需要使用重要的水平缩放技术来实现高效处理的数据。"

综上所述，大数据是指超出了传统方式分析和处理能力的数据集，很难适用于既有的数据库架构（黄永勤，2014），传统的软件工具难以进行数据的捕捉、存储、管理和分析，必须考虑新的处理模式和管理工具。同时，大数据的数据获取由传统的抽样转变为所有数据，对数据的分析更注重对数据的关联关系的探索和对事物的未来发展趋势的预测。

随着大数据的发展，大数据的概念也在不断得到充实和发展。大数据已经成为当今知识管理、商业智能领域最热门的话题之一。全球互联网巨头都已意识到了在"大数据"时代数据的重要意义，包括 EMC、惠普、IBM、微软、Oracle、SAP、Teradata 在内的全球 IT 巨头纷纷通过收购大数据相关厂商来实现技术整合，可见其对大数据的重视（赵金明，2013）。受益于"大数据"概念的行业业内人士表示"大数据"产业链条包含了数据生成、数据存储、数据处理和数据展示等一个完整的生态系统之中的多个环节。

完整的生态系统还应当包括大数据处理结果的应用，与大数据相关的公司有以下几类。

（1）与海量数据的存储和处理相关的公司。

（2）与数据中心建设与运营维护相关的公司。

（3）与视频化应用相关的公司。

（4）与智能化和人机交互概念相关的公司（表 1.1）。

表 1.1　大数据概念股一览表

股票代码	股票名称	投资亮点
002230	科大讯飞	提供语音技术
300229	拓尔思	提供语义技术
300036	超图软件	提供地理地图处理技术
300182	捷成股份	提供视频处理解决方案
600588	用友软件	具备企业管理领域大数据技术
300302	同有科技	提供大数据存储设备

大数据有四个典型的特征，具体如下。

1）数据容量大

EMC 公司 2014 年发布了最新的数据宇宙报告《充满机会的数字宇宙：丰富的数据和物联网不断增长的价值》，这是业界唯一的，量化并预测年度数据产生量的研究报告。报告显示，2013 年全球数据量为 4.4ZB，在接下来的十年，全球数据量仍将保持 40%的速度增长，每两年翻一番（马建堂，2015），2013~2020 年全球数据量将增长 9 倍，由 4.4ZB 增至 44ZB[①]。

社会的数据量已经由 TB、PB 级别跃升至 EB、ZB 级别。这是一个什么概念呢？先来回顾一下各数据衡量单位之间的换算关系：

1 kilobyte（KB）=10^3byte

1 megabyte（MB）=10^6byte

1 gigabyte（GB）=10^9byte

1 terabyte（TB）=10^{12}byte

1 petabyte（PB）=10^{15}byte

1 exabyte（EB）=10^{18}byte

1 zettabyte（ZB）=10^{21} byte

1 yottabyte（YB）=10^{24}byte

1 nonabyte（NB）=10^{27}byte

1 doggabyte（DB）=10^{30}byte

《充满机会的数字宇宙：丰富的数据和物联网不断增长的价值》中将其形容为，假设一个字节的数据是一加仑（1 加仑≈3.785 升）水的话，仅十秒就会有足够的数据填满一个普通房子。到 2020 年，这一过程将仅花费两秒时间。假设将 2013 年全球的数据用 iPad 来存储，它们叠加起来的长度会超过三分之二的地球到月球的距离（253 704 千米）。到 2020 年，全球的数据总量将填满 6.6 个地球到月球距离的堆栈。

2）数据类型多

按照数据结构，数据可以划分为结构化数据、半结构化数据和非结构化数据（王晓波，2014）。数据不仅仅单纯指人们在互联网上发布的信息（包括网络日志、社会数据、互联网文本和文件；互联网搜索索引；呼叫详细记录、天文学、大气科学、基因组学、生物和其他复杂或跨学科的科研、军事侦察、医疗记录、摄影档案馆视频档案、大规模的电子商务等信息），也包括全世界的工业设备、汽车、电表上无数的数码传感器随时测量和传递的有关位置、运动、震动、温度、湿度

① 数据来源：The Digital Universe of Opportunities：Rich Data and the Increasing Value of the Internet of Things. http://china.emc.com/leadership/digital-universe/2014iview/executive-summary.htm。

乃至空气中化学物质的变化等海量的数据信息。

结构化数据是我们传统的使用习惯上的数据形式，基本是表格式的数据。目前对结构化数据的处理技术已经相当成熟，一般用关系型数据库进行结构化数据的处理。

相对于便于存储的结构化数据，在企业和人们日常生活中接触到的半结构和非结构化数据越来越多，高清图像、视频、音频等多媒体文件都属于非结构化数据。在大数据环境下，非结构化大数据对存储、管理和处理这些复杂的多形态的数据对的能力提出了更高要求。Hadoop 的流行简化了非结构化数据的处理难度，对非结构化数据的处理将是大数据挖掘的重要方向。

