教育大数据采集机制与关键技术研究

柴唤友[1]，刘三女牙[1,2]，康令云[1]，张雅娴[1]，李卿[2]，刘智[2]
1. 华中师范大学国家数字化学习工程技术研究中心，湖北 武汉 430079；
2. 华中师范大学教育大数据应用技术国家工程实验室，湖北 武汉 430079

摘要
数据采集是实现教育大数据应用价值潜能的基础，因此对于教育大数据建设与应用至关重要。阐述了教育大数据的采集内容、采集方式、采集手段及标准与规范，并结合当前教育大数据建设与应用中的实际问题，分别从平衡数据共享与隐私保护、驱动数据治理与人才创新、创新采集机制与相关技术3个方面，对教育大数据采集研究提出对策与建议。

关键词
教育大数据；数据采集；数据伦理；数据治理

中图分类号：G-4 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020051

Research on the mechanism and key technologies for big data collection in education

CHAI Huanyou[1], LIU Sannyuya[1,2], KANG Lingyun[1], ZHANG Yaxian[1], LI Qing[2], LIU Zhi[2]
1. National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China
2. National Engineering Laboratory for Educational Big Data, Central China Normal University, Wuhan 430079, China

Abstract
The mechanism and its related technologies of data collection are the foundation of realizing the valuable potential of applying big data in education, and therefore are vital for the construction and application of big data in education. The contents, methods, means, standards and specifications of data collection of big data in education were outlined. Combining with the practical problems in the construction and application of big data in education, guidance on research on big data in education was provided, from the perspectives about balancing data sharing and privacy protection, driving data governance and talent acquisition, and improving the mechanisms and related technologies respectively.

Key words
educational big data, data collection, data ethic, data governance

1 引言

中共中央、国务院印发的《中国教育现代化2035》和《加快推进教育现代化实施方案（2018—2022年）》明确指出，加快教育现代化是赶超世界先进教育水平的重要战略部署。为实现这一目标，教育工作者和研究者需要充分运用新理念、新模式、新技术激发教育创新能力，培养适应时代发展的创新型人才[1]。作为现今新兴信息技术发展的重要构成部分，大数据已成为驱动新一轮教育变革和发展的核心力量，强力赋能我国的现代化教育事业。

教育大数据指在所有教育活动过程中产生的以及依据教育需求采集到的，一切用于教育发展并能创造巨大潜在应用价值的数据集合[2]。作为大数据的一个子集，教育大数据特指教育领域的大数据，具有驱动教育决策科学化、学习方式个性化、教学管理人性化和评价体系全面化的价值潜能[3]。在教育大数据的建设与应用过程中，如何对相关数据进行采集、分析和应用是三大核心研究问题[4]。其中，数据采集是基础，决定着教育大数据分析和应用的质量，并最终影响着教育大数据价值潜能的实现程度[5]。然而，当前教育界学者对教育大数据的采集机制和技术尚未达成共识，存在一些有待解决的关键问题，如：教育大数据究竟包含哪些内容？如何采集？涉及哪些关键技术？需要遵守哪些标准与规范？针对上述问题的研究有助于推动大数据在教育领域中的大规模成功应用，并深化我国教育现代化的改革与发展。

2 教育大数据采集内容

教育大数据涉及的数据内容普遍存在场景多样、量化困难、汇聚复杂等特点。具体而言，场景多样是指教育大数据来源于众多与教育或学习相关的场景，如教学活动、科研活动、社交活动等相关场景；量化困难源于教育场景的多样性、人的不确定性以及人、机、物之间交互的复杂性等因素；汇聚复杂是因为教育大数据具有来源多样化、结构异质化和内容复杂化等特点[6]。

由于上述特点的存在，教育大数据的采集内容框架呈现出基于不同分类标准的多元化特点[7]，目前较为常见的是依据数据采集场景来区分不同类别的教育大数据。根据数据采集场景的差别，教育大数据一般可被分为教育管理数据、教育教学数据、科学研究数据、室外学习数据、校园生活数据、成长经历数据6个类别。每种类别的教育大数据分别涉及不同的数据主体、数据来源和数据内容，见**表1**。

教育管理数据来源于各种不同类型的教育管理活动，即管理者通过组织协调教育队伍并借助教育内部各种有利条件，高效达成教育管理目标的活动过程。该过程通常涉及学生、教师、学校和其他相关机构等主体，可产生学校管理信息（如特等教师数量、教职工学历信息等）、行政管理信息（如教育行政部门设置的大学专业门类信息）、教育统计信息（如班级规模、性别分布信息等）等。

