



“十三五”江苏省高等学校重点教材

大数据管理与应用专业系列教材

大数据 管理与应用导论

曹 杰 李树青 主编

非外借



科学出版社

编号: 2017-2-024



“十三五”江苏省高等学校重点教材
大数据管理与应用专业系列教材

大数据管理与应用导论

主 编: 曹 杰 李树青

副主编: 蒋伟伟 陈俊鹏 刘光徽 苏 震

郑怀丽 周挽澜



科 学 出 版 社

北 京



编号：2017-2-024

内 容 简 介

大数据管理与应用主要以信息科学、计算机科学和管理科学等学科为理论基础，其研究内容包括大数据科学基础理论、大数据预处理、大数据计算、大数据管理和分析等。本书力图通过对大数据科学相关数据管理方面内容的综合介绍，面向大数据时代的电子商务智能数据计算领域，从大数据采集、大数据预处理、大数据存储与计算、多源异构大数据分析、大数据知识融合技术和大数据的应用管理等方面说明大数据管理与应用的主要研究内容 and 应用方向。

本书可以作为高等学校大数据管理与应用、计算机应用、信息管理等专业的本科生和研究生教材，还可以作为 MBA、软件工程硕士等专业的教学用书，也可以供相关技术人员、管理人员参考。

图书在版编目（CIP）数据

大数据管理与应用导论 / 曹杰，李树青主编. —北京：科学出版社，2018.11

大数据管理与应用专业系列教材

ISBN 978-7-03-058688-9

I. ①大… II. ①曹… ②李… III. ①数据管理-高等学校-教材
IV. ①TP274

中国版本图书馆 CIP 数据核字（2018）第 202523 号

责任编辑：郝 静 / 责任校对：郑金红

责任印制：吴兆东 / 封面设计：蓝正设计

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>

北京虎彩文化传播有限公司印刷

科学出版社发行 各地新华书店经销

*

2018 年 11 月第 一 版 开本：787×1092 1/16

2018 年 11 月第一次印刷 印张：10 3/4

字数：254 000

定价：49.00 元

（如有印装质量问题，我社负责调换）

序

随着大数据技术的快速发展，各行各业对大数据管理应用人才的需求日益旺盛。全球管理咨询公司麦肯锡（McKinsey）分析报告显示，预计在2018年，数据科学家的缺口在14万~19万人，而懂得如何利用大数据做决策的分析师和经理的岗位缺口将达到150万人。

相应的，在教育领域，2017年在教育部公布的高校新增专业名单中，有250所高校成功申请了“数据科学与大数据技术”本科专业。同时，该专业名单中也增设了“大数据管理与应用”的新本科专业。大数据管理与应用专业最早的名称是商务数据科学，由管理科学与工程类专业教学指导委员会于2015年设立，在调研课题“新专业建设研究”中提出。课题组经过近两年的工作，详细调查研究了国际教育动态和国内大数据分析人才需求情况，并将课题成果分别在教育部高等学校教学指导委员会2015年年会和2016年年会上进行报告和讨论，征求全体委员意见和建议。在此基础上，形成了“管理科学与工程类专业教学指导委员会关于增设商务数据科学专业的建议”，并于2017年3月以专家建议的形式上报教育部高等教育司。考虑到大数据应用范围的广泛性，在《2017年度普通高等学校本科专业备案和审批结果》中，最终确定名称为“大数据管理与应用”。

作为互联网大数据环境下新设的专业方向，大数据管理与应用主要在两个方面形成了较为明显的区别和特点：一是在专业知识方向，侧重于大数据环境下相关数据的应用处理，与当前社会企业需求现实结合更为紧密，因此对学生掌握相关新数据分析工具和方法的要求较高，并能适应海量数据计算的实际应用要求；二是在专业方向定位上，侧重于数据处理这一专门领域，撤除了传统信息管理与信息系统专业中信息系统分析设计和其他信息管理类课程，如信息咨询、信息资源管理等，在数据处理方向，对数据获取、数据整理、数

