

教育大数据研究综述

杜婧敏¹, 方海光^{1,2}, 李维杨¹, 仝赛赛¹

(1. 首都师范大学 教育技术系, 北京 100048;

2. 北京师范大学智慧学习研究院, 北京 100875)

摘要:教育大数据不仅仅是大数据在教育领域应用,而且通过教育领域反向驱动大数据技术分化为独立的分支,从而带来了传统教育技术领域长期研究问题解决的新途径,甚至可以跨越传统个性化学习的精确逻辑推理过程而直接分析全样本学习者特征。这样来促进教育管理科学化变革、促进教学模式改革、促进个性化教育变革、促进教育评价体系改革、促进科学研究变革等。这些主要依托于教育大数据的两项关键技术:教育数据挖掘和学习分析。但是我国的教育大数据研究与应用尚处于起步阶段,缺乏前期的规范化和系统化顶层设计,面临许多层次的挑战,如应用挑战、安全与隐私问题、运营挑战等,因此,特别需要从教育大数据整体进行规划和由上至下的系统综合研究。

关键词:教育大数据;教育变革;数据挖掘;学习分析

中图分类号:TP393

文献标志码:A

文章编号:1673-8454(2016)19-0001-04

一、引言

移动互联网技术相比较 PC 时代的互联网有着本质的区别,伴随着数据时代带动的非线性大数据增长过程,社会生活和城市建设过程中的教育系统也在高速发展过程中,这完全不同于过去的二十年间的信息技术对教育教学的改变进程。联合国在 2012 年发布的大数据白皮书《Big Data for Development: Challenges & Opportunities》中指出大数据的出现将会对社会各个领域产生深刻影响^[1]。无论国内外都开始了对大数据的研究,起步略有早晚之分,但从政策到区域再到机构的发展方式是共识。《国民经济和社会发展的第十二个五年规划纲要》中提出:“重点研究……海量信息处理及知识挖掘的理论与方法……”。美国教育部于 2012 年 10 月发布的《通过教育数据挖掘和学习分析促进教与学》中指出通过对教育大数据的挖掘与分析,促进美国高等院校及 K-12 学校教学系统的变革^[2]。2015 年 8 月 31 日国务院发布的《促进大数据发展行动纲要》中指出“数据已成为国家基础性战略资源”,并在启动的十大工程之一“公共服务大数据工程”中明确提出要建设教育大数据。目前,教育大数据的重要性已经提升到国家战略层面,教育大数据中心建设和区域数据共建共享应用成为主要的建设模式,对个体学习全过程数据分析和精准数据评测学习成为主要的应用场景,从而通过宏观和微观的两方面建设,带动教育领域的广泛关注和重视。

二、相关概念

大数据的 3V 定义基本是从大数据的特征出发,即:规模性(Volume)、多样性(Variety)和高速性(Velocity)。目前比较认可的是 4V 定义,如国际数据公司(IDC)认为,大数据是符合 4V 特征的数据集,即在 3V 基础上增加价值性(Value)^[3]。而 IBM 认为大数据必然具有真实性(Veracity)^[4]。维基百科对大数据的定义是:“巨量资料(Big Data),或称大数据、海量资料,指的是所涉及的资料量规模巨大到无法透过目前主流软件工具,在合理时间内达到撷取、管理、处理、并整理成为帮助企业经营决策更积极目的资讯”^[5]。

另一方面,教育大数据的定义最早从产生教育大数据的主体出发,将教育大数据分为广义的和狭义的两类:广义的教育大数据泛指所有来源于日常教育活动中人类的行为数据;狭义的教育大数据是指学习者行为数据^[2]。也有研究指出教育大数据指整个教育活动过程中所产生的以及根据教育需要采集到的,一切用于教育发展并可创造巨大潜在价值的数据集合^[6]。

基于以上研究,可以认为教育大数据的定义包含三层含义:第一个含义,教育大数据是教育领域的大数据,是面向特定教育主题的多类型、多维度、多形态的数据集合;第二个含义,教育大数据是面向教育全过程的数据,通过数据挖掘和学习分析支持教育决策和个性化学习;第三个含义,教育大数据是一种分布式计算架构方

式,通过数据共享的各种支持技术达到共建共享的思想。也就是说,我们把教育大数据定义为:面向教育全过程时空的多种类型的全样本的数据集合。教育大数据不仅仅是建设教育大数据中心,不仅仅是分析全过程学习数据,更多的是一种共享的生态思想。

与用传统方法收集的教育数据相比,教育大数据有更强的实时性、连续性、综合性和自然性,并使用不同的应用程序来分析和处理不同复杂度和深度的数据。传统教育数据收集的大多是阶段性的数据,而且大多在用户知情的情况下收集,使用的分析方法也通常是简单的统计分析方法。教育大数据收集的是整个教育教学过程中静态和动态的所有数据,可以在不影响教师和学生活动的情况下,连续记录整个教学活动的全部数据,如教学资料、互动反映和学生在每个知识点上停留的时间等。