教育教学数据是指师生在（线上或线下）教和学的活动过程中产生的数据，通常涉及学生、教师、教育资源和教育设备等主体。通过学生、教师与教育资源、教育设备间的交互，教学场景可以产生学生和教师的行为和状态信息（如学生的学习策略、学习动机，教师的课前准备度和教学策略等）、教育资源信息（如PPT课件、微课、软件等）、教育设备运行信息（如设备损耗、故障信息等）等。

科学研究数据是指学生（特别是研究

表1 教育大数据的数据内容采集框架

场景	来源	主体	内容
教育管理	教育管理活动	学生、教师、学校和其他相关机构等	学校管理信息、行政管理信息、教育统计信息等
教育教学	教学活动	学生、教师、教育资源和教育设备等	学生和教师的行为和状态信息、教育资源信息、教育设备运行信息等
科学研究	科学研究活动	学生、教师、论文、科研设备和科研材料等	科研设备操作信息、论文发表信息、科研材料与消耗信息、导师指导信息等
室外学习	教室外的教育活动	学生、客观环境或对象等	学习者与客观环境或对象之间的交互信息，如感知内容、互动记录、活动体验等
校园生活	校园非学习活动	学生、网络、健身设备、刷卡机、社交工具等	餐饮消费信息、上机上网信息、健身洗浴信息、社会交往活动信息等
成长经历	个体成长活动	学生、家长、教师、社会环境等	同个人成长经历有关的环境信息，如家庭经历、校园经历和社会环境等

生）在开展科学研究活动时产生的一系列数据内容，通常涉及学生、教师、论文、科研设备和科研材料等主体。相应地，科研活动中可以产生科研设备操作信息（如错误操作类型及数目等）、论文发表信息（如实际贡献、发表时间、发表期刊名称及影响因子等）、科研材料与消耗信息（如化学或生物试剂等）、导师指导信息（如论文修改意见等）等。

室外学习数据来源于学习者在教室外参与的一系列教育活动，如在动植物园中的生物习性研究、参观各种场馆、野外探险等。该活动通常由学习者主动发起，并由学习者自身进行调控和负责，涉及学习者以及与其交互的客观环境或对象。在室外学习场景中，研究者通常可以采集学习者与客观环境或对象之间的交互信息，如感知内容、互动记录、活动体验等。

校园生活数据是指学习者在校园非学习活动（如餐饮、上网、健身、社交等）中产生的各类数据，通常涉及学生、网络、健身设备、刷卡机、社交工具等主体。通过参与上述非学习活动，学习者可以产生餐饮消费信息（如饮食类型及价格、就餐时间等）、上机上网信息（如上网时间、网络活动类型等）、健身洗浴信息（如健身和洗浴的时间和频率等）、社会交往活动信息（如好友数量、联系频率等）等。

成长经历数据是指伴随学生成长（从出生到现阶段）而产生的各种环境（包括家庭环境、社会环境、校园及班级环境）数据，涉及学生、家长、教师、社会环境等诸多主体。在成长过程中，学习者可以产生一系列同个人成长经历有关的环境信息，如家庭经历（如家长文化素养、职业特点）、校园经历（如学校规章制度、教师特点）和社会环境（如社会风气、社会期望）等。

总而言之，上述6个类别的数据相辅相成、相互促进，共同构成了教育大数据全面且丰富的采集内容。

3 教育大数据采集方式

由于数据来源多样（如国家、区域、学校、班级和个体等不同来源）且形式不一（如结构化、半结构化以及非结构化数据共存），教育大数据的采集方式也相应具有多样化特点。总体而言，教育大数据的采集方式主要包括集中式采集、伴随式采集和周期性采集3种。其中，集中式采集侧重

于数据采集的统一性，伴随式采集侧重于数据采集的实时性，而周期性采集侧重于数据采集的连续性。

（1）集中式采集

集中式采集是指教育管理机构借助教育管理活动而统一开展的数据获取方式。例如，对学生在家庭情况、校园生活和学习环境3方面的成长经历数据进行统一采集。在教育大数据视域下，不同机构、不同单位采集的不同层次、不同类型的信息不再相互割裂，而是可以得到整合和管理，因此有助于研究者获得针对特定分析对象的全面且丰富的理解[8]。集中式采集的教育大数据主要以结构化和结果性数据为主，具有覆盖面广、标准化程度高、关注层面相对宏观的基本特点。其中，覆盖面广是指相关数据内容涵盖广泛，包括学生个体层次、家庭层次和学校层次等多方面的内容；标准化程度高是指相关数据内容一般具有统一的采集标准，易于分析和处理；关注层面宏观是指相关数据内容通常指向特定分析单元的教育发展整体状况，具有宏观性。