据存储、数据分析和数据管理等大数据处理五大关键环节专门进行了深入探讨和强调。

本书是在作者科研团队多年讲授相关课程和从事相关课题研究的基础上凝练而成，同时也吸收了国内外学者的相关成果。本书从大数据管理应用的各个常见方面，主要对相关技术应用、理论研究、主要问题和典型案例进行了较为全面的阐述，具体包括大数据采集、大数据预处理、大数据存储与计算、多源异构大数据分析、大数据知识融合技术和大数据的应用管理等内容。在理论阐述上力求简洁扼要、深入浅出，在应用介绍上则力求清晰、详尽而不累赘。因此，本书是一本适合管理人员、技术人员、相关专业本科及硕士生学习大数据管理和应用的参考书。全书共分7章，其中第1章由刘光徽执笔，第2章和第4章由周挽澜执笔，第3章由蒋伟伟执笔，第5章由陈俊鹏执笔，第6章由苏震执笔，第7章由郑怀丽执笔。全书最后由曹杰和李树青统一审定和校对。

本书的写作得到了南京财经大学信息工程学院同仁的大力支持，科学出版社也给予了很大的帮助，在此一并表示感谢！

限于作者水平，书中不足之处在所难免，殷切期望有关专家和广大读者批评指正。

作 者

2018年5月

目 录

第 1 章 引论 \ 1

- 1.1 大数据背景下的智能商务概述 \ 1
- 1.2 大数据在智能商务中的应用 \ 11
- 1.3 商务大数据的挑战性问题 \ 22

第 2 章 大数据采集 \ 25

- 2.1 网络爬虫技术 \ 25
- 2.2 社交平台大数据采集方法 \ 33
- 2.3 轨迹大数据采集 \ 41

第 3 章 大数据预处理 \ 48

- 3.1 数据预处理技术 \ 48
- 3.2 模式/本体对齐 \ 55
- 3.3 实体链接 \ 61
- 3.4 大数据溯源技术 \ 65

第 4 章 大数据存储与计算 \ 70

- 4.1 大数据的基础架构 \ 70
- 4.2 大数据的存储方案 \ 76
- 4.3 大数据计算方案 \ 86

第 5 章 多源异构大数据分析 \ 92

- 5.1 大数据建模 \ 93
- 5.2 复杂语义抽取 \ 102
- 5.3 大数据群智计算与商务智能应用 \ 110

第 6 章 大数据知识融合技术 \ 117

6.1 知识模型与建模 \ 117

6.2 知识表示与演化规律研究 \ 123

6.3 个性化知识融合 \ 133

6.4 面向网络大数据的知识融合实例 \ 145

第 7 章 大数据的应用管理 \ 149

7.1 大数据在互联网领域的应用——推荐系统 \ 149

7.2 大数据在互联网领域的应用——精准营销 \ 154

7.3 大数据在物流配送领域的应用 \ 157

参考文献 \ 160

第1章 引 论

1.1 大数据背景下的智能商务概述

1.1.1 大数据时代的背景

近年来，随着互联网、云计算、移动通信和物联网的迅速发展，数以亿计的用户使互联网服务时时刻刻都在产生巨量的交互。互联网、移动互联网、物联网、车联网、GPS（global positioning system，全球定位系统）、医学影像、安全监控、金融等领域都在疯狂产生着数据：全球每秒钟发送 290 万封电子邮件；每天上传到 Youtube 的视频总时长达 28 800 个小时；Twitter 上每天发布 5 000 万条消息；亚马逊上每天将产生 630 万笔订单；网民每个月在 Facebook 上要花费 7 000 亿分钟；Google 上每天需要处理的数据达 24PB（1PB 等于 100 万 GB）。

根据国际数据公司（International Data Corporation，IDC）做出的估测，数据一直都在以每年 50% 的速度增长，也就是说每两年就增长一倍（大数据摩尔定律），并且大量新数据源的出现导致非结构化、半结构化数据爆发式增长，这意味着人类在最近两年产生的数据量相当于之前产生的全部数据量。预计到 2020 年，全球将总共拥有 35 亿 ZB（1ZB 等于 1 万亿 GB）的数据量，与 2010 年相比，数据量将增长近 30 倍。

随着云计算、物联网和移动互联网、社交媒体等新兴信息技术和应用模式的快速发展，信息技术与人类世界政治、经济、军事、科研、生活等方方面面不断交叉融合，全球数据量急剧增加，推动人类社会迈入大数据（big data）时代。大数据时代的到来迅速引起了科技界和企业界甚至世界各国政府的关注。