三、教育大数据的业务分析

教育大数据是一种新的架构和思维基础,可以弱化分析模型直接进入研究目标进行分析,还可以通过数据聚集对“小现象”进行规律研究,还可以对个体进行多维度数据融合研究,再可以对目标对象进行时间序列和空间维度动态分析。面对具体教育业务,教育大数据技术对教育的管理、教学、学习、科研和评价等都产生了很大的影响。教育大数据的业务分析如图1所示。

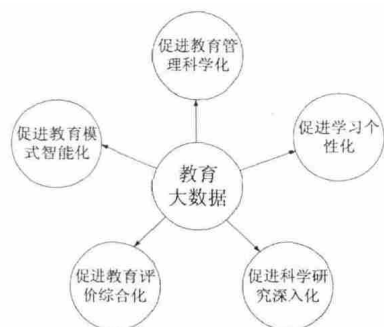


图1 教育大数据的业务分析

1. 教育大数据对教育管理的支持

传统的教育决策制定形式常被形象的称为“拍脑袋”决策,是指决策者常常不顾实际情况,以自己有限的理解、假想、推测依据直觉、冲动或趋势来制定政策。这种来自决策者“头脑发热”决策,经常处于朝令夕改的尴尬境地,教育大数据正可以帮助解决这种不足。

大数据时代,教育者将更加依赖于数据和分析,而不是直觉和经验;同样,教育大数据还将改变领导力和管理的本质。服务管理、数据科学管理将取代传统的行政管理、经验管理。技术不断发展,教育数据挖掘与分析不断深入,不仅要着眼于已有的确定关系,更要探寻隐藏的因果关系。利用大数据技术可以深度挖掘教育数据

中的隐藏信息,可以暴露教育过程中存在的问题,提供决策来优化教育管理。大数据不仅可以运行和维护各教育机构的人事信息、教育经费、办学条件和服务管理的数据,而且可以长期积累所有类型教育机构的数据,利用统计分析、应用模型等技术将数据转换为知识,最终为教育者和学习者提供科学的决策。

2. 教育大数据对教学模式的支持

教育大数据推进实现智慧学习。教师在智慧教学环境下,利用大数据技术可以更深入的了解每一个学习者的学习状况,并且与学习者的沟通更加通畅,教师的整个教学过程和学习者的学习过程更加精准化和智能化。教师对教学过程的掌握从依靠经验转向以教育数据分析为支撑,学生对于自己学习状况的了解从模糊发展到心中有数,可以更好的认识自我、发展自我、规划自我。大数据技术可以帮助教师及时调整教学计划和教学方法,有利于教师自身能力提高和职业发展。

3. 教育大数据对个性化学习的支持

除了学生学习的行为可以被记录下来外,学生在资源上的数据也可以被精确记录下来,如点击资源的时间、停留多久、问题回答正确率、重复次数、参考阅读、回访率和其他资源信息,通过大数据可以定制个人学习报告,分析学习过程潜在的学习规律,还可以找到学生的学习特点、兴趣爱好和行为倾向,并一目了然教育状态信息。大数据技术使教育围绕学习者展开,使传统的集体教育方式转向为个性学习方式。同时还伴随着教育者和学习者思维方式的改变,进一步朝着个性化学习方向迈出重要的一大步,使得精准的个性化学习成为可能。

4. 教育大数据对教育评价的支持

教育评价正在从“经验主义”走向“数据主义”,从“宏观群体”评价走向“微观个体”评价,从“单一评价”走向“综合评价”^[7]。教育大数据下教育评价的变化,不仅表现在评价思想,还包括评价方法,不仅包括对学生的评价,还包括对教学管理、评估质量等具体水平的评价。教学评估不再仅仅是由考试成绩和纪律帮助教师评价的主观传统意义上的感受,而由大量的数据感知得到,为实现教学评价的公正提供了依据,优化了教学方向。教育评价可以是多元化的,而不是仅停留在知识掌握程度这一单一维度。

5. 教育大数据对科学研究的支持

教育大数据使得从追求单向因果性转向追求复杂的多元相关性,并用直观的图形等表达方式体、系统、清晰、简洁地展现。这种新理念、新思维的创生,是实现教育创新和发展不可缺少的手段、工具和方法论^[8]。教育大数据时代,科学研究将从随机抽样、探讨因果关系走向全部数

据、寻找相关关系。大数据技术减少了研究资金的浪费,在某些问题上,数据分析为研究人员提供了个性化的服务,可以提高了研究的效率和成果的可靠性。大数据依赖于自动、连续的记录和搜集的数据比传统调查数据更加客观和中立。大数据还将改变传统学术研究的过程,信息系统依赖于自动同步,连续获得持续的行为数据,这意味着学术研究和信息技术、课题研究与实践联系在一起。