（2）伴随式采集

伴随式采集是指借助教育信息管理系统（如特定课程管理系统）应用在管理过程中实时产生教育基础数据而开展的数据获取方式。例如，学习（或课程）类系统会全程记录学习者的在线行为数据，如学习时长、鼠标点击次数及频率、论坛读帖和发帖的次数和时间、作业和考试次数等；管理类系统会有效记载学校的资产和人事信息，如学籍管理、教学设备、教务科研、财务人事以及校园安全与生活等数据。在教育大数据视域下，智能化数据采集除了关注学生的在线表现，还重视学生线下的学习、练习或实践等过程性数据，例如，利用可穿戴设备可自然真实地抽取学生实践练习中的生理表征和行为习惯，而无须过多的人工干预。通过全域式网络架构与学生随身携带的新型便携式智能传感器，新型数据采集系统可实现伴随式采集学生学习的全过程数据（除了学生的常规学习过程信息，还包括个人提交的作品信息、社会实践相关信息等）的目标[9]。伴随式采集的教育大数据以过程性数据为主，普遍具有密集性、动态性、复杂性、全面性等特点。其中，密集性是指相关数据内容产生的速度和数量级别均远远高于常规总结式采集方式，动态性是指相关数据内容一直处于持续、动态的定位与追踪之中，复杂性是指相关数据内容通常类型多样、结构异质，全面性是指相关数据内容能够完整记录所有与学生学习相关的信息。

（3）周期性采集

周期性采集是指利用特定教育管理软件对学习环境、教学过程、教育质量等进行周期性监控和测量的数据获取方式。例如，学生在入校之初会被统一要求登记身心健康信息、家庭基本信息；学校会定期更新全体教职工基础信息、教育设备运行信息、行政管理信息、人事资产信息和学校管理信息等。在教育大数据视域下，个体、专业、学校等不同层次不同类型的数据内容皆可被纳入周期性采集的对象范围内[10]。周期性采集的教育大数据在数据类型上同时包含过程性和结果性数据，在分析层次上以整体性层次（较少关注学生个体的教育发展水平）为主，具有连续性、规范性和充分性的基本特点。其中，连续性是指相关数据内容应多次采集，以确保客观评估；规范性是指相关数据内容的采集应符合特定情况下的技术规范，以保证后续数据的一致化分析和处理；充分性是指相关数据内容的采集可从多个路径和渠道获得，以保证数据的多样性。

4 教育大数据采集手段

构建多样化的数据采集手段有助于扩展教育大数据采集的广度和深度。目前，教育大数据的采集手段主要有平台采集（针对在线人机交互时产生的学习过程数据）、视频录制（针对线下教学环境中学习者交互的视频音频数据、校园安全数据等）、图像识别（针对学习过程中的图像类数据）、物联感知（针对校园环境下产生的学习者的学习生活数据及个人生理数据）等。

（1）平台采集

平台采集是指借助各种与教育或学习相关的移动或桌面应用平台，获取教育数据内容的方法或手段。随着教育信息技术的不断发展，越来越多的移动或桌面应用平台被应用在教育领域中，利用这些平台进行教育数据采集也随之成为可能。目前，基于平台采集的教育数据采集技术主要涵盖平台自动记录技术、日志搜索分析技术、移动App技术和网络爬虫采集技术等。

平台自动记录技术是指基于在线学习与管理平台内的嵌入式数据采集系统，自动记录并获取学习者的在线学习行为数据（如平台登录次数、驻留时间等）的技术。由于在线学习与管理平台使用人数的迅猛增长，基于该技术开展教育数据挖掘已成为当前教育大数据研究领域的一大热点[11]。例如，来自斯坦福大学的Bihani A等人[12]在进行在线学习成绩预测时，从Piazza在线论坛中挖掘了学生登录总天数、查看帖子数、提出问题数、回答问题数等量化数据。

日志搜索分析技术是指针对教育或学习应用平台中发生的所有事件（如学习者访问记录、运维工作记录等）进行记录并分析的技术。基于数据驱动或理论驱动方法，教育研究者可以利用该技术发现学习者的在线表现特点及其规律[13]。例如，悉尼大学的McBroom J等人[14]在考察周练习任务中学习者行为与其期末考试成绩之间的联系时，对来自该学校初级计算机科学数据结构课程的494名学生的程序评估提交和测试应用（programming assessment submission and testing application，PASTA）平台的练习日志数据进行了长期行为分析。