最早提出大数据时代到来的是全球知名咨询公司麦肯锡。麦肯锡称：“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。”大数据时代的到来,使要处理的数据量巨大且增速极快,而业务需求和竞争压力对数据处理的实时性和有效性又提出了更高要求,传统的常规技术手段根本无法应对。

在电子商务领域,随着大数据的日益兴起和全方位的发展,相关实践和研究均日益呈现出一些具有重要意义的变化趋势。近十余年来,电子商务的发展大致经历了三次变革。2000年左右,门户网站和在线零售兴起,电子商务逐渐成为人们生活中不可或缺的购物方式,称为.com 浪潮。2005年,随着社会网络和移动网络的迅速崛起,紧密联系在一起的社会化商务和移动商务为电子商务带来了深层次的变革,企业与消费者之间的关系趋于平等、互动和相互影响,大量新型商务模式和基于位置的服务如雨后春笋般涌现。线上电子商务的巨大成功促使传统零售业企业开始思索如何转型发展,而线上企业也逐渐陷入单纯线上交易时商业模式难以进一步拓宽的困境。2015年8月,京东以43亿元战略入股永辉超市,获取了永辉10%的股份;阿里以283亿元战略入股苏宁云商,以19.99%的占股成为苏宁的第二大股东。这两个事件说明线上电商企业正在积极寻求线下实体店资源,而线下零售商也正在寻求与线上电商企业在业务、战略、资本等多个层面进行合作。这两种趋势均促使O2O(online-to-offline,线上到线下)模式的迅速崛起与发展,使得O2O模式同时得到拥有在线业务的企业及传统零售企业的广泛关注,并成为富有前景和竞争力的电子商务新兴发展模式。

1.1.2 大数据的概念

大数据是近年来的一个技术热点,历史上,数据库、数据仓库、数据集市等信息管理领域的技术,很大程度上也是为了解决大规模数据的处理问题。被誉为数据仓库之父的Bill Inmon早在20世纪90年代就经常提及大数据。2011年5月,在以“云计算相遇大数据”为主题的EMC World 2011会议中,EMC公司提出了大数据概念。

大数据是指大小超出了常用软件工具在运行时间内可以承受的收集、管理和处理数据的能力的数据集。大数据是由于目前常用软件的存储模式与能力、计算模式与能力不能满足存储与处理现有数据集规模而产生的相对概念。

数据中隐藏着许多有价值信息，在以往需要相当多的时间和成本才能提取这些信息，如沃尔玛或 Google 这类企业都要付出高昂的代价才能从大数据中挖掘出有用信息。而当今的各种资源如智能硬件、云架构和开源软件使得大数据的处理更为方便，成本也大幅度降低，即使是在车库中创业的公司也可以用较低的价格租用云服务时间。

对于企业来讲，大数据的价值体现在两个方面：分析使用和二次开发。对大数据进行分析能揭示其中隐藏的信息，如零售业中对门店销售情况、地理位置和社会信息的分析能提升对客户理解。对大数据的二次开发则是那些成功网络公司的长项。例如，Facebook 通过结合大量用户信息，推出高度个性化的定制服务，并创造出一种新的广告模式。这种通过大数据创造出新产品和服务的商业行为并非巧合，Google、Yahoo!、亚马逊和 Facebook 都是大数据时代的创新者。

大数据的特征可以用四个“V”来形容。

(1) 规模性 (volume)。大数据的起始计量单位是 P (1 000 个 T)、E (100 万个 T) 或 Z (10 亿个 T)。非结构化数据的规模和增长速度，比结构化数据增长快 10~50 倍，是传统数据仓库的 10~50 倍。

(2) 多样性 (variety)。大数据的类型可以包括网络日志、音频、视频、图片、地理位置信息等，具有异构性和多样性的特点，没有明显的模式，也没有连贯的语法和句义，多类型的数据对数据处理能力提出了更高的要求。

(3) 高速性 (velocity)。处理速度快，时效性要求高，需要实时分析而非批量式分析，数据的输入、处理和分析连贯地进行，这是大数据分析有别于传统数据挖掘最显著的特征。

(4) 价值密度低 (value)。大数据价值密度相对较低。随着物联网的广泛应用，信息感知无处不在，信息海量，但价值密度较低，存在大量不相关信息。因此，需要对未来趋势与模式做可预测分析，利用机器学习、人工智能等进行深度复杂分析。而如何通过强大的机器算法更迅速地完成数据的价值提炼，是大数据时代亟须解决的难题。