半结构化数据是介于结构化数据和非结构化数据之间的一种数据，它是结构化的数据（黄远鸣，2014），但是结构变化很大，不能完全按照非结构化或者结构化数据的处理方式来进行分析处理。

3）商业价值高

对于大数据的价值，一方面是通过大数据挖掘，发现以往没有发现的新规律和新知识；另一方面是新的结果能够直接应用到相关的生产经营当中，产生直接的经济效益。通常价值密度的高低与数据总量的大小成反比（廖仕东，2015）。以遍布城市各地的监控视频为例，一部 1 小时的视频，在连续不间断的监控中，有用数据可能仅占一两秒。2013 年，数字宇宙中仅 22%的信息被视为有用数据，但实际上仅有不到 5%的有用数据得到了分析。到 2020 年，由于物联网带来的数据增长，所有数据中 35%的数据将被视为有用数据。如何通过强大的机器算法更迅速地完成对数据的价值“提纯”成为目前大数据背景下亟待解决的难题（朱建平和李秋雅，2014）。

4）处理速度快

所谓 1 秒定律指的是对处理速度的要求，一般要在秒级时间范围内给出分析结果，如果时间太长就失去了价值，因为客户的体验就在一秒之间。这是大数据区分于传统数据挖掘的最显著特征（阎巍和李俭，2015）。在面对蕴含巨大商业价值的海量数据时，处理数据的效率就是企业的生命。传统的数据处理方式已经无法满足如此海量的数据的高效处理需求，大数据时代对数据驾驭能力提出了新的挑战，也为人们获得更为深刻和全面的潜在价值提供了机遇（龚文峰，2014）。

1.1.2　大数据的发展背景及历程

1. 企业级应用

随着企业信息化应用的逐渐深入，信息处理系统随之产生了大量的数据。在企业的经营管理过程中，企业的内部业务企业资源计划（enterprise resource planning，ERP）系统、财务系统、办公自动化（office automation，OA）系统、

客户关系管理（customer relationship management，CRM）系统、供应链管理系统等产生了大量存储于数据库中的数据，同时也产生了众多文档、交易记录、操作日志、客户反馈等非结构化数据、传感器数据及图像视频监控文件等实时多媒体数据。一些企业已经意识到这些数据的潜在价值，并通过数据挖掘方法对客户的交易过程、业务处理流程等方面进行了分析和预测。但是企业所处的信息化环境正在发生着变化，企业应用和互联网、移动互联网的融合越来越快，来自企业外部的非结构化数据在大大增加。大数据技术开始向传统企业及组织的 IT 应用领域渗透，一些 IT 相关业务领先的企业和组织开始尝试大数据技术实验性部署。对于这些数据的分析和应用将促使企业的基础 IT 架构、数据处理、应用软件的开发和管理模式等领域产生新的变革。因此，国内一些硬件厂商也纷纷开始布局大数据，如联想通过与全球知名的存储公司 EMC 合作，正式进入大数据的企业级应用领域，随后国内的其他厂商也纷纷推出基于大数据的产品，如华为在统一存储领域中推出了面向企业级应用的四款 T 系列的 OceanStor 产品，提高了其在存储领域的地位。

2. 网络信息的海量增长

早期的网络主要是提供电子邮件和网页服务，而文本分析、数据挖掘和网页分析技术也相应地应用于挖掘电子邮件内容和构建搜索引擎等领域中，网络数据占据了一大部分的全球数据量。而今，Web 充满了各种不同类型的数据，如文本、图片、视频等，大量应用于半结构化或者非结构化的技术应运而生。随着社交媒体的发展，各类论坛、博客、社交网站为用户创建、上传和分享数据创造了更为便捷的方式，社交数据开始急速增加。Google 上每天需要处理超过 20PB 的数据。截至 2013 年 12 月新浪微博拥有 1.291 亿的月活跃用户，61 400 万的平均日活跃用户，每天发布和转发的信息超过 2 亿条（郁芹和钱姮，2013）。截至 2013 年，淘宝网拥有近 5 亿的注册用户数，每天有超过 6 000 万的固定访客，同时每天的在线商品数已经超过了 8 亿件，平均每分钟售出 4.8 万件商品，在 2014 年“双十一”，支付宝交易笔数高达 1.058 亿笔，通过无线设备支付的订单共近 900 万笔（李鹏程，2015）。