移动App技术是指利用教育App采集学生学习（过程性或结果性）数据的技术。典型的相关技术工具有国外的化学实验模拟类ChemCrafter、数学教程视频类Virtual Nerd mobile Math、新闻阅读类Newsela、教育互动类eduClipper等，以及国内的小猿搜题、猿辅导、扇贝单词等不同类型的教育App。移动App可被用于辅助传统教育，因此基于该类应用采集的数据可作为教育大数据分析内容的强力补充[15]。

网络爬虫采集技术一般是指依据一定准则，借助特定程序或者脚本自动捕获网页信息的技术。目前应用较多的爬虫框架采集方法包括基于Hadoop平台开发的Chukwa、基于Facebook的Scribe、基于LinkedIn的Kafka以及基于Cloudera的Flume等。在教育领域中，该技术可被用于捕获并分析教育应用平台中的文本信息，如学生在异步论坛中发布的帖子、校园贴吧中的舆情信息等[16]。

（2）视频录制

视频录制是指对源于计算机硬件终端和计算机视窗环境内的视频内容加以录制的方法或手段。典型的录制模式包括捕捉摄像头、摄像机、数码相机、硬盘录像机等硬件视频，以及可录制计算机视窗内容的游戏视频和电影视频等。目前，视频录制手段涉及的教育数据采集技术主要有视频监控技术和视频录播技术等。

视频监控技术是指借助视频监控设备检测、监视特定物理区域，实时展示、记录现场图像，或支持搜索和展示历史图像的技术。在教育领域中，该技术可被用于监控校园环境，提供关于校园安全的数据信息。例如，一些企业开发的校园网格化监控系统可实现实时监控校园环境的目标。

视频录播技术一般是指可在教师现场授课的同时，自动产生课堂教学实况录像，并完整录制教师授课全过程的技术[17]。该技术可在无须专人操作控制的条件下录制整个教学过程，因此极大地方便了视频课程资源的制作和记录。例如，国内一些公司开发的便携录播视频工具能够实现基于无线摄像机的全场景拍摄目标。

（3）图像识别

图像识别是指对特定物理图像进行对象检测，以识别各种不同模式的目标和对象的技术[18]。作为人工智能的重要研究领域之一，图像识别在教育领域有广泛的应用，如网评网阅技术、点阵数码笔技术和拍照搜题技术等。

网评网阅技术是指以电子扫描技术和计算机网络技术为基础，将多年来人工阅卷积累的丰富经验与现代信息技术相整合的一种先进、科学、高效的自动化评分方式。相比传统人工评阅方法，网评网阅技术能够极大地降低广大教师的工作负担，并支持更为精准科学的教育教学评价。例如美国教育考试服务中心开发的TextEvaluator以及科大讯飞开发的智能阅卷技术，后者已于2017年在襄阳中考中率先使用，目前已被广泛应用于上海、青岛等城市的中考阅卷。

点阵数码笔技术是指一种通过数码笔前端的高速摄像头实时捕捉笔尖在印刷了一层隐形点阵图案的纸张上的运动轨迹，同时压力传感器将压力数据传回数据处理器，然后将相关信息通过蓝牙或者USB向外传输的新型书写技术。不同于传统纸笔书写，该技术能够记录纸张类型、笔尖坐标、笔尖压力等信息，并支持本地存储及远程传播功能。根据应用类型的不同，点阵数码笔技术可被划分为：支持个人笔记作业管理的DoTnore数码笔，其书写内容可被同步保存到电脑、平板和手机上；支持教学课堂交互的Symphony数码笔，其特点是可以多人同时使用，而且结果可被同步到教师电脑上；支持远程教学会议的Tnote数码笔，该技术能够突破基于视频、语音、键盘的传统交互方法，打破时间空间限制，从而提供纸面书写的交流方式。

拍照搜题技术是指通过拍照、语音等方式帮助用户快速找到疑难问题的答案的技术。该技术融合了扫描、识别、检索等技术手段和海量题库大数据，有助于学生提升学习效率，并实时采集学生作业练习数据。目前国内作业帮、小猿搜题、学霸君、网易有道词典等教育产品均可实现该功能，其中作业帮还根据学生、家长、老师三大群体进行了拍照搜题功能的细化区分。

