大数据技术与应用专业规划教材
教育部-阿里云产学合作专业综合改革项目规划教材

互联网大数据处理技术与应用

◎ 曾剑平　编著

清華大學出版社

大数据技术与应用专业规划教材
教育部-阿里云产学合作专业综合改革项目规划教材

互联网大数据处理技术与应用

◎ 曾剑平　编著

清华大学出版社
北京

内 容 简 介

本书是教育部阿里云产学合作项目规划教材，书中全面介绍互联网大数据处理的主要理论和技术。全书共分为四大部分：概述、互联网大数据的获取、互联网大数据的结构化处理与分析技术、综合应用。在第1部分“概述”中首先对信息时代的技术变迁进行了回顾和归纳，指出了人类进入大数据时代的必然性及其基本特征，然后分析了互联网大数据的特点，接着对互联网大数据的相关技术进行了归纳和分析，最后指出了互联网大数据技术的发展。第2部分是“互联网大数据的获取”，包括原始数据的获取和数据提取技术，对网络爬虫的内核技术、主题爬虫技术、动态Web页面获取技术、微博信息内容获取技术、DeepWeb数据获取技术、反爬虫技术、反反爬虫技术、Web页面内容提取技术进行介绍。在第3部分“互联网大数据的结构化处理与分析技术”中，全面介绍结构化处理技术、大数据语义分析技术、大数据分析的模型与算法、大数据隐私保护、大数据技术平台，内容涵盖了互联网大数据处理与分析的主要方面。第4部分是关于互联网大数据技术的综合应用，以个性化新闻推荐为应用背景，运用阿里云大数据技术平台将本书介绍的一些关键技术、模型和平台贯穿在一起。

本书可作为高等院校计算机、信息、软件、大数据等相关专业研究生和高年级本科生的教材，也可作为计算机、信息、软件、大数据等领域研究人员和专业技术人员的参考书。

图书在版编目(CIP)数据

互联网大数据处理技术与应用/曾剑平编著. —北京：清华大学出版社，2017
(大数据技术与应用专业规划教材)
ISBN 978-7-302-46371-9

Ⅰ. ①互… Ⅱ. ①曾… Ⅲ. ①数据处理 Ⅳ. ①TP274

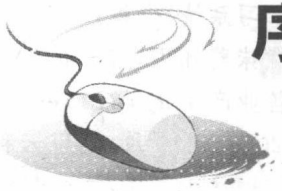
中国版本图书馆CIP数据核字(2017)第021601号

责任编辑：黄　芝　王冰飞
封面设计：刘　键
责任校对：梁　毅
责任印制：沈　露

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课 件 下 载：http://www.tup.com.cn，010-62795954

印 装 者：北京国马印刷厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：19　　**字　　数**：464千字
版　　次：2017年4月第1版　　**印　　次**：2017年4月第1次印刷
印　　数：1～2000
定　　价：49.00元

产品编号：072183-01



序言

DT 时代的数据思维与智能思维

本套云计算大数据丛书出版正值信息科技领域进入新一轮巨变，中国经济面临转型机遇的特殊时期。全球信息科技行业伴随着云计算、大数据、物联网、人工智能的发展即将进入一个泛智能的时代，云计算成为数字经济的基础设施；数据驱动、泛在智能成为各行各业转型升级的基础，不仅传统的 IT 从业人员面临能力升级，大多数在校大学生也面临新一轮知识体系的更新，各个垂直行业面临新一轮的人才升级。新一代人才教育与培训，需要一套产学一体的培训课程体系，这是阿里云愿意投身云计算大数据网络安全人才培养体系的时代背景。云计算、大数据、网络安全不仅关乎网络强国的大使命，也逐步成为各行各业专业人才的“元学科”，会逐步成为高等与职业教育的通识课程，一些发达国家已经在中小学立法普及编程课，已经开始指向这个趋势。“懂云计算，有数据思维，理解智能化”，未来可能是每一个工程技术人员与专业人士的必要素质。

