浙江科技学院学报，第30卷第6期，2018年12月
Journal of Zhejiang University of Science and Technology
Vol. 30 No. 6, Dec. 2018

doi：10.3969/j.issn.1671-8798.2018.06.009

教育大数据智能分析平台研究与实践

任东晓[1,2]，王中华[3]

（1.浙江科技学院 曙光大数据学院，杭州 310023；2.电子科技大学 计算机科学与工程学院，成都 611731；
3.北京中电普华信息技术有限公司，北京 100085）

摘　要：中国教育资源虽丰富但数据分散、数据收集与分析手段落后，为此，研究了教育大数据智能分析平台及关键技术。提出多源异构教育大数据的集成和融合方案，搭建教育大数据智能分析平台以打破教育信息孤岛，实现信息共享。研究结果可为后续教育质量综合分析、教育质量预警和教育决策支持等提供参考，从而推进教育现代化发展。

关键词：教育大数据；数据融合；机器学习

中图分类号：TP311.13　　**文献标志码**：A　　**文章编号**：1671-8798(2018)06-0501-05

Research and practice of intelligent analysis platform of educational big data

REN Dongxiao[1,2], WANG Zhonghua[3]

(1. School of Sugon Big Data Science, Zhejiang University of Science and Technology, Hangzhou 310023, Zhejiang, China; 2. School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, Sichuan, China; 3. Beijing China-Power Information Technology Co., Ltd., Beijing 100085, China)

Abstract: Although China's educational resources are abundant, there are problems such as data dispersion, backwardness of data collecting and analytical methods. In view of these problems, the educational big data were studied from the intelligent analysis platform and key technologies. The integration and fusion schemes were proposed in terms of multi-source heterogeneous educational big data and an intelligent analysis platform of educational big data was established, which broke the island of educational information and realized information sharing. The results could provide a reference for follow-up comprehensive analysis of educational quality, educational quality warning and educational decision support so as to promote modernization of education.

Keywords: educational big data; data fusion; machine learning

收稿日期：2018-09-29

基金项目：浙江省教育厅一般科研项目(Y201839622)

通信作者：任东晓(1982—　)，女，河南省南阳人，高级工程师，博士，主要从事大数据分析和处理研究。E-mail：rendx29@163.com。

大数据正在实现人类工作、生活与思维的大变革，其威力也强烈冲击着整个教育系统[1-2]。中国教育科研网、现代远程教育、校校通、班班通等工程的实施，“泛在学习”“移动学习”“智慧校园”“微课”“慕课”“翻转课堂”“信息化可穿戴设备”等应用[3-7]的普及，在促进中国教育信息化进程的同时，产生了大量类型多和应用价值高的教育大数据。教育大数据是指在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[8]。换言之，教育大数据是由教育者和受教育者在教学活动和教学管理过程中所产生的有关教学行为和学习行为的大量数据，具有广泛的应用价值。中国政府高度重视教育大数据及研究应用，将教育大数据上升到国家战略层面，并提出“探索发挥大数据对变革教育方式、促进教育公平、提升教育质量的支撑作用”。在《教育信息化“十三五”规划》[9]中，教育大数据在学习空间应用及教育管理平台建设中的重要作用被多次强调。中国教育数据丰富，2016年全国共有学校51.2万所，各级各类学生近3.2亿人，专任教师共计1 578万人。其中，中国高等教育在全球高等教育所占比例高达20%，在学规模有3 699万人。庞大的教育基数随之产生了巨大的教育数据和伴生数据，即教育大数据，构成了国家的重要核心数据之一[10]。