（4）物联感知

物联感知是指基于现有和正在发展中的可互操作的信息通信技术，通过互连（物理和虚拟）事物来实现测评特定对象的一种全球性基础设施或技术增强型解决方案[19]。由于物联网具有无处不在的特性，学校和学术机构正在寻求将物联感知纳入教育活动，以解决教育部门的各种模式、目标、主题和观念问题，最终使学生、教师和整个教育系统受益[20]。现有教育领域内的物联感知采集手段主要包括物联网感知技术、可穿戴技术、非接触式感知技术、校园一卡通技术和多模态融合技术等。

物联网感知技术一般是指被用于物联网底层（即物理世界中发生的具体物理事件）感知信息的技术，在教育领域主要指多媒体信息采集技术。通过多媒体信息采

集技术，多媒体计算机系统中的主机能够随时采集各种多媒体外接设备的状态（视频或音频）信息，从而为相关（教学）设备的精确调整提供信息支撑。例如，Cook C等人[21]通过使用自动语音识别设备对来自两个州7所学校14名教师的132堂课进行音频录制，能够实现课堂效果评估和学生成绩预测的目标。前谷歌工程师Ventilla M创办的Altschool中的Alt Video系统通过各种传感器、摄像头和麦克风综合采集学生课堂行为数据，有助于改进教学过程和教学系统。

可穿戴技术是指利用可直接穿戴在用户身上或嵌入用户衣饰或配件内的设备（如智能手环、谷歌眼镜）开展数据采集的技术。通过可穿戴设备，学习者个体的生理状态及学习行为数据能够得到实时的记录和存储。例如，在学生的语音指令下，集成了麦克风、耳机以及微型摄像头的谷歌眼镜可以开展拍照摄像，从而实现及时保存教师板书内容的功能。

非接触式感知技术是以光电、电磁等技术为依托，在不接触被测对象的情形下，获取其基本信息的科学技术或手段。在教育领域中，该技术强调在不产生干扰的情况下采集学习者的生理与行为数据，有助于实现针对学习者信息（认知、行为及情感）的自动化和非侵扰式采集。例如，为了分析学生的注意力，Millsa C等人[22]、Stewart A等人[23]借助Logitech C270摄像头和Tobii TX 300眼动仪，对观看《红气球》电影的60名参与者进行了生理和行为数据采集。

校园一卡通技术是指基于将智能卡物联网技术、计算机网络的数字化理念融合于校园日常管理而开展的统一管理身份认证、人事、学工等信息的应用解决方案。该技术能够统一记录并采集学习者的金融消费、图书借阅和考勤等校园生活信息，是构建“数字化校园”和“智慧校园”的重要组成部分[24]。例如，华东师范大学率先利用学生的一卡通餐饮消费数据，对经济困难的学生提供情感抚慰和助学金支持，这体现了基于物联感知数据的人性化关怀。

多模态融合技术一般是指联合图像、文本、语音等多模态信息进行目标检测或识别的技术。在教育领域中，该技术可被用于分析与学习者相关的多维度数据，以识别和解释内在学习过程、特征和变化，最终助力学习者学习体验和学习绩效的提升。其中，情感识别技术被认为是多模态融合技术在教育领域中的典型应用。如何基于教学视频中的视频、音频和文本等多样化信息判断学习者学习过程中的情感状态，是教育领域内相关学者正在关注且亟须解决的关键问题[25]。例如，Wampfler R等人[26]使用触控笔、数位板、生物传感器（Empatica E4、Shimmer GSR、GoPro HERO3）等产生的多模态数据，分析88位参与者在解决数学任务时的情感状态；Vail A等人[27]在分析学生参与Java编程课程时的情感反应时，综合采集了学生的手势、姿势、面部表情变化和皮肤电活动等信息。

5 教育大数据采集标准与规范

出于教育科学研究和大数据研究的学术目标和伦理要求，许多研究机构或组织针对教育大数据的不同方面制定了一系列基本标准与规范，如描述学习者信息的朋友的朋友（friend of a friend，FOAF）规范、面向教学内容的学习目标元数据（learning object metadata，LOM）标准、全国信息技术标准化技术委员会教育技术分技术委员会（China E-Learning Technology Standardization Committee，CELTSC）构建的教学评价标准等。但总