规模性、多样性、高速性、价值密度低是大数据的显著特征，或者说，只有具备这些特点的数据，才是大数据。

面对大数据的全新特征，既有的技术架构和路线已经无法高效地处理如此庞大的数据，而对于相关组织来说，如果投入巨大成本采集的信息无法得到及时处理以反馈出有效信息，那将得不偿失。可以说，大数据时代对人类的数据驾驭能力提出了新的挑战，也为人们获得更为深刻、全面的洞察能力提供了前

所未有的空间。

1.1.3 大数据融合

大数据融合 (big data fusion) 是一种数据处理过程, 在遵循一定规则的情况下, 运用计算机技术对感知数据进行分析、处理, 以满足相关决策和任务的需求。

大数据融合包含了三层含义: ①数据的全空间。即数据分为确定的和模糊的、全空间的和子空间的、同步的和异步的、数字的和非数字的数据, 它是复杂的多维多源的数据, 覆盖全频段。②数据的融合。融合不同于组合, 组合指的是外部性, 融合指的是内部性, 它是系统动态过程中的一种数据综合加工处理。③数据的互补过程。数据表达方式的互补、结构的互补、功能的互补、不同层次的互补是数据融合的核心, 只有互补数据的融合才可以使系统发生质的飞跃。

大数据融合的本质是运用合适的数据处理算法和融合模式, 对多维数据进行处理、融合分析, 进一步提高数据质量, 从而实现知识提取。

在大数据环境下需要的关键技术主要针对海量数据的存储和运算。大数据融合是最大限度发挥大数据价值的一种手段, 它的实现可以使人类对世界的探索 and 认识向新的深度和广度拓展。不同于传统的数据集成或知识库技术, 探索大数据融合需要大跨度、深层次和综合性的研究方法。

面向大数据的数据融合仍然是一个需要深入研究的问题, 具体而言, 需要在现有的数据集成与融合技术的基础上, 结合大数据的异构性、冗余性和相关性等特点, 研究大数据的数据融合和集成方法, 以有效地解决大数据获取的全面性和一致性问题。

1.1.4 大数据的应用

大数据的类型大致可分为三类: ①传统企业数据 (traditional enterprise data), 包括 CRM Systems 的消费者数据、传统的 ERP (enterprise resource planning, 企业资源计划) 数据、库存数据以及账目数据等; ②机器和传感器数据 (machine-generated/sensor data), 包括呼叫记录 (call detail records)、

智能仪表、工业设备传感器、设备日志、交易数据等；③社交数据（social data），包括用户行为记录、反馈数据等，如 Twitter、Facebook 这样的社交平台上产生的数据。

大数据挖掘商业价值的方法主要分为四种：①客户群体细分，然后为每个群体量身定制特别的服务；②模拟现实环境，发掘新的需求同时提高投资的回报率；③加强部门联系，提高整条管理链条和产业链条的效率；④降低服务成本，发现隐藏线索进行产品和服务的创新。

大数据应用是利用大数据分析的结果，为用户提供决策辅助，发掘潜在价值的过程。从理论上来看，所有产业都会从大数据的发展中受益。但由于数据缺乏以及从业人员本身的原因，第一、第二产业的发展速度相对于第三产业来说会迟缓一些。

1. 大数据在互联网企业的应用

IBM 大数据提供的服务包括数据分析、文本分析、蓝色云杉（混搭供电合作的网络平台）等业务事件处理以及 IBM Mashup Center 的计量、监测和商业化服务。IBM 的大数据产品组合中，最新系列产品的 InfoSphere bigInsights 基于 Apache Hadoop。该产品组合包括 bigInsights 和 bigsheet 软件。bigInsights 是打包的 Apache Hadoop 的软件和服务，用于初始大数据的分析；bigsheet 软件用于帮助客户从大量数据中轻松、简单、直观地提取和批注相关信息，是为金融、风险管理、媒体和娱乐等行业量身定做的行业解决方案。

阿里巴巴基于对大数据价值的沉淀，将集团旗下的阿里金融与支付宝两项核心业务合并，成立了阿里小微金融。另外，为了便于在内部解决数据的交换、安全和匹配等问题，阿里集团还搭建了一个数据交换平台。在这个平台上，各个事业群可以实现数据的内部流转，实现价值最大化。