目前的教育系统信息资源和实体资源被各部门、主体之间的边界和壁垒所分割，资源的组织是零散的，信息空间与物理空间分离，学校与家庭、社会不易协同。教育系统的零散分布使得教学与学习活动的灵活性受到限制，以致在一定程度上阻碍了教育的发展。并且，教育大数据还存在数据分散、数据收集和分析手段落后等问题。从横向来看，经费监管、学生在学和就业、科研、继续教育、学生资助、留学和回国等数据分属于不同的单位管理；从纵向的行政区划上看，各级地方政府的教育数据也多为独立王国[11-12]。教育信息存在诸多孤岛，未能实现有效融合和数据共享。教育大数据类型繁多，包括结构化、半结构化和非结构化数据，不同类型数据的集成质量受到集成准确性差并且冗余度高的简单数据集成系统的影响，实现不同类型数据的有效集成和共享是非常重要的。因此，切实有效的数据集成和融合方案，可以去掉冗余和错误数据，提高数据质量，为准确的数据分析和挖掘奠定基础[13-15]。

综上所述，教育大数据是中国基础性的战略资源之一。运用大数据的思想和方法对教育数据进行深度分析和挖掘，找到教育现象、教育内容及教育规律之间的关联性，以符合教育事业发展的内在逻辑性，是时代发展的迫切要求。因此，本文针对教育大数据多源异构等特点，主要研究教育大数据的集成融合和智能分析平台的建设，并给出具体实践，为教育大数据的深度应用提供参考。

1 教育大数据智能分析平台的设计

1.1 教育大数据概述

教育大数据主要产生于教学活动和教育管理过程，收集的是整个教育教学和管理过程中静态和动态的所有数据，既包括教务管理、图书管理、学生管理、财务管理、科研管理、后勤服务等系统的结构化数据，又包括课堂教学、教室和实验室使用、社会实践、宿舍能耗、校园生活、安全、网络课件、讲课音视频、图片、交互记录、学习痕迹等非结构化和半结构化数据。除此之外，教育大数据还包括家庭、社区、博物馆、图书馆等非正式环境下学习活动产生的数据，以及智能设备、社交媒体等“伴随式收集”的教育动态和即时数据。教育大数据来源分散、类型繁多、质量良莠不齐和标准不一致，不同数据源间可能存在重复数据，数据冗余度高。从来源广泛且类型繁多的教育大数据中勾勒学生画像，分析和挖掘学生潜质特征、自我价值倾向、学业趋势等具有重要的意义。例如，借助学生基本信息、上课情况、与教师互动记录、借阅图书、一卡通、门禁、网上课件下载记录和停留时间等数据，通过数据关联分析和大数据挖掘等技术可以了解学生行为轨迹和真实状态，发现学生的兴趣爱好和行为倾向，明确学生的学习类型和风格，得知学生的知识掌握情况，及时预警学生学业，为学生提供多样化和个性化的帮助，从而提高教学质量，促进智慧教育的发展。

1.2 教育大数据智能分析平台建设思路

本文针对教育大数据现存问题，利用分布式计算、大数据分析、数据挖掘、机器学习等先进技术，研究

多源异构数据的集成和融合、学生精准画像和教育过程动态监控管理，以打破教育信息孤岛，充分挖掘教育大数据的应用价值，按照数据来源、关键技术和平台搭建的思路展开研究，研究思路概括如图 1 所示。

