

Dashuju Dui Gaoxiao Jiaoyu De Yingxiang

大数据

对高校教育的影响

主编◎王鹤蒙 宋 征

中国矿业大学出版社

China University of Mining and Technology Press

大数据对高校教育的影响

主 编 王鹤蒙 宋 征
副主编 钱建新 叶丛如
刘红卫

中国矿业大学出版社

图书在版编目(CIP)数据

大数据对高校教育的影响 / 王鹤蒙, 宋征主编. —

徐州: 中国矿业大学出版社, 2016. 6

ISBN 978-7-5646-3133-8

I. ①大… II. ①王… ②宋… III. ①信息技术—应用—高等教育—研究—中国 IV. ①G649.2

中国版本图书馆 CIP 数据核字(2016)第 129885 号

书 名 大数据对高校教育的影响

主 编 王鹤蒙 宋 征

责任编辑 陈振斌

出版发行 中国矿业大学出版社有限责任公司

(江苏省徐州市解放南路 邮编 221008)

营销热线 (0516)83885307 83884995

出版服务 (0516)83885767 83884920

网 址 <http://www.cumtp.com> **E-mail:** cumtpvip@cumtp.com

印 刷 江苏徐州新华印刷厂

开 本 880×1230 1/32 **印张** 4.25 **字数** 88 千字

版次印次 2016 年 6 月第 1 版 2016 年 6 月第 1 次印刷

定 价 16.00 元

(图书出现印装质量问题, 本社负责调换)

前 言

近年来,随着互联网和物联网的快速发展和广泛应用,世界处于各种数据指数式增长的环境中,进入大数据时代。大数据日益成为推动各个领域变革的强劲力量。从高等教育领域来看,大数据在教学、科研和管理方面引发的创新与变革日益显现。如何能有效地将巨量的数据资源转化为丰硕的教育研究成果,应用于改善教育的决策与实践,对高等教育界来说意义重大。

本书介绍了大数据的由来、大数据出现的背景及大数据的作用,阐述了大数据对高校教育的影响,并指出大数据在高校教育中应用的可能以及如何使用大数据处理高校 IT 安全问题,同时提出高校大数据发展的若干可能。通过阅读本书,读者可以认识到什么是大数据、大数据有什么作用、身边的大数据在哪。

本书主要有下列特色:

(1) 精心选择国内外使用大数据的例子,让读者了解到底什么是大数据,大数据能有什么作用。

(2) 由浅入深地介绍大数据对高等教育的影响和推动作用,符合读者阅读心理。

(3) 数据信息是高等学校的重要资产,对数据的有效管理是数字化校园建设的重要内容之一。本书以高校信息化建设为背景,探讨了未来大学校园的若干可能以及智慧校园的发展目标。

(4) 在大数据时代,海量数据的分析将对教育信息化产生深刻的影响和冲击,本书指出数据分析是高校发展的关键使命,高校可依靠信息系统的作用来适应社会变化。

由于编者水平有限,加上时间仓促,书中难免存在疏漏之处,恳请读者批评指正。编者邮箱:wanghm266@163.com。

王鹤蒙

2016年4月

目 录

第一章 大数据出现的背景及走向	1
第一节 大数据的出现	1
第二节 大数据在国外的应用	4
第三节 大数据在中国的应用	9
第四节 大数据的走向	10
第二章 认识大数据	14
第一节 什么是大数据	14
第二节 传统数据与大数据的区别	16
第三节 高等院校如何提高数据质量	19
第三章 大数据影响了高校教育	24
第一节 从可汗学院到慕课	24
第二节 从平常的教育中看大数据	28
第三节 让教育发现每一个学生	31
第四节 技术重新定义教育	35
第五节 高校教师在大数据时代如何站稳脚跟	38
第六节 影响高校信息化建设的因素	41

第七节 大数据时代的教育变革	44
第四章 大数据在高校教育中的应用	53
第一节 高校在大数据浪潮中扮演着重要 角色	53
第二节 高校中的大数据	55
第三节 大数据在高校教育信息化建设中的 应用	58
第四节 大数据在高校教育信息化应用中的 挑战	60
第五节 大数据在高校教育信息化应用中的 策略	65
第五章 让沉睡的数据说话	67
第一节 高校教育的大数据在哪里	67
第二节 高校教育研究如何跟上大数据时代	70
第三节 大数据给高校教育带来怎样的可能	73
第四节 数据分析是关键	76
第六章 大数据如何处理高校 IT 安全问题	85
第七章 高校大数据发展的若干可能	93
第一节 未来的大学校园怎么样	93
第二节 大数据可能给高校教育领域带来的 威胁	98