2016 年开始，全球信息科技进入一个新的加速爆发周期，可能发生的大概率事件是：二十年之内，有一半的人类知识工作者会被人工智能替代，有服务能力的机器人会诞生，全世界的产业工人会少于机器人；虚拟现实和增强现实会替代今天的智能手机，变成一个新的入口；各行各业都会需要基于物联网的智能化，“中国制造”会成为广泛意义的“中国智造”。

新一轮科技带来了生活方式的变革、生产方式的变革，还有学习方式的变革，这几个趋势的背后，是云计算作为一种普惠科技的基础设施，大数据成为新能源，智能化成为一种新常识。

2016 年，全世界的短视频总量增长了 6 倍，直播业务在中国增长了 10 倍，远在偏远小镇的青年可以通过直播做电子商务，转化率可以提升十倍以上。当一个技术的使用成本趋近于零的时候，会带来广泛的社会效应。十年以前的直播只有电视台能做，需要专门的摄像机等设备，而今天的直播只需要一个手机，而且是多对多带互动的。无论是短视频，还是直播，背后都有云计算作为普惠科技的支撑作用，由此带来的，所有与知识传播有关的教育，包括整个内容行业，都会被它改变，随着大数据和人工智能的加入，人类学习的方式交互性会

更强，“学习系统”会根据不同人的理解程度做个性化的推荐与辅导。

这意味着知识生产与知识传播方式的根本性转变，这个恰恰是云计算、人工智能等科技与各行各业产生化学反应的交叉点，数据是这个转变的新能源。

在2016年10月，阿里云和法院系统合作，发布了一个面向法律服务的智能应用“法小淘”，通过把数千万份法律判例文本化，“法小淘”智能应用可以为普通老百姓以及初级律师提供“打官司”的咨询服务，根据用户输入的案件信息给出建议，包括推荐合适的律师。貌似与科技远离的法律服务也用上了人工智能，这是垂直行业泛智能化的一个小例子。

中国制造进入智能时代

在工业界，阿里云跟中石化合作，协助他们做了企业的电商平台；与徐工合作，推动工厂基于工业云的智能化；与上汽合作，推出具有智能服务的互联网汽车，都收到积极的市场反馈。中国制造，面临智能化的产业机遇，借助互联网人口和产业布局两大优势成为未来的第一个智能产品制造国。

在接下来的几年，互联网+智能制造的叠加会在很多个垂直领域出现，数据智能与制造业结合，产生“跨界重混”的效果，甚至制造业就不是以制造为主，而是以服务化为主。这个巨大的重构背后依赖云和大数据。也因为这个需求，我们可预见工业企业对云计算大数据人才的需求会越来越强烈。

“创业化生存”与共享经济的兴起

创业化，会成为一种常态，越来越多的年轻人开始告别公司，兴起中的数字经济体都是基于云平台的网络化协作组织；云计算成为共享经济的超级容器，催生新一代创业者和“斜杠青年”。十年以后，或许一半以上的从业者都是“斜杠青年”，今天美国就有数千万人是跨工作、跨公司的“斜杠青年”。

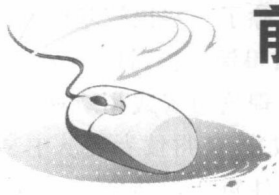
过去十年，云计算使得创业公司的创业门槛降低了10倍，没有云计算，Airbnb、NetFlix、推特、Uber等公司不可能这么快成长壮大，新一代创业者的一个核心能力就是要懂技术，理解数据和算法的价值，缺少技术理解力的创业者将面临更大的同质化压力。一句话，无论是草根创业，还是做一个“斜杠青年”，必要的数据思维是生存本能。

创业化和共享经济的崛起，有赖于云计算作为基础设施，大数据作为新能源的全新范式，新一代创业公司需要大量的科技人才。

在未来的经济环境里，普惠云科技的基础设施化、制造的智能化、软件的泛化以及数据无处不在，是一个大趋势，并且不断向各行各业渗透。本套丛书就是希望在这个普惠科技与各行各业深度融合的时代为下一代科技人才的培养提供更多产业界的经验与实践。