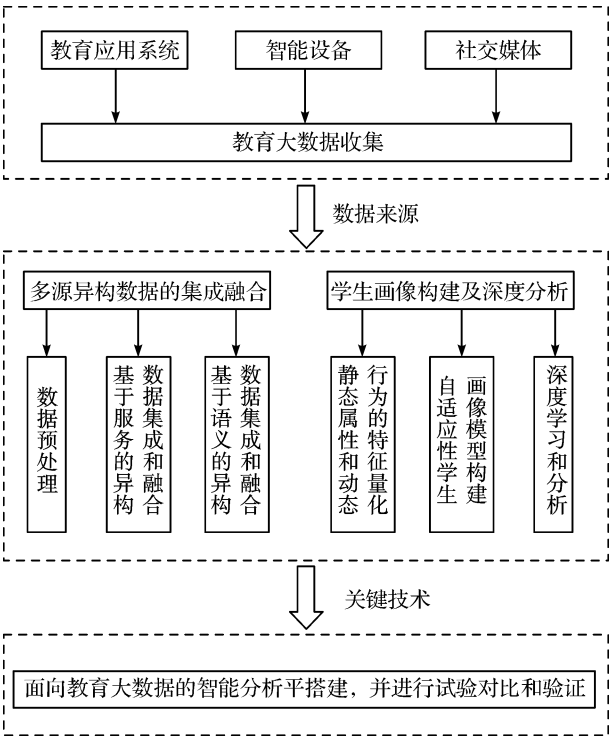

图 1　教育大数据智能分析平台建设采用的研究思路

Fig. 1　Research concept for construction of intelligent analysis platform of educational big data

从图 1 可以看出，在数据收集阶段，教育大数据主要来源于教育应用系统和智能设备及社交媒体。教育应用系统中的数据一般集中存储在各系统的数据库中，易于获取，但可能存在大量重复数据或者质量不高数据，例如数据存在缺失值或异常值。智能设备和社交媒体中的教育数据，一般可通过 API 或者爬虫工具获取，但可能是半结构化或非结构化数据。基于教育大数据来源的广泛性，在教育大数据智能分析平台建设中针对不同问题采用不同的技术处理手段：

1）不同教育系统间的数据关联性较大且存在大量重复数据，依据数值缺失机制，基于极大似然估计、随机森林、遗传算法等模型预测缺失值，以提高数据预处理质量；研究重复数据删除算法，去掉冗余数据，减缩占用存储空间。

2）教育大数据中的结构化、半结构化和非结构化数据并存，采用基于服务的逻辑数据集成和融合技术，利用 HDFS、HBase 存储非结构数据，关系型数据库存储结构化数据和数据分析结果，不同数据之间利用数据服务接口实现逻辑集成和融合，打破教育信息孤岛。

3）研究基于语义的异构数据整合技术。采用分布式计算并利用 MapReduce 技术和 Hadoop 分布式计算框架等提高数据处理速度；通过语义技术将各种异构数据表达为语义资源，然后发布到语义库中，进而实现语义层面上的数据查询和数据计算。

2　教育大数据智能分析平台建设实践

2.1　教育大数据智能分析平台系统架构

教育大数据的智能分析平台的系统架构自下而上分四层：数据层、模型层、应用层和展示层。系统架构具体如图 2 所示。数据层采集数据并进行数据预处理。模型层针对预处理后的数据，设计并构建数据分析模型。应用层利用模型进行教育大数据分析，主要支持教育质量综合分析、教育质量预警和教育决策支持等三方面的应用。展示层利用可视化的方法，将大数据分析的结果进行展示。