第三节	大数据背景下高校信息化面临 新挑战	99
第四节	大数据背景下的高校智慧校园建设	101
第八章	高校如何在大数据时代掘宝	109
第一节	高校可借大数据自我提升	109
第二节	高校可依靠信息系统适应社会变化	111
第三节	高校要大胆拥抱慕课革命	119
参考文献		125

第一章 大数据出现的背景及走向

第一节 大数据的出现

数据是人类对客观世界的记录,人类世界的本质就是数据。随着信息技术的进步,数据的收集、保存、维护、使用等任务,成为横跨各个领域的现象和挑战。2011年5月,全球知名咨询公司麦肯锡(Mckinsey)发布了《大数据:创新、竞争和生产力的下一个前沿领域》报告,首次提出了“大数据”的概念,并在报告中指出“数据已经渗透到每一个行业和相关职能领域,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”。麦肯锡关于“大数据”的报告已经上过《纽约时报》《华尔街日报》的专栏封面,进入美国白宫官网的新闻,现身在国内一些互联网主题的讲座沙龙中,甚至被嗅觉灵敏的国金证券、国泰君安、银河证券等写进了投资推荐报告。在我们的日常生活中,人人都在创造数据,也在利用和分享数据。视频、音频、图像、数字等多种交互方式的丰富,让我们已经进入了数据信息爆炸的阶

段。哈佛大学社会学教授加里·金称：“这是一场革命，庞大的数据资源使得各个领域开始了量化进程，无论学界、商界还是政府，所有领域都将开始这种进程。”美国、欧洲、日本政府纷纷提出了“大数据的研究和发展计划”，更是将“大数据”提升到了全球性战略发展的高度。随着“大数据”时代的到来，人们在实践中很快认识到，通过数据的开放、整合和分析，能够发现新的知识、创造新的价值，从而促进社会进步和发展。在这个快速发展的时代，大数据给高等教育带来哪些影响和挑战，非常值得关注。

进入 2012 年，大数据(Big Data)一词越来越多地被提及，人们用它来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术发展与创新。

数据正在迅速膨胀并变大，它决定着企业的未来发展。虽然现在企业可能并没有意识到数据爆炸性增长带来问题的隐患，但是随着时间的推移，人们将越来越多地意识到数据对企业的重要性。“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与潜力。

正如《纽约时报》2012 年 2 月的一篇专栏中所称，“大数据”时代已经降临，在商业、经济及其他领域中，决策将日益基于数据分析而作出，而并非基于经验和直觉。

哈佛大学社会学教授加里·金说：“这是一场革命，庞大的数据资源使得各个领域开始了量化进程，无论学术界、商界还是政府，所有领域都将开始这种进程。”

“大数据”在物理学、生物学、环境生态学等领域以及

军事、金融、通讯等行业存在已有时日,却因为近年来互联网和信息行业的快速发展而引起人们关注。

随着云时代的来临,大数据也吸引了越来越多的关注。著名云台的分析师团队认为,大数据通常用来形容一个公司创造的大量非结构化和半结构化数据。这些数据在下载至关系型数据库用于分析时会花费过多的时间和金钱。大数据分析常和云计算联系到一起,因为实时的大型数据集分析需要像 MapReduce 一样的框架来向数十、数百或甚至数千的电脑分配工作。

“大数据”在互联网行业则指的是这样一种现象:互联网公司在日常运营中生成、累积的用户网络行为数据。这些数据的规模是如此庞大,以至于不能用 G 或 T 来衡量,大数据的起始计量单位至少是 P(1000 个 T)、E(100 万个 T)或 Z(10 亿个 T)。

大数据到底有多大?一组名为“互联网上一天”的数据告诉我们:一天之中,互联网产生的全部内容可以刻满 1.68 亿张 DVD;发出的邮件有 2940 亿封之多(相当于美国两年的纸质信件数量);发出的社区帖子达 200 万个(相当于《时代》杂志 770 年的文字量);卖出的手机为 37.8 万台,高于全球每天出生的婴儿数量 37.1 万……

截至 2012 年,数据量已经从 TB(1 TB=1024 GB)级别跃升到 PB(1 PB=1024 TB)、EB(1 EB=1024 PB)乃至 ZB(1 ZB=1024 EB)级别。根据国际数据公司(IDC)监测统计结果,2011 年全球数据总量已经达到 1.8ZB,到 2020 年全球将总共拥有 35ZB 的数据量。

1.8 ZB 有多大? 它等于 1.8 万亿 GB。如果以每部 2 小时的高清电影来计算,1.8 ZB 可容纳超过 2000 亿套高清电影,即使全中国人聚在一起,24 小时不休息也要 5.9 年才能看完,而 35 ZB 的数据则需要花 114.7 年,超过了人类平均寿命。