Oracle 大数据机与 Oracle Exalogic 中间件云服务器、Oracle Exadata 数据库云服务器及 Oracle Exalytics 商务智能云服务器一起组成了甲骨文面向大数据领域的集成化的系统产品组合。

EMC 获得了纽约证券交易所和 Nasdaq 的数据，提供的大数据解决方案包含 40 多个产品。

2. 大数据在金融行业的应用

作为新兴的信息技术，大数据的使用突破了传统金融机构解决信息不对称

的方式,互联网金融借助大数据在信息挖掘、分析和预测上的强大功能,抢占传统金融业务、开拓新的市场需求。大数据在金融行业创造着重要的业务价值。随着大数据在金融行业中价值的不断提升,金融体系也逐渐进行着重构。

中国人民银行个人信用评分模型就是大数据挖掘技术在风险管理中的典型应用。该信用评分模型系统称为中国评分(China score)。它由我国信贷结构的七组评分模型组成,目前在各大商业银行运行良好。该评分系统利用全国各大金融机构的所有个人信贷账户的住房贷款、汽车贷款、信用卡等业务的历史信息(人数超过6 000万人,数据积累超过3年),运用先进的数据挖掘和统计分析技术,通过对消费者的人口特征、信用历史记录、行为记录、交易记录等大量数据进行系统的分析,挖掘出蕴含在数据中的行为模式。

工商银行则运用大数据技术进行客户流失分析和管理。客户流失分析的目的是通过分析现有客户使用产品的各种信息,预测客户在之后一段时期是否会流失,从而为其提供针对性的服务,实施挽留措施。在客户流失分析中,客户的特征主要由活期存款、定期存款、中间业务、贷款业务、贷记卡业务、国际贷记卡业务和客户基本资料这七类信息描述。其中包括客户使用各业务的产品特性、交易行为描述和客户自身的年龄性别等。

大数据在金融行业应用范围较广,其他的典型案例如花旗银行利用IBM沃森电脑为财富管理客户推荐产品;美国银行利用客户点击数据集为客户提供特色服务,如有竞争的信用额度;招商银行对客户刷卡、存取款、电子银行转账、微信评论等行为数据进行分析,每周给客户发送针对性广告信息,里面有顾客可能感兴趣的产品和优惠信息。

大数据在金融行业的应用可以总结为以下五个方面:①精准营销,依据客户消费习惯、地理位置、消费时间进行推荐。②风险管控,依据客户消费和现金流提供信用评级或融资支持,利用客户社交行为记录实施信用卡反欺诈。③决策支持,利用决策树技术进行抵押贷款管理,利用数据分析报告实施产业信贷风险控制。④效率提升,利用金融行业全局数据了解业务运营薄弱点,利用大数据技术加快内部数据处理速度。⑤产品设计,利用大数据计算技术为财富客户推荐产品,利用客户行为数据设计满足客户需求的金融产品。

3. 大数据在物联网行业的应用

物联网不仅是大数据的重要来源,还是大数据应用的主要市场。在物联网中,现实世界中的每个物体都可以是数据的生产者和消费者。由于物体种类繁多

多，物联网的应用也层出不穷。

在物联网大数据的应用上，物流企业应该有深刻的体会。UPS（United Parcel Service，联合包裹速递服务公司）快递为了使总部能在车辆出现晚点的时候跟踪到车辆的位置以及预防引擎故障，在公司的货车上装有传感器、无线适配器和不间断电源。同时，这些设备也方便了公司监督管理员工并优化行车路线。UPS 为货车定制的最佳行车路径是根据过去的行车经验总结而来的。2011 年，UPS 的驾驶员少跑了近 4 828 万千米的路程。

智慧城市是一个基于物联网大数据应用的热点研究项目，迈阿密戴德县就是一个基于物联网大数据的智慧城市样板。美国佛罗里达州迈阿密戴德县与 IBM 的智慧城市项目合作，将 35 种关键县政工作和迈阿密市紧密联系起来，帮助政府领导在治理水资源、减少交通拥堵和提升公共安全方面制定决策时获得更好的信息支撑。IBM 使用云计算环境中的深度分析向戴德县提供智能仪表盘应用，帮助县政府各个部门实现协作化和可视化管理。智慧城市应用为戴德县带来多方面的收益，如戴德县的公园管理部门 2016 年因及时发现和修复跑冒滴漏的水管而节省了 100 万美元的水费。