四、教育大数据转换层次模型

教育大数据具有转换层次的四个过程特征:数据、信息、知识、智慧。教育大数据技术采集的海量教育元数据,经过抽取、转换、加载,联机分析处理和统计分析等过程,转换为教育信息,又经过数据可视化将教育数据呈现出来,最后形成教育决策来指导教育者和学习的教学,这个过程数据完成了从数据到信息、知识和智慧的演变。教育大数据转换层次模型如图2所示。



图2 教育大数据转换层次模型

目前在教育大数据领域中有很多研究领域,比较主要的是数据挖掘和学习分析。两者既有相同之处,又存在一定差异。

1. 教育大数据和数据挖掘

数据挖掘就是从大量的、有噪声的、不完全的、模糊的、随机的实际应用数据中提取有效的、新颖的、潜在有用的知识的非平凡过程^[9]。目前认可度较高的一种处理模型是 Fayyad 等人设计的多处理阶段模型,在此不再赘述。在大数据时代处理数据理念上有三大转变:抽样到全体、绝对精确到效率、因果到相关^[10]。大数据时代的数据挖掘继承于传统数据挖掘又不同于传统数据挖掘。首先是分析对象不同,传统数据分析的对象是在已知的数据范围中易处理的数据,而基于大数据的数据挖掘对象时包括非结构化数据的全部数据,不能保证原始数据是完整的、清洗过和没有错误的。其次是分析基础,传统分析是建立在关系数据模型之上的,是在系统内就创立的主题之间的关系基础上进行的,而在大数据分析中,绝大部分分析基于纵裂数据库之外。再其次是分析效率,传统数

据处理过程中需要等待抽取、转换、加载等工作完成。而大数据分析是对数据进行实时分析。最后是硬件差别,在传统数据处理中所使用硬件比较昂贵。而大数据处理过程加入了对分析软件的使用,所以硬件成本较低。

近几年出现了许多有关教育数据挖掘(EDM)的文献。研究者对 EDM 研究的主题范围集中在使用数据挖掘提高机构效率和促进学生的学习过程。但是教育数据挖掘领域内还存在更广泛的话题,例如,在课程管理系统内(CMS)学生的退出和保留、个性化推荐系统、以及学生学习评价。

教育数据挖掘可以向教育者、学习者、教育管理者、家长及教育研究者提供决策帮助,最终提高学习者的学习水平。基于教育大数据的数据挖掘过程即将学习结果、学习内容、学习资源与教学行为等教育原始数据使用多种技术,如聚类、关系挖掘和模型构建等,最终达到预测学习者发展趋势、促进有效学习的目标。具体过程如图3所示。

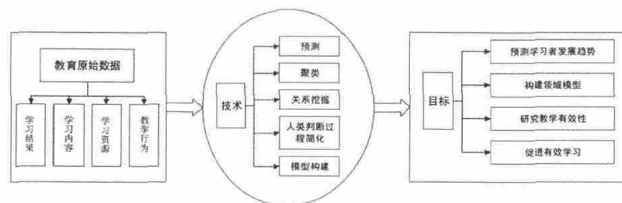


图3 基于教育大数据的数据挖掘过程

教育大数据公认的特征之一是规模性,面对海量数据,可以采取分布式文件的系统进行并行运算。对于半结构化或非结构化的数据,可以采用自然语言理解和信息抽取等方式将其转化为结构化数据。对于杂质较多的数据,可以在数据挖掘时进行数据清洗。对于实时产生的数据可以使用自动获取效率优先的方式来采集数据^[11]。

2. 教育大数据和学习分析

新媒体联盟(New Media Consortium)将学习分析定义为:利用松散耦合的数据收集工具和分析技术,研究分析学习者学习参与、学习表现和学习过程的相关数据,进而对课程、教学和评价进行实时修正^[2]。总之,学习分析运用多种方法采集、存储和分析学习者数据,如移动终端的数据和现场智能数据等,再使用多种技术来分析处理这些数据,最终应用于教育者和学习者,产生评估、预测和干预。基于教育大数据的学习分析过程如图4所示。

教育大数据学习的具体应用体现在评估、预测和干预。评估是指基本统计分析及其可视化、发现问题学生、学生社交网络分析与应用,预测是指学生分类、学生模型的构建、预测学生成绩,干预是指对教师的教学方法提供改进意见。教育大数据在学习分析中还有其他应