第二节 大数据在国外的应用

一、在 PRADA 的试衣间

PRADA(普拉达,一个意大利的奢侈品牌)在纽约的旗舰店中每件衣服上都有 RFID(射频识别)码。每当一个顾客拿起一件 PRADA 衣服进试衣间,RFID 会被自动识别。同时,数据会传至 PRADA 总部。每一件衣服在哪个城市,哪个旗舰店什么时间被拿进试衣间,停留了多长时间,数据都被存储起来加以分析。如果有一件衣服销量很低,以往的做法是直接“干掉”。但如果 RFID 传回的数据显示这件衣服虽然销量低,但进试衣间的次数多,也许这件衣服的下场就会“截然不同”:对这件衣服的某个细节作微小改变就会重新创造出一件非常流行的产品。

二、对美国总统大选的支撑

就在美国总统奥巴马成功击败对手罗姆尼,再次赢得美国总统选举的当天,《时代》杂志撰写了一篇文章,描述了奥巴马总统获胜背后的秘密——数据挖掘。用竞选工

作组发言人 Ben LaBolt 的话来形容：奥巴马团队拥有“核代码”，数据是能够击败罗姆尼的最根本优势！

的确，美国总统奥巴马的再次当选创造了奇迹：在他获胜前的 70 年时间里，没有一位美国总统能够在全国失业率高于 7.4% 的情况下连任成功；而他与对手罗姆尼的一系列“激战”也让整个竞选过程变得扑朔迷离！我们且不论谁在政治上更英明，但是说到如何获得更多选民支持以及如何让他们掏腰包，奥巴马团队绝对比罗姆尼团队更加聪明：奥巴马与罗姆尼均获得了近 10 亿美金筹款，而其网络筹款是罗姆尼的 2 倍；奥巴马在整个竞选过程中的花销不到 3 亿美金，而罗姆尼花了近 4 亿美金却仍然败选；奥巴马最终以 332 票赢得选举，高出罗姆尼近 100 张投票，而在大选前一周的一项民调中显示，55% 的被调查选民都认为罗姆尼比奥巴马更具有未来视野！

这一连串的数字显然推翻了美国历史上总统选举的定律：谁筹的钱越多谁胜出可能性越大，谁花的钱越多谁就会赢。奥巴马团队能取得颠覆性的胜利，是因为他们做到了三个最根本的目标：让更多的人掏更多的钱，让更多的选民投票给奥巴马，让更多的人参与进来！这些都是因为他们对选民的认知达到了“微观”层面：每个选民最有可能被什么因素说服？每个选民在什么情况下最有可能掏腰包？什么样的广告投放渠道能够最高效获取目标选民？

通过这些分析，奥巴马团队制定了相应的策略，并赢得了大量草根阶层选民的支持和捐赠。一项民调显示：80% 的美国选民认为奥巴马比罗姆尼让他们感觉更加重

视自己。结果是,奥巴马团队筹得的第一个1亿美金中,98%来自于小于250美金的小额捐款,而罗姆尼团队在筹得相同数额捐款的情况下,这一比例仅为31%。

让这一切“微观智能”成为可能的是数据。正如竞选总指挥吉姆·梅西纳(Jim Messina)所说,在整个竞选活中,没有数据做支撑的假设很少存在。奥巴马团队运用数据挖掘技术在美国政坛上取得的胜利,告诉了我们一个不争的事实:今天,我们已然进入了一个“微竞争”的时代,在激烈的市场竞争中,谁能够深入地了解他的每一个用户的个性化需求,谁就能在竞争中击败对手,获取胜利。

事实上,许多国际商业巨头早已悄悄地实践着“微竞争”:深入了解他们的每一个用户,并为他们提供个性化服务。他们在市场上获得巨大成功的缘由,与美国总统奥巴马获选同出一辙:数据挖掘。

为了筹到10亿美金的竞选款,奥巴马的数据挖掘团队使用两年的时间搜集、存储和分析了大量数据。他们注意到影星乔治·克鲁尼(George Clooney)对美国西海岸40岁至49岁的女性具有非常大的吸引力。她们无疑是最有可能为了在好莱坞与克鲁尼和奥巴马共进晚餐而不惜自掏腰包的一个群体。克鲁尼在自家豪宅举办的筹款宴会上,为奥巴马筹集到数百万美元的竞选资金。

之后,当奥巴马团队决定在东海岸物色一位对于这个女性群体具有相同号召力的影星时,数据团队发现莎拉·杰西卡·帕克(Sarah Jessica Parker,《欲望都市》的女主角)的粉丝们也同样喜欢竞赛、小型宴会和名人。于是,一