4. 大数据在医疗健康方面的应用

医疗健康数据是持续、高增长的复杂数据，蕴含的信息价值也丰富多样。对其进行有效的存储、处理、查询和分析，可以开发出其潜在价值，深远地影响人类的健康。例如，安泰保险为了改善对代谢综合征的预测，从千名患者中选择了 102 人来完成实验。在一个独立的实验室内，通过对患者一系列代谢综合征的检测试验结果进行分析，利用最后的结果制订一个高度个性化的治疗方案。这样，患者可以通过服用他汀类药物及减重 5 磅等方式来减少未来 10 年内 50% 的发病率。如果患者体内含糖量高于 20%，医生会建议患者降低体内甘油三酯总量。

西奈山医学中心（Mount Sinai Medical Center）是美国最大最古老的教学医院，也是重要的医学教育和生物医药研究中心。该中心使用来自大数据创业公司 Ayasdi 的技术分析大肠杆菌的全部基因序列（包括超过 100 万个 DNA 变体）来了解菌株为什么会对抗生素产生抗药性。Ayasdi 的技术使用了一种全新的数学研究方法——拓扑数据分析（topological data analysis），来了解数据的特征。

微软的 HealthVault 是一个出色的医学大数据的应用软件，它于 2007 年发

布,目标是管理个人及家庭的医疗设备中的个人健康信息。目前已经可以通过移动智能设备录入健康信息,而且还可以从第三方机构导入个人病历记录。此外,其通过提供 SDK 以及开放的接口,支持与第三方应用的集成。

5. 大数据在交通运输方面的应用

目前,交通的大数据应用主要有两个方面。一方面,可以利用大数据传感器数据来了解车辆通行密度,合理进行道路规划,包括单行线路规划;另一方面,可以利用大数据来实现即时信号灯调度,提高已有线路运行能力。科学地安排信号灯是一个复杂的系统工程,必须利用大数据计算平台才能计算出一个较为合理的方案。科学的信号灯安排将会使已有道路的通行能力提高 30%左右。在美国,政府依据某一路段的交通事故信息来增设信号灯,降低了 50%以上的交通事故率。机场可以依靠大数据安排航班起降来提高航班管理的效率,航空公司利用大数据可以提高上座率,降低运营成本。铁路部门利用大数据可以有效安排客运列车和货运列车,提高效率、降低成本。

6. 大数据在教育方面的应用

大数据不仅可以帮助教师在课堂上改善教育教学,在重大教育决策制定和教育改革方面,大数据更是发挥着重要作用。美国利用数据来诊断处在辍学危险期的学生,探索教育开支与学生学习成绩提升的关系,探索学生缺课与成绩的关系。例如,美国某州公立中小学的数据分析显示,在语文成绩上,教师高考成绩和学生成绩呈现显著的正相关。也就是说,语文教师的高考成绩与他们现在所教学生的学习成绩有很明显的关系,教师的高考成绩越好,学生的语文成绩也越好。对于这个关系,可以进一步探讨其背后真正的原因。其实,教师高考成绩高低某种程度上是教师的某个特点在起作用,而正是这个特点对教好学生起着至关重要的作用,教师的高考分数可以作为挑选教师的一个指标。如果有了充分的数据,便可以发掘更多的教师特征和学生成绩之间的关系,从而为挑选教师提供更好的参考。

大数据还可以帮助家长和教师甄别出孩子的学习差距和有效的学习方法。例如,美国的麦格劳-希尔教育出版集团就开发出了一种预测评估工具,帮助学生评估他们已有的知识和达标测验所需程度的差距,进而指出学生有待提高的地方。评估工具可以让教师跟踪学生学习情况,从而找到学生的学习特点和方法。有些学生适合按部就班,有些则更适合图式信息和整合信息的非线性学习。这些

都可以通过大数据搜集和分析很快识别出来,从而为教育教学提供坚实的依据。在国内,尤其是北京、上海、广州等城市,大数据在教育领域已经有了非常多的应用,如慕课、在线课程、翻转课堂等,其中就应用了大量的大数据工具。

7. 大数据在政府机构的应用

目前,世界上一些国家已经在政府部门开始推广大数据应用。通过分析和比较这些先发国家的大数据应用,我们能了解当前和未来需要大数据应用聚焦和服务的地方,并为其他国家开展大数据应用提供借鉴。