感谢清华大学出版社出版本套云计算与大数据方面的系列教材。感谢各位高校老师的辛苦努力和用心付出，使得本系列教材能够付梓出版。

——阿里云业务总经理　刘松



前言

互联网技术及应用进入一个高速发展时期，那些随手可得的互联网应用深刻地影响着社会经济的发展，改变了人们衣食住行、吃喝玩乐的生活方式，人们对互联网的依赖度逐年提升。网络数字化生活形态的形成，促进了互联网数据的累积，大数据由此成为互联网技术应用的新鲜血液，并将成为今后很长一段时期内各方关注的焦点。互联网大数据处理的理论、技术及其应用与社会经济各个领域的融合越来越密切，相关领域的专业技术人员迫切需要建立完整的互联网大数据分析应用的知识体系，以适应今后发展趋势的要求。

本书作者及其科研团队近十年来一直从事互联网内容分析挖掘、网络舆情、大数据、信息内容安全技术和应用方面的科研工作。在包括国家自然科学基金项目在内的各类科研项目支持下，对互联网信息获取和提取方法、互联网信息内容结构化处理技术、语义分析技术、数据挖掘的模型与算法、社交媒体中的用户行为及互联网金融等应用领域开展了大量研究，积累了一定的经验，强烈希望把科研工作中的体会和理解整理出来。此外，作者从 2011 年开始先后为复旦大学信息安全专业的本科生、研究生开设了《信息内容安全》《大数据安全》等课程，经过多年的教学实践，了解了学生的学习需求，积累了较为充足的讲义和素材。2016 年 5 月，教育部联合阿里云计算有限公司等单位发起了产学合作专业综合改革项目，确定了包括大数据在内的多个新技术方向的教材编写目标，以产学结合来推动高校教材和课程的改革。本书的编写正是在该综合改革项目的支持和推动下进行的，是第一本系统讲述互联网大数据处理技术及应用的教材和专业参考书。

本书在知识结构上，试图覆盖互联网大数据处理与应用的完整知识体系；在内容上，尽量做到深入浅出，既考虑知识的基础性，也兼顾技术发展方向和前沿。本书全面介绍互联网大数据处理与应用中的主要理论和技术，分为概述、互联网大数据的获取、大数据的结构化处理与分析技术和综合应用四大部分，涉及互联网大数据处理技术的各个方面，侧重于基本原理和实践技术的介绍，特别是较为系统全面地介绍互联网大数据获取、分析挖掘的各种技术，并融合了阿里云计算大数据平台的一些先进思想和业界的实践经验。

本书作为一本产学兼顾的教材，具有如下特色。

(1) 针对互联网大数据，从大数据的获取到可视化展示与发布的整个过程，帮助学生建立完整的知识体系。侧重于非结构化数据处理与分析，由于传统的结构化数据分析技术相对比较成熟，因此这种安排将有助于读者接触到更多的大数据核心关键技术。