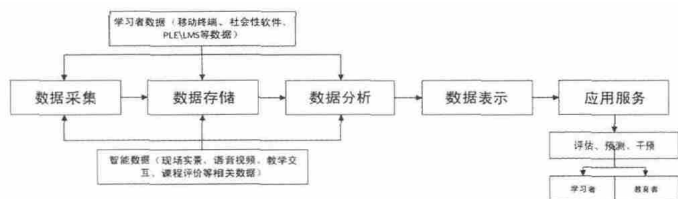


图4 基于教育大数据的学习分析过程

用,如表1所示^[12]。除此之外,大数据在学习分析中的应用还包括学生分组与协作、社交网络分析、开发概念图、课件制作、规划和调度等。

表1 大数据在学习分析中的应用

序号	应用	具体解释
1	个性分析	学生的表现可以通过分析学生与其他学生和老师的互动来预测。
2	数据可视化	随着教育数据的规模不断增长,教育数据报表变得越来越复杂。数据可视化技术使数据直观可视,仅仅需要查看数据视觉报告,可以很容易地识别数据的趋势和关系。
3	即时反馈	学习系统可以针对学生的投入提供智能和即时的反馈,这可以促进学生的互动和表现。
4	课程推送	新课程可以被推荐给有兴趣的学生,通过分析学生活动可以辨别其学习兴趣。

五、教育大数据的进一步挑战

1.教育大数据的技术挑战

教育大数据在实际应用存在很多技术瓶颈,如:在数据的采集时数据挖掘和学习分析的重要环节,在这一环节有很多技术挑战;面对海量教育数据,数据的存储、处理和分析都存在技术考验;另外由于目前没有统一的数据规范,不同系统之间的兼容也是一大问题。从全局考虑,我国的教育大数据系统应遵循顶层设计原则,由教育部对数据格式、数据存储等问题制定统一规范,下级企业、学校按照统一规范去设计自己的系统,这带来了新的技术挑战。

2.教育大数据的安全与隐私挑战

在伦理道德方面,教育大数据面临着数据安全与保护隐私的挑战。教育大数据不仅是一种宝贵的教育收益,而且有关学习者和教育工作者的隐私权,其潜在的安全和隐私甚至是无法用常用办法评估出来,甚至无法通过独立的方法进行评估。如果教育大数据处理不当将导致严重的安全漏洞,甚至是影响区域教育政策的制定。教育数据所有权究竟属于学习者还是运行平台或是双方共享,这是互联网教育和教育大数据发展所面临的必然问题。如何保证所有权归属、如何判定所有权归属,这些仍是需要深虑的问题。

3.教育大数据的价值挑战

教育大数据技术挖掘出教育数据的巨大价值,但

是数据的权属问题不是传统的财产、知识产权等可以涵盖的,数据成为国家间争夺的资源,人口红利、地大物博、经济实力、文化优势等都体现为数据资源储备和数据服务影响力。教育大数据是一种无形资产,国家应当保障教育大数据不外泄并不被恶意使用,而且还要兼顾部分数据向公众开放,发挥其应用的社会价值,做到适当的综合评估来进行价值平衡,这也是前所未有的挑战。

参考文献:

[1]Big Data for Development: Challenges & Opportunities[DB/OL].2012-05-01.http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGloabalPulse-June2012.pdf.

[2]徐鹏,王以宁,刘艳华等.大数据视角分析学习变革——美国《通过教育数据挖掘和学习分析促进教与学》报告解读及启示[J].远程教育杂志,2013(6):11-17.

[3]Barwick H. The “four Vs” of Big Data. Implementing Information Infrastructure Symposium[EB/OL].2012-10-02.http://www.computerworld.com.au/article/396198/iiis_four_vs_big_data/.

[4]IBM.What is big data/[EB/OL].2012-10-02.http://www-01.ibm.com/software/data/big_data/.

[5]大数据[EB/OL].http://zh.wikipedia.org/wiki/大数据,2013-06-28.

[6]杨现民,唐斯斯,李冀红.发展教育大数据:内涵、价值和挑战[J].现代远程教育研究,2016(1):50-61.

[7]余胜泉.突破与转型:数字校园及其智慧化发展趋势[J].中小学管理,2015(5):4-8.

[8]胡弼成,王祖霖.“大数据”对教育的作用、挑战及教育变革趋势——大数据时代教育变革的最新研究进展综述[J].现代大学教育,2015(4):98-104.

[9]王惠中,彭安群.数据挖掘研究现状及发展趋势[J].工矿自动化,2011,37(2):29-32.

[10]维克托·迈尔·舍恩伯格,肯尼思·库克耶.大数据时代[J].中国电力教育,2013(5).

[11]魏顺平.基于大数据的教育决策支持案例分享[EB/OL].2016-01-14.http://learning.sohu.com/20160114/n434551706.shtml.

[12]Sin K, Muthu L. APPLICATION OF BIG DATA IN EDUCATION DATA MINING AND LEARNING ANALYTICS-A LITERATURE REVIEW[J].Ictact Journal on Soft Computing,2015.

(编辑:王晓明)