体而言，这些标准与规范大多针对教育大数据的不同主体、不同层次和不同教育过程，缺少针对教育大数据采集方面的标准与规范。

依据教学活动的不同构成部分，可将教育大数据采集标准划分为下述5类：教学主体类、教学评测类、教学资源类、教学管理类和教学过程类。教学主体类标准是指针对学生、家长、教师、教研员和教学管理者等制定的采集标准，包括伦理（即针对学生隐私保护的规范）和权益（如学生的知情同意权、自由参与权）方面的规范等。教学评测类标准是指针对教学目标、知识能力、信息素养、教学能力等的评测而制定的采集标准，如术语方面的规范（如对评测指标的命名方式及其特点的定义方式等）。教育资源类标准是指为统一描述、封装与重组不同形式、不同粒度、不同格式的教学资源而制定的采集标准，如格式（如资源数据的记录方式）方面的规范。教学管理类标准是指针对指向管理需求的一系列基本信息和管理数据（如学生教师数据、学校数据和基础设施数据等）而制定的采集标准，如过程（即依据管理活动类型而确定的数据采集流程）方面的规范。教学过程类标准是指为描述教学过程中教学主体与教学内容（如课程、资源等）、教学环境（如传统教室、户外学习环境）及其他教学活动参与者之间的交互经历而制定的采集标准，如支撑技术（如采集工具类型及其使用方式）方面的规范。5种类别的采集标准通过有机结合，共同构成了教育大数据采集标准与规范的复杂内涵。

6 挑战与展望

数据采集是教育大数据建设与应用的基础和关键，针对其机制与技术的研究不仅关系着教育大数据采集的数量与质量，还影响着后续的分析及应用过程，因此对教育大数据发挥其教育潜能至关重要。然而目前，教育大数据采集机制与技术研究仍存在许多问题，在教育数据伦理、教育数据治理和教育数据采集规范等方面面临诸多挑战，因此需要研究者加以关注并解决。

（1）数据共享与隐私保护的平衡挑战

作为一种特殊的大数据资源，教育大数据需要适度向社会和公民开放，但由此会产生隐私泄露、数据滥用等潜在风险，因此教育大数据的隐私保护和安全问题必须得到重视和解决[28]。首先，应明确采集用户或研究目的。其次，有必要采取特定措施，以确保数据主体对数据采集享有知情同意权，防止侵犯个人隐私。应开发更先进、安全系数更高的技术手段来保障教育数据安全，以避免教育隐私数据泄露和数据滥用等问题。同时，应规范数据开放与共享的流程，以防止操作不当带来的数据泄露问题。最后，应加快制定“教育大数据安全管理办法”，从制度层面保障数据主体的隐私安全。

（2）数据治理与人才创新的驱动挑战

在教育信息化2.0时代，各种教育场景无时无刻不在产生海量的、多来源的、多种结构类型的数据，如何协同教育相关部门开展高效的数据治理，是教育大数据建设与应用过程中必须面临和解决的一大核心问题[29]。首先，应充分发挥各类教育相关主体（包括政府、学校、企业等）在建设与应用教育大数据上的独特优势，驱动教育大数据采集的来源多样性和内容全面性；其次，应尽快确立教育大数据治理的相关方法和机制，推动教育大数据治理的规范化、制度化和常态化；最后，应大力培养教育大数据治理人才，鼓励支持技术创新，助力教育大数据治理的有效性和高效性。

（3）采集机制与相关技术的创新挑战

借鉴（广义）大数据领域的相关研究成果，教育大数据已在数据采集方面积累了一定的知识与经验。但和其他领域的相关研究类似，教育大数据采集研究也存在一些机制与技术上的“通病”，有待未来研究者加以关注和解决[4,28]。首先，针对教育领域内的密集型数据，如何保证数据采集的可靠性、如何确保采集的数据质量、如何避免出现重复数据等，都需要通过更新现有的采集机制和相关技术加以解决。其次，现有的教育大数据采集技术虽然在一定程度上能够解决传统数据采集方式的一些缺陷，但也会带来一些新的问题，如测量精度不高、受环境影响较大等。因此，未来有必要探索更为稳定且精准的新型采集技术。再次，考虑到教育场景的多样性和复杂性，在针对特定教育场景开展数据采集时，有必要选择合适的采集机制与技术，“因地制宜”，以确保数据采集的有效性和可靠性。最后，为了使教育大数据能够发挥最大效力，未来还可考虑采集并融合学习者的“个体性数据”，以支持和推动个性化学习服务，达成“因材施教”的核心教育目标。

7 结束语

随着现代信息技术的迅猛发展，教育大数据正以全新方式驱动教育决策、学习方式、教学管理和评价体系的智能化和信息化。其中，数据采集作为教育大数据建设与应用中的基础性过程，是实现教育大数据价值潜能的关键。本文概述了教育大数据的采集内容、采集方式、采集手段及采集基本规范。在此基础上，进一步探讨了教育大数据采集研究面临的问题和挑战，并指出了未来的潜在研究方向。未来要继续深入探讨教育大数据建设与应用中的数据采集问题，强力推动大数据赋能教育事业，从而确保实现教育现代化的最终目标。

参考文献:

[1] 沈阳, 田浩, 曾海军. 大数据时代的教育：若干认识与思考——访中国科学院院士梅宏教授[J]. 电化教育研究, 2020, 41(7): 5–10.
SHEN Y, TIAN H, ZENG H J. Education in the age of big data: some insights and reflections: an interview with professor MEI Hong, academician of Chinese Academy of Sciences[J]. e–Education Research, 2020, 41(7): 5–10.