此外，网络应用产生的数据不仅仅来自于互联网，由于传统互联网到移动互联网的转变，移动宽带网速的迅速提升，产生数据的终端由 PC（personal computer）机转向了包括 PC、功能手机、智能手机在内的多样化终端，移动互联网也成为网络数据的重要来源（杨义恒，2012）。个人智能手机和平板电脑的快速普及，使越来越多的人、设备和传感器通过数字网络连接起来，而通过产生、分享和访问数据，移动互联网正在逐渐渗透到人们工作和生活的各个领域之中，移动终端逐渐演变成一个提供了通话管理、游戏娱乐、办公记录、网页浏览、购物理财、视频

分享等各类应用在内的运行环境。根据国际数据公司（International Data Corporation，IDC）最新报告（表 1.2），2014 年第三季度全球智能手机出货量超 3.2 亿部，在新兴市场需求量大增的推动下，2014 年第三季度智能手机出货量相较 2013 年同期增长了近 25.2%。

表 1.2　2014 年第三季度智能终端厂商出货量和市场份额

厂商	2014 年第三季度		2013 年第三季度		2014 年第三季度较 2013 年第三季度变化率/%
	出货量/百万部	市场份额/%	出货量/百万部	市场份额/%	
三星	78.1	23.8	85.0	32.5	-8.1
苹果	39.3	12.0	33.8	12.9	16.3
小米	17.3	5.3	5.6	2.1	208.9
联想	16.9	5.2	12.3	4.7	37.4
LG	16.8	5.1	12.0	4.6	40.0
其他	159.2	48.6	113.0	43.2	40.9
合计	327.6	100.0	261.7	100.0	25.2

3. 云计算的出现

云计算是 IT 领域继 PC、互联网之后的第三次革新浪潮。2006 年，Google 推出“Google101 计划”，首次正式提出了云计算的概念。短短数年间，云计算给信息领域带来了巨大的变革。目前，各国纷纷制订了云计算发展的国家计划，近年在国内也掀起了兴建云计算基地的热潮；国内外的知名 IT 企业也竞相推出云计算的产品和系统；学术界也对云计算技术积极开展深入研究（蔡京等，2012）。云计算是一种基于互联网的相关服务的增加、使用和交付的模式，通常涉及通过互联网来提供动态易扩展且经常是虚拟化的资源（崔嘉伟和孟波，2012）。在云计算出现之前，我们的数据大多保存在个人计算机和远程服务器当中。面对海量的 Web 数据，云平台将所有的数据集中存储到云端，用户通过浏览器或者专用的应用程序访问云端数据。云计算作为一种新型计算模式，体现了网格计算、分布计算、并行计算、效用计算等技术的融合与发展。随着以云计算为代表的新型信息技术在国民经济、国家安全、科学研究、社会民生等各个领域的不断深化应用，社会生活模式、工作模式和商业模式也在发生着重大转变，以云计算为代表的信息产业通过其基础设施即服务（infrastructure-as-a-service，IaaS）、平台即服务（platform-as-a-service，PaaS）和软件即服务（software-as-a-service，SaaS）等服务模式正带动着众多产业形态的创新和改革（蔡丽霞，2015）。目前，Google 云计算已经拥有了 100 多万台服务器，Amazon、IBM、微软、Yahoo 等的“云”

均拥有几十万台服务器。企业私有云一般拥有数百上千台服务器。现在个人用户也可以将文档、照片、视频等文件上传到“云”中永久保存。

从技术上看，大数据与云计算密不可分。大数据的特点在于对海量数据的挖掘，其必然无法用单台的计算机进行处理，需要依托云计算的分布式处理、分布式数据库、云存储和虚拟化技术（郝志刚，2014）。云计算是大数据的 IT 基础和平台，大数据是云计算的重要应用，两者相辅相成，缺一不可。

4. 物联网的应用

物联网是互联网的延伸和扩展，通过局部网络或者互联网等通信技术将红外感应、全球定位系统、激光扫描器、射频识别（radio frequency identification，RFID）、蓝牙等信息传感设备进行信息交换和通信，实现对物体的智能化识别、定位、跟踪和监控管理，形成人与物、物与物、人与人之间的互联，从而实现信息化、远程管理控制和智能化的网络，它包括互联网以及互联网上的所有资源，物联网的用户端延伸、扩展到物与物，机器与机器之间进行信息交换和通信，因此是新一代信息技术的重要组成部分（李洋，2011）。