个与奥巴马共进晚餐的“竞争”便诞生了,那就是争夺在杰西卡·帕克的纽约 West Village 豪宅美餐的机会。“克鲁尼效应”被成功地复制到了东海岸。

通过对粉丝们的信息分析,奥巴马团队成功地满足了她们与钟爱的明星共进晚餐的愿望,也成功地让她们打开钱包。同样,聪明的商家通过用户的购买历史记录分析来建立模型,为他们量身预测未来的购物清单,进而设计促销活动和个性服务,让他们源源不断地为之买单。

在本次竞选中,奥巴马团队的投票动员绝不是千篇一律的:对于不同的用户,他们动员的渠道及采取的互动方式都因人而异。当奥巴马决定在社交新闻网站 Reddit 上回答问题时,多名总统高级助理并不清楚此事。一名官员表示:“我们为何将奥巴马放在 Reddit 上? 因为我们发现很大一部分目标选民在 Reddit 上。”而在“摇摆州”的电话动员上,他们也发现,一个摇摆州志愿者打来电话的效果要优于一个从非摇摆州(如加州)志愿者打来的电话。

此外,奥巴马团队还首次利用 Facebook 进行大规模的投票动员,这模仿了现场组织者挨家挨户敲门的方式。在竞选活动的最后几周,下载某一款应用的用户收到了多条消息,其中包含他们在“摇摆州”好友的照片。他们被告知,可以通过点击按钮,呼吁这些目标选民采取行动,例如进行投票注册、更早的投票,以及参与到投票中。奥巴马竞选团队发现,大约 1/5 收到 Facebook 请求的选民做出了响应,这在很大程度上是因为请求来自他们熟悉的人。

三、在连锁超市中的应用

在商业上,以数据分析为支撑的决策,也让商家获得了不少好处,Tesco 就是一个例子。作为全球利润第二大的零售商(仅次于沃尔玛),这家英国超级市场巨人从用户行为分析中获得了巨大的利益。从其会员卡的用户购买记录中,Tesco 可以了解一个用户是什么“类别”的客人,如速食者、单身、有上学孩子的家庭,等等。这样的分类可以帮助 Tesco 设计个性化的服务,比如,通过邮件或信件寄给用户的促销可以变得十分个性化,店内的上架商品及促销也可以根据周围人群的喜好、消费的时段来更加有针对性,从而提高货品的流通。这样的做法为 Tesco 获得了丰厚的回报,仅在市场宣传一项,就能帮助 Tesco 每年节省 3.5 亿英镑的费用。

在 2012 年年初,一名男子闯入了他家附近的 Target 店铺(Target 是一家美国零售连锁超市):“你们怎么能这样!”男人向店铺经理大吼,“你们竟然给我 17 岁的女儿发婴儿尿片和童车的优惠券,她才 17 岁啊!”店铺经理不知道发生了什么,立刻向来者道歉,表明那肯定是个误会。然而,经理并没有意识到,公司正在运行一套数据挖掘系统。

一个月后,这个愤怒的父亲打来电话道歉,因为 Target 发来的婴儿用品促销广告并不是误发,他的女儿的确怀孕了。事实上,Target 创建了一套女性购买行为在怀孕期间产生变化的模型,不仅如此,如果用户从他们的店铺中购买了婴儿用品,Target 在接下来的几年中会根据婴儿的生

长周期情况定期给这些顾客推送相关产品,使这些客户形成长期的购买习惯。

四、预测甲型 H1N1 流感

2009 年,波及美国及全世界的甲型 H1N1 流感爆发的几周前,互联网公司——谷歌的工程师们在《自然》杂志发表的论文,令公共卫生官员们感到震惊。文中解释了谷歌为何能够预测流感传播,而且具体到特定的地区。

谷歌通过分析人们上网搜索记录,例如“咳嗽”、“发烧”等词条,总共处理了 4.5 亿个不同的数字模型,并将得出的预测与 2007 年、2008 年美国疾控中心记录的实际流感病例进行对比。谷歌公司发现,他们的预测与官方数据相关性高达 97% 且判断及时。

第三节 大数据在中国的应用

一、在粮食统计方面的应用

中国的粮食统计是一个老大难的问题。传统的统计办法,依靠统计人员层层上报,水分很大,数据的真实性令人怀疑。在一次会议上,国家统计局原总经济师姚景源讲述了他们是如何进行粮食统计的。他们采用遥感卫星,通过图像识别,把中国所有的耕地标示、计算出来,然后把中国的耕地网格化,对每个网格的耕地抽样进行跟踪、调查和统计,再按照统计学的原理,计算(或者说估算)出中国