(2) 除了一些比较基础性的知识外，在各个章节还融入了作者在教学和科研中所积累的一些值得深入探讨的问题和观点，具有一定的启发性。

(3) 理论与实践相结合，各个章节既包含技术原理介绍，也包含实现技术、开源架构等方面的叙述，使得读者能从中掌握技术应用及实现方法。

(4) 注重产学结合，基于阿里云及其大数据平台，构建了综合应用实例，有效地集成运用了本书的一些关键技术，帮助读者深入理解大数据处理技术。

全书由曾剑平负责内容安排、统稿，由互联网大数据处理技术和应用研究领域的一线人员参与编写。书中各章的编写人员安排：第 1 章由曾剑平、段江娇编写，第 2 章由曾剑平、段江娇、胡源编写，第 3 章由曾剑平、胡源编写，第 4 章由曾剑平、张硕编写，第 5 章由曾剑平、段江娇、毛天昊编写，第 6 章由曾剑平、张硕、段江娇、毛天昊编写，第 7 章由张泽文、吴爽、曾剑平编写，第 8 章由曾剑平、王欣编写，第 9 章由曾剑平、黄智行编写。另外，黄智行对第 5 章的 CRF 应用实例的部分程序及第 9 章的个性化新闻推荐系统进行了实现。本书在编写过程中，得到了阿里云计算有限公司的李妹芳女士的大力支持，在产学合作教材编写项目申请、立项、跟踪、结题、应用案例构建，以及相关的文字表达方面给予了很多帮助和指导。阿里云计算有限公司的宁尚兵先生在阿里云平台和大数据平台的使用、开发方面也给了大力的支持和帮助，阿里云计算有限公司的多位技术专家对本书的结构和知识安排提出了有益的建议。清华大学出版社的编辑们为本书的出版和编辑花费了很多心思。复旦大学计算机科学技术学院汪卫教授、中国科学院计算技术研究所靳小龙副研究员对本书进行了审阅，提出了宝贵的意见。此外，在本书的编写过程中，参考和引用了许多作者发表的各种论文、技术报告，我们均已在参考文献中列出。在此，一并表示衷心的感谢。

由于互联网大数据处理与应用技术所涉及的内容广泛，许多技术仍在不断发展中，所以本书在内容选择及编写上从深度和广度做了精心的安排。尽管编写组成员最近 5 个月来全身心投入，对每个技术要点尽量清楚地描述，但由于时间仓促及作者的学识水平限制，书中难免存在不足之处和疏忽，恳请读者不吝批评指正，以利于再版修订完善。

作　者

2017 年 1 月

目录

第1部分 概 述

第2部分 互联网大数据的获取

第 3 部分 互联网大数据的结构化处理与分析技术

第4部分 综合应用

第1部分

概　述

第1章 互联网大数据

本章对互联网大数据及其相关技术进行了概述，总结了信息技术发展过程和规律，指出人类社会进入数据时代，发展数据技术的必然性；从大数据的特点、思维、产业等方面做了一些探讨和分析，归纳并描述了互联网大数据的特点；着重对互联网大数据处理的技术体系进行了阐述，解释了其中的相关技术，包括数据采集、结构化处理、模型与算法、平台技术等；对互联网大数据技术的发展进行了展望，最后给出了本书的章节安排说明。

1.1 从 IT 走向 DT

自从信息技术开始运用于解决各种日常事务处理问题以来，各种软硬件设备和网络建设投资规模不断增加，各行业构建了具备一定规模的信息化系统，有力地促进了企事业单位生产经营和管理革新。在这个过程中，互联网的出现为信息化注入新的思维，极大地拓展了原有信息系统的时空维度。从 Web 1.0 到 Web 2.0，互联网应用中所形成的新思维更是导致这种时空维度发生了质的变化，互联网技术与各行业的生产经营和管理过程进行了深度融合，并成为当今时代最为活跃的资本市场热点。信息化、Web 1.0、Web 2.0 的演化进程中累积了大量数据，在信息化建设进入相对稳定状态时，人们开始意识到这些数据的价值，期待从数据中发现一些人工分析所无法得到的知识，从而推动经营管理进入新的阶段。

1.1.1 信息化与 Web 时代

在 20 世纪八九十年代信息化建设初期，人们面对的是复杂而又重复的业务和管理流程，迫切需要运用信息技术手段来将人们从这种重复劳动中解放出来，简化操作流程，提高工作效率。因此，在这个阶段，一些复杂且重复的工作任务成为信息化的首选目标。企业资源规划(ERP)、企业生产过程管理、人力资源管理、财务管理等是这个阶段的典型代表。为此，人们构建了支撑这类信息化系统运行所必需的软硬件设备、企业内部互联网络(Intranet)，开发了相应的应用软件，有力地保障了信息系统的运行。随着信息化分工的进一步细化，大量的基础网络技术、基础软硬件系统成为一个独立的行业逐步得到发展，各种操作系统、数据库系统、主机设备、行业软件系统等成为这个阶段最主要的产品，而它们反过来也使得人们在构建信息技术平台时不需要从头开始，而可以实现一种组装式的构建方法，