在西班牙首都马德里,整合警察、消防、医疗系统,使救援时间大幅度缩短,巡逻队、消防车、救护车能够在8分钟内到达81%的突发事件现场;在新加坡,智能交通综合信息管理平台在预测交通流速和流量方面有高达85%的准确率,能通过有效的引导和干预,显著提升高峰时段的车辆通行效率;在我国江苏省苏州市,覆盖城乡的信息化防控网络,在警力与人口配比不足万分之十的情况下,使打击处理案件数、刑拘转捕率、技术支撑率均为江苏省最高,实现了“以十抵万”的办案效率。

在公共管理领域,国内外一些先行者已经在运用大数据的方法,通过多渠道的数据采集和快速综合的数据处理,提升治理社会的能力,实现政府公共服务的技术创新、管理创新和服务模式创新。大数据在公共管理领域的应用,不仅使传统难题变得迎刃而解,更成为新时期应对新挑战、解决新问题的必然选择。

利用大数据治国,美国政府早已先行一步,奥巴马认为,数据在未来将是陆权、海权、空权之外的另一种国家核心资产。美国白宫科技政策办公室在2012年3月发布《大数据研究和发展计划》,同时组建“大数据高级指导小组”,以协调政府在大数据领域的2亿多美元投资,这标志着美国把大数据提高到国家战略层面,形成全体动员的格局。

根据麦肯锡的报告,大数据技术可为欧盟23个最大的政府公共部门管理活动的成本提供15%~20%的下降空间,在未来10年每年创造1500亿~3000亿欧元的价值,并将公共部门的预计效率提高0.5个百分点。

对各个国家和地区大数据实践的研究表明,在以下几个方面,大数据的应用可以进一步协助政府机构发挥职能作用。

(1) 重视应用大数据技术,盘活各地云计算中心资产,把原来大规模投资产业园、物联网产业园从政绩工程改造成智慧工程。

(2) 在安防领域,应用大数据技术提高应急处置能力和安全防范能力。

(3) 在民生领域,应用大数据技术提升服务能力和运作效率,以及提供个性化的服务,如医疗、卫生、教育等部门。

(4) 解决在金融、电信等领域中数据分析的问题,虽然一直得到极大的重视,但由于存储能力和计算能力的限制,只局限于对交易型数据的统计分析。

1.1.5 大数据电子商务模式

电子商务是将传统的商务活动转移到网络平台的方式,通过安全和高效的网络信息技术实现了电子商务的各项活动。一般来说双方并不见面,更多是通过 App 或者 Web 浏览器来完成交易。所以,这种商务活动的特点也注定其与大数据时代有着密不可分的关系,而大数据时代的到来也预示着电子商务的新一轮技术革新,如何在技术革新下保持服务水平的增长也是电子商务整个行业急需解决的问题。众所周知,电子商务经济活动的开展需要以市场为导向,其最大目的就是吸引用户,实现自身的经济效益。这就要求电子商务企业在开展经济活动的过程中借助大数据分析总结出一套行业应用垂直标准,从企业的文化战略、产品营销、市场开拓、技术管理等内容出发,促进自身产业结构的转型。

大数据时代下电子商务服务模式有以下几点创新。

1. 个性化的数据导购模式

大数据时代的到来,使得社会对数据的搜集能力有了质的提升,数据的搜集方式也发生了根本性的变化,所搜集到的数据非常准确地反映出市场的真实情况。因此,电子商务在运营过程中一定要充分合理地利用数据信息,加强数据导购,实现个性化的数据导购模式。用户在利用互联网浏览网页的时候,会留下消费痕迹,这些消费痕迹蕴含着非常丰富的数据信息,电子商务平台必须要利用这些数据信息,挖掘出其中隐藏的价值,进行个性化导购。第一,个性化广告。个性化广告是指在分析用户的网页浏览记录、消费记录的情况下,利用数据信息库自动对用户的消费理念和消费习惯进行调查,在相关页面向用户推荐同类型的产品。第二,个性化推荐。用户在网络环境下,面对各式各样的商品,虽然选择很多,但是选择起来却非常困难。个性化推荐要求为用户推荐更多有价值的产品,尽可能地缩短用户选购的时间,将用户从烦琐的商品信息中“拯救”出来。