[2] 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.
YANG X M, TANG S S, LI J H. The definition, potential value and challenges of big data in education[J]. Modern Distance Education Research, 2016(1): 50–61.

[3] 程玉, 胡凡刚, 吴运明. 教育大数据价值体现、问题反思与发展路径[J]. 软件导刊, 2020, 19(5): 281–284.
CHENG Y, HU F G, WU Y M. Reflections on the values, problems and development path of big data on education[J]. Software Guide, 2020, 19(5): 281–284.

[4] DANIEL B. Big data and analytics in higher education: opportunities and challenges[J]. British Journal of Educational Technology, 2015, 46(5): 904–920.

[5] LI Y, ZHAI X. Review and prospect of modern education using big data[J]. Procedia Computer Science, 2018, 129: 341–347.

[6] 刘三女牙, 杨宗凯, 李卿. 计算教育学：内涵与进路[J]. 教育研究, 2020, 41(3): 152–159.
LIU S N Y, YANG Z K, LI Q. Computational education: connotations and approaches[J]. Educational Research, 2020, 41(3): 152–159.

[7] VATSALA V, JADHAV R R S. A review of big data analytics in sector of higher education[J]. International Journal of Engineering Research and Applications, 2017, 7(6): 25–32.

[8] 何普亮, 张战胜. 大数据时代的教育数据挖掘: 方法、工具与应用[J]. 中国教育技术装备, 2019(23): 7–10.
HE P L, ZHANG Z S. Educational data mining in big data era: methods, tools and applications[J]. China Educational Technology & Equipment, 2019(23): 7–10.

[9] 王小根, 单必英. 基于动态学习数据流的“伴随式评价”框架设计[J]. 电化教育研究, 2020, 41(2): 60–67.
WANG X G, SHAN B Y. Design of “Accompanying Evaluation” framework based on dynamic learning data flow[J]. e–Education Research, 2020, 41(2): 60–67.

[10] AMAXILATIS D, AKRIVOPOULOS O, MYLONAS G, et al. An IoT–based solution for monitoring a fleet of educational buildings focusing on energy efficiency[J]. Sensors, 2017, 17(10): 2296.

[11] 徐岸峰, 杨仲基, 王波. 基于智慧教育平台大数据的学生学习行为分析[J]. 高教学刊, 2020(22): 16–19.
XU A F, YANG Z J, WANG B. Analysis of students’ learning behavior based on big data of intelligence education platform[J]. Journal of Higher Education, 2020(22): 16–19.

[12] BIHANI A, PAEPCKE A. QuanTyler: apportioning credit for student forum participation[C]// The 11th International Conference on Educational Data Mining. New York: ACM Press, 2018: 106–115.

[13] JANSEN B J. Search log analysis: what it is, what’s been done, how to do it[J]. Library & Information Science Research, 2006, 28(3): 407–432.

[14] JEFFRIES B, KOPRINSKA I, MCBROOM J, et al. Mining behaviours of students in AutoGrading submission system logs[M]. Heidelberg: Springer, 2016.

[15] CHARLAND A, LEROUX B. Mobile application development: web vs. native[J]. ACM Queue, 2011, 54(5): 49–53.

[16] LI C, MA J. Research on online education teacher evaluation model based on opinion mining[C]// The 2012 National Conference on Information Technology and Computer Science. Paris: Atlantis Press, 2012: 1061–1064.

[17] ELLIOTT C, NEAL D. Evaluating the use of lecture capture using a revealed preference approach[J]. Active Learning in Higher Education, 2016, 17(2): 153–167.

[18] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770–778.

[19] PEI X L, WANG X, WANG Y F, et al. Internet of things based education: definition, benefits, and challenges.[J]. Applied Mechanics and Materials, 2013, 411: 2947–2951.

[20] KASSAB M, DEFRANCO J, LAPLANTE P. A systematic literature review on Internet of things in education: benefits and challenges[J]. Journal of Computer Assisted Learning, 2020, 36(2): 115–127.