因此极大地加快了信息化进程。

在20世纪90年代前后，中国开始小规模接入互联网(Internet)。在初始阶段，互联网应用以浏览型业务为主，即所谓的Web 1.0时代。众多的互联网内容提供商(ICP)构建了专门的网站，为网民提供信息服务。典型的就是各类新闻网站、信息聚合网站和企事业单位的门户网站，它们提供了国内外新闻信息、企事业单位介绍、活动安排、政策规定等信息发布场所。众多的互联网服务提供商(ISP)为网民提供各种接入互联网的途径，如PSTN拨号、专线、分组技术等，但是由于接入技术的限制，带宽一般都不高，大都只能达到K级，因此，ICP提供的信息内容形式上比较单一，相对于图片、视频等多媒体内容而言，文本内容的占比要高得多。

随着用户端接入技术的发展和带宽的提高，人们不满足于简单的信息浏览，而是对信息交互提出了越来越高的需求。用户体验、用户为中心的思维得到了ICP的关注，由此，从Web 1.0转而进入Web 2.0时代。作为一个概念，Web 2.0是2004年始于出版社经营者O'Reilly和MediaLive International之间的一场头脑风暴论坛。相比于Web 1.0，Web 2.0则更注重用户的交互作用，用户既是网站内容的浏览者，也是网站内容的创造者。一个直观的例子就是从Web 1.0的个人网站到Web 2.0的博客进化。个人网站提供了关于个人信息的介绍和发布(Homepage)，但个人并无法在网站上与浏览者进行交流沟通，也无法得到浏览者的反馈信息。博客虽然也是个人信息发布的场合，但是个人不但可以灵活地在网站上修改和发布各种信息，而且浏览者可以进行评论，并与博客主人进行交互。因此，作为Web 2.0应用，它允许用户参与网站内容生成，实现用户与网站、用户与用户之间的交互。典型的Web 2.0应用有网络论坛、公告板(BBS)、博客(Blog)、信息聚合(RSS)、百科全书(Wiki)、社会网络(SNS)、对等网络(P2P)、即时信息(IM)等。然而Web 2.0初期的这些应用看起来只是为了满足人们日益增长的娱乐和休闲需求，与企事业信息化系统之间的联系并不很紧密。因此，提供服务的网站除了广告收入以外，就难以有其他的收入来源，网站自身的生存也成了最主要的问题。

寻找新的商业模式关乎到生存发展，就成为互联网应用提供商急切需要解决的问题，眼光自然而然就投向了传统的生产经营等领域。将传统的零售过程搬到互联网上则是第一道突破，电子商务由此成为一种比较早期的被挖掘出来的新应用。1997年年底在加拿大温哥华举行的第五次亚太经合组织非正式首脑会议(APEC)上美国总统克林顿提出敦促各国共同促进电子商务发展的议案，其引起了各国首脑的关注。在国内，各个行业所涉及的销售等商业过程都成为电子商务的现实样例，1999年3月8848等B2C网站正式开通，标志着网上购物进入实际应用阶段。此后，B2B、B2C、C2C、B2M、O2O等各种形式的电子商务层出不穷，成为互联网资本市场追踪的目标。

Web 2.0时代所形成的共享、开放、免费、去中心化等优秀的互联网思想，继续不断地指引着各个传统行业依靠互联网创造新的商业模式。最近几年来，出现了各种各样的网上服务，如网上交易、网上订餐、网上约车、网上租房、网上婚恋等。在这种背景下，传统行业找到了与互联网结合的有效模式，已有的信息化成果也就与互联网有了越来越密切的联系，即所谓的线下线上融合。至此，信息化与Web 1.0、Web 2.0的进化终于走到了同一条战线上，信息化进入了一个崭新的时代。