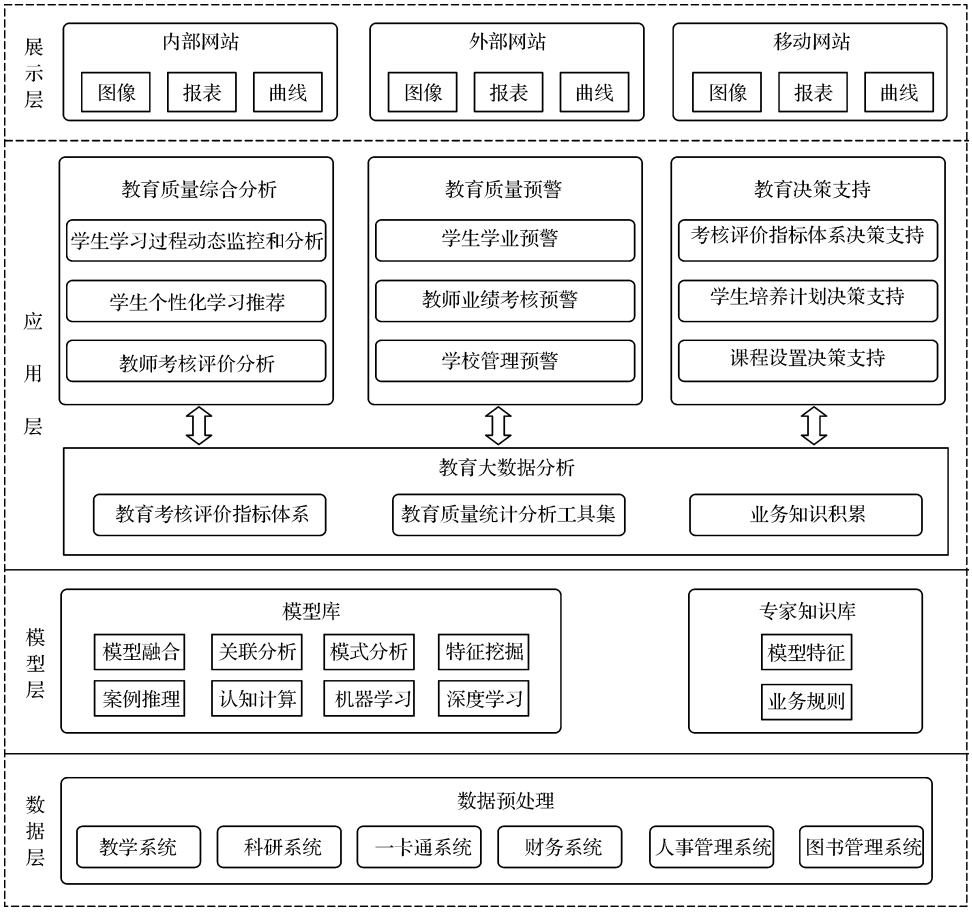

图 2 教育大数据智能分析平台系统架构

Fig. 2 Architecture for intelligent analysis platform of educational big data

2.2 教育大数据智能分析平台技术架构

按照数据收集、数据集成、数据存储、数据分析、数据查询等数据处理流程，教育大数据智能分析平台在建设实践时，包括深度分析场景、统计分析场景、查询检索场景、在线分析处理(online analytical processing，OLAP)场景，如图 3 所示。不同应用场景采用的技术手段和具体内容有以下几个方面：

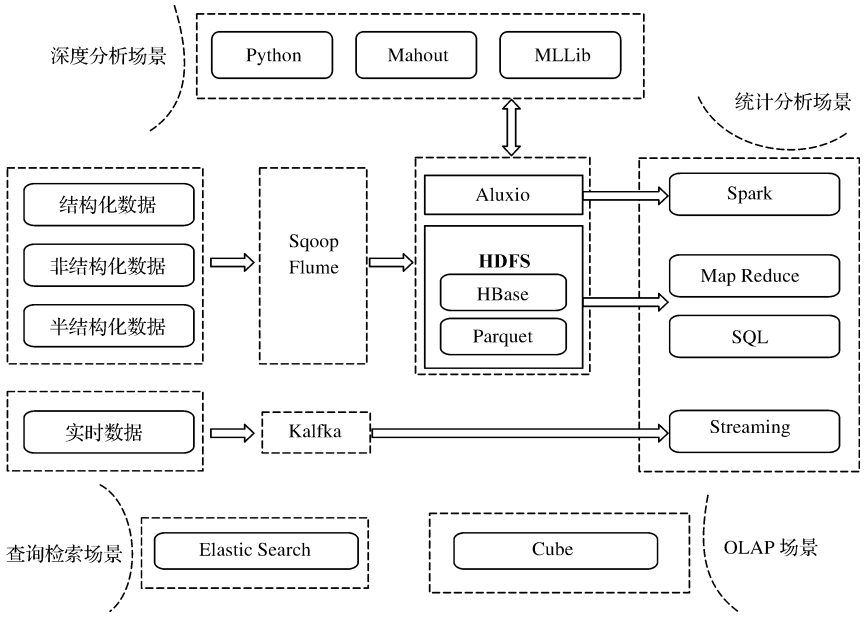

图 3 教育大数据智能分析平台的应用场景

Fig. 3 Application scenario of intelligent analysis platform of educational big data