目前，物联网在智能工业、智能农业、智能交通、智能电网、安全监控等行业都有了一定的应用，巨大的网络连接使得网络上的流通数据大幅度增加。根据 IDC 公布的数据，在 2005 年机器对机器产生的数据就占据了全世界数据总量的 11%，预计到 2020 年，这一数值将增加到 42%，届时物联网将会产生更多的数据（杜兵，2012）。

物联网背景下的大数据成倍增长主要体现如下：联网终端数量逐渐增多；应用数据需求不断增长；数据采集数量的大规模增加；感知层的多样化数据需求。目前物联网领域采用的感知技术主要包括传感器、RFID、红外技术、蓝牙技术等短距离传输技术，广义上还包括音频、视频采集技术、文字与数据采集技术等大量相关的采集技术。传感网技术等的使用是物联网最为重要的技术类型，这些技术为工业、农业、医疗、交通等各行业提供了更为广泛的数据来源支撑，为物联网应用的数据采集提供了丰富的数据源泉。

早在 1980 年，著名未来学家阿尔文·托夫勒在其代表作《第三次浪潮》中将大数据誉为“第三次浪潮的华彩乐章”，准确预言了信息化时代的到来（刘晓英，2015）。

2001 年 META 集团分析师 Doug Laney 发布研究报告“3D Data Management：Controlling Data Volume，Velocity，and Variety”，首次提出“3V”模型，现在仍然是大数据的定义和特征的重要组成部分。

继知名科学杂志 *Nature* 在 2008 年推出大数据专刊，围绕大数据所面临的技术挑战进行了讨论之后，*Science* 也于 2011 年推出大数据专刊，从互联网技术、

超级计算、环境科学、生物医药等方面讨论了科学研究中的大数据处理问题（王元卓等，2013）。

2011 年 5 月，EMC 公司在拉斯维加斯举行以“云计算相遇大数据”为主题的 EMC World 会议，抛出了大数据的概念。同年，美国知名咨询公司麦肯锡在其报告“Big data：the next frontier for innovation，competition，and productivity”中对大数据做出了明确的定义，并指出“数据已经渗透到每一个行业和业务职能领域，逐渐成为重要的生产因素；而人们对于海量数据的运用将预示着新一波生产率增长和消费者盈余浪潮的到来”（张兴军，2014）。因此 2011 年被很多人认为是大数据元年。

2012 年 3 月，美国奥巴马政府宣布了“大数据研发计划”，并设立了 2 亿美元的启动资金，希望增强海量数据收集、分析萃取能力，根据该计划，美国国家科学基金会（National Science Foundation，NSF）、国立卫生研究院（National Institute of Health，NIH）、国防部（Department of Defense，DOD）、能源部（Department of Energy，DOE）、国防高级研究计划局（Defense Advanced Research Projects Agency，DARPA）、地质勘探局（United States Geological Survey，USGS）6 个联邦部门和机构将共同提高收集、储存、保留、管理、分析和共享海量数据所需的核心技术，扩大大数据技术开发和应用所需人才的供给。该计划还强调，大数据技术事关美国国家安全、科学和研究的步伐，将引发教育和学习的变革（段黎萍，2013）。欧盟方面也有类似的举措。过去几年欧盟已对科学数据基础设施投资 1 亿多欧元，并将数据信息化基础设施作为“Horizon 2020”计划的优先领域之一。联合国推出的“全球脉动”项目，希望利用大数据预测某些地区的失业率或疾病爆发等现象，以提前指导援助项目。2012 年 5 月，联合国发布了一份大数据白皮书，提出大数据对于联合国和各国来说是一个历史性的机遇，并讨论了如何利用大数据更好的服务公民（孙强和张雪峰，2014）；12 月，国际达沃斯论坛发布了“Big Data, Big Impact：New Possibilities for International Development”的报告，从金融服务、教育、健康、农业等多个方面分析了大数据所能带来的新的机遇。

国内也充分认识到了大数据的重要性，并高度重视大数据的开发利用工作。2011 年工业和信息化部发布了《物联网“十二五”发展规划》，将关键技术创新列为重点工程之一，其中的信息处理技术包含了海量数据存储、数据挖掘及图像视频智能分析，皆为大数据的重要组成部分（许爱装，2013）。从 2011 年底到 2012 年初，中国资本市场连续发布了三篇大数据主题研究报告《大数据时代即将到来》、《大数据时代三大发展趋势和投资方向》及《以数据资产为核心的商业模式》，系统阐述了大数据的商业图景，掀起了资本市场大数据的投资热潮。为了推动大数据产业的良性发展，2012 年 10 月，中国计算机学会组织来自大学、科研单位、企业和政府部门的 110 位专家成立了大数据专家委员会，其宗旨是探讨大数据的