[21] COOK C, OLNEY A M, KELLY S, et al. An open vocabulary approach for estimating teacher use of authentic questions in classroom discourse[C]// The 11th International Conference on Educational Data Mining. New York: ACM Press, 2018: 116–126.

[22] MILLSA C, BIXLERA R, WANG X, et al. Automatic gaze–based detection of mind wandering during narrative film comprehension[C]// The 9th International Conference on Educational Data Mining. New York: ACM Press, 2016: 30–37.

[23] STEWART A, BOSCH N, MELLO S K D. Generalizability of face–based mind wandering detection across task contexts[C]// The 10th International Conference on Educational Data Mining. New York: ACM Press, 2017: 88–95.

[24] 鲁鸣鸣, 张丹, 王建新. 基于校园一卡通数据

好友发现及应用[J]. 大数据, 2017, 3(2): 78–91.
LU M M, ZHANG D, WANG J X. Smart-card based campus friend mining and its applications[J]. Big Data Research, 2017, 3(2): 78–91.

[25] DOLIANITI F S, IAKOVAKIS D, DIAS S B, et al. Sentiment analysis techniques and applications in education: a survey[C]// The International Conference on Technology and Innovation in Learning, Teaching and Education. Heidelberg: Springer, 2019: 412–427.

[26] WAMPFLER R, KLINGLER S, SOLENTHALER B, et al. Affective state prediction in a mobile setting using wearable biometric sensors and stylus[C]// The 12th International Conference on Educational Data Mining. New York: ACM Press, 2019: 198–207.

[27] VAIL A, WIGGINS J, GRAFSGAARD J, et al. The affective impact of tutor questions: predicting frustration and engagement[C]// The 9th International Conference on Educational Data Mining. New York: ACM Press, 2016: 247–254.

[28] 刘三女牙, 杨宗凯, 李卿. 教育数据伦理: 大数据时代教育的新挑战[J]. 教育研究, 2017, 38(4): 15–20.
LIU S N Y, YANG Z K, LI Q. Education data ethic: the new challenge of education in big data era[J]. Educational Research, 2017, 38(4): 15–20.

[29] DANIEL B K. Big data and data science: a critical review of issues for educational research[J]. British Journal of Educational Technology, 2019, 50(1): 101–113.

作者简介

柴唤友（1990– ），男，华中师范大学国家数字化学习工程技术研究中心博士生，主要研究方向为教育数据挖掘、学习分析、教学心理与行为分析。

刘三女牙（1973– ），男，博士，华中师范大学教授、人工智能教育学部副部长，国家数字化学习工程技术研究中心、教育大数据应用技术国家工程实验室常务副主任。教育部新世纪优秀人才支持计划和湖北省新世纪高层次人才工程入选者，湖北省政府专项津贴专家，主要研究方向为教育大数据、智能教育及教育技术，目前担任教育部高等学校教学信息化与教学方法创新指导委员会教育技术专业教学指导分委员会委员、中国教育发展战略学会教育大数据专业委员会副理事长、全国信息技术标准化技术委员会教育技术分技术委员会委员、《大数据》期刊编委等。先后主持国家重点研发计划、国家科技支撑计划、国家自然科学基金、国家社会科学基金等项目20余项，荣获高等学校科学研究优秀成果奖（科学技术）科学技术进步奖一等奖1项，湖北省科技进步奖一等奖2项、二等奖1项，高等教育国家教学成果奖二等奖1项，湖北省高等学校教学成果奖一等奖1项。

康令云（1995– ），女，华中师范大学国家数字化学习工程技术研究中心博士生，主要研究方向为学习分析、在线协作学习。

张雅娴（1996– ），女，华中师范大学国家数字化学习工程技术研究中心硕士生，主要研究方向为大数据分析、图像识别。

李卿（1982– ），女，博士，华中师范大学教育大数据应用技术国家工程实验室副教授，主要研究方向为教育科学战略、教育大数据与感知计算。

刘智（1986– ），男，博士，华中师范大学教育大数据应用技术国家工程实验室副教授，主要研究方向为教育数据挖掘、情感计算与学习行为分析。

收稿日期： 2020-09-08

通信作者： 李卿，viven_a@mail.ccnu.edu.cn

基金项目： 国家自然科学基金资助项目（No.61807012）；中央高校基本科研业务费专项资金资助项目（No.CCNU20QN027，No.CCNU20TS032）

Foundation Items: The National Natural Science Foundation of China (No.61807012), The Project of Special Funds for Basic Scientific Research Operating Expenses of Central Universities (No.CCNU20QN027, No.CCNU20TS032)