1）在数据收集阶段，根据结构化、半结构化数据和非结构化数据的特点，采用 Sqoop 和 Flume 导入数据和日志文件；利用 kalfka 采集实时数据；通过智能设备和社交媒体提供的 API 或者爬虫工具获取外部数据。

2）在数据存储阶段，分析不同数据存储方案的优缺点，以 HDFS、HBase、关系型数据库为存储主体，HDFS、HBase 存储非结构数据，关系型数据库存储结构化数据和数据分析结果，提高模型的可扩展性。

3）在数据分析阶段，根据不同的应用场景，采用 Spark、MapReduce、Storm 等计算框架实现批处理和流式处理，采用 Spark MLIib、Mahout 等数据建模工具实现聚类、分类、推荐、过滤、频繁子项挖掘等智能分析功能；数据查询采用 HiveQL 查询语句提高数据抽取、转化、加载的效率。

3　结　论

基于教育大数据的现状分析，本文提出了教育大数据智能分析平台的建设思路，实现多源异构教育大数据的集成和融合，打破教育信息孤岛；以先进技术为手段，搭建教育大数据智能分析平台，实现精确学情诊断、及时预警学业、个性化学习推荐和智能决策支持等，提高教育管理过程的智能性。在后续工作中，我们将进一步研究教育大数据智能分析平台的隐私保护等问题，以提高数据的安全性和平台的可靠性。

参考文献：

[1] LI Y, ZHAI X N. Review and prospect of modern education using big data[J]. Procedia Computer Science, 2018, 129(1): 341.

[2] SCHWERDTLE P, BONNAMY J. Big data in nurse education[J]. Nurse Education Today, 2017, 51(1): 114.

[3] 于长虹，王运武，马武. 智慧校园的智慧性设计研究[J]. 中国电化教育，2014(9)：7.

[4] LIPPENYI Z, GERBER T P. Inter-generational micro-class mobility during and after socialism: the power, education, autonomy, capital, and horizontal (PEACH) model in Hungary[J]. Social Science Research, 2016, 58(1): 80.

[5] 张鸷远. “慕课”(MOOCs)发展对我国高等教育的影响及其对策[J]. 河北师范大学学报(教育科学版)，2014，16(2)：1.

[6] ZHOU M M. Chinese university students' acceptance of MOOCs: a self-determination perspective[J]. Computers & Education, 2016, 92/93: 194.

[7] 范文翔，马燕，李凯，等. 移动学习环境下微信支持的翻转课堂实践探究[J]. 开放教育研究，2015，21(3)：5.

[8] 维克托·迈尔-舍恩伯格. 大数据时代[M]. 盛扬燕，周涛，译. 杭州：浙江人民出版社，2013：1.

[9] 教育部. 教育部关于印发《教育信息化“十三五”规划》的通知[EB/OL]. (2016-06-24)[2018-07-15]. http://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1778/201701/t20170119_295319.html：1.

[10] 邹国伟，成建波. 大数据技术在智慧城市中的应用[J]. 电信网技术，2013，4(4)：122.

[11] 裴莹，付时秋，吴锋. 我国教育大数据研究热点及存在问题的可视化分析[J]. 中国远程教育，2017，12(1)：46.

[12] 周庆，牟超，杨丹. 教育数据挖掘研究进展综述[J]. 软件学报，2015，11(1)：3026.

[13] ASIF R, MERCERON A, ALI S A, et al. Analyzing undergraduate students' performance using educational data mining[J]. Computers & Education, 2017, 113: 177.

[14] FENG T, CHENG Y. Research on algorithm recommended by online education for big data[C]//SHS Web of Conferences. Paris: EDP Sciences, 2015.

[15] JOHNSON L, ADAMS S, CUMMINS M. The NMC horizon report: 2012 higher education edition[J]. New Media Consortium, 2012, 24(4).