文章编号：2096-1472(2018)-05-34-04
DOI:10.19644/j.cnki.issn2096-1472.2018.05.011

基于大数据应用技术的学情分析系统架构分析与设计

李 强[1]，赵晨杰[2]，罗先录[1]
(1.广东东软学院，广东 佛山 528225；
2.江苏传智播客教育科技股份有限公司北京分公司，北京 100096)

摘 要：目前基于信息化、体验式的教学线上和线下课堂，可通过移动端、网页端、嵌入式设备端等捕获大量的学情行为数据。如何采集这些线上线下产生的各种学情数据，利用采集数据的特点结合主流的大数据应用技术进行处理、分析和挖掘，并对受教育者或教育机构提供有用的决策信息成为很多研究机构的研究主题。本文基于目前学情分析系统的发展，引入大数据技术，设计了以Hadoop为核心的学情分析系统，提出了基于学情分析系统的数据挖掘并行算法分析平台设计，实现了一种基于数据的智慧校园平台。

关键词：学情数据；大数据技术；数据挖掘；平台设计
中图分类号：TP301 **文献标识码**：A

Analysis and Design of the Academic Analysis System Based on Big Data Application Technology

LI Qiang[1],ZHAO Chenjie[2],LUO Xianlu[1]
(1.*Department of Computer Science,Guangdong Neusoft University,Foshan* 528225,*China*;
2.*Beijing Branch,Jiangsu Chuan Chi Podcast Education Technology Co.,Ltd.,Beijing* 100096,*China*)

Abstract: At present,information-based and experiential teaching online and offline classrooms can capture a large amount of academic behavior data through mobile terminals,web pages,and embedded devices.The topics in many research institutions includes how to collect the various academic data generated online and offline,and how to use the characteristics of the collected data in combination with mainstream big data application technologies to process,analyze,and mine useful information for decision-making by educators or educational institutions.Based on the development of the current academic behavior analysis system,this paper introduces big data technology,designs an academic behavior analysis system based on Hadoop,proposes a design of an analysis platform of data mining parallel algorithm based on academic behavior analysis system,and implements a data-based smart campus platform.

Keywords:academic data;big data technology;data mining;platform design

1 引言(Introduction)

随着信息技术的发展，数据无时无刻不在产生，特别是教育大数据，已经成为推动教育行业的提升和变革的强大力量。基于信息化、体验式的教学线上和线下课堂，可通过移动端、网页端、嵌入式设备端等捕获大量的学情行为数据，这些数据符合大数据4V特性：海量(Volume)、多样性(Variety)、时效性(Velocity)和有效性(Veracity)，给传统的教育数据存储、分析和处理都带来了极大的挑战。在与其他行业相比，教育界对大数据的广泛接纳比其他成熟行业稍晚。但如今大数据已经慢慢走进教育的各个角落。产生了更多的教育机构和企业开始对教育大数据深入研究并构建可交互的大数据平台。教育的大数据不仅影响学校内部治理的改革，而且会驱动整个教育领域的变革，利用大数据平台构建每一位受教育者的用户画像，针对每一位受教育者给出合理的建议，从而使得教育和关爱每一个孩子成为可能。

“大数据”这一概念已经在各行业的应用获得了极大的成功，也应运而生了“数据科学”这一崭新科学领域，通过大数据理论基础和框架技术可解决教育和大数据融合中所产生的问题。本文提出了教育大数据背景下运用大数据技术处理和分析教育行业数据的技术架构，并基于Hadoop技术生态圈设计了学情分析系统的技术架构及数据挖掘平台，将其应用于学院教学质量监控。

2 学情分析概述(Overview of academic behavior analysis)

学情分析指的是学生在学习方面有何特点、学习方法怎样、习惯怎样、兴趣如何，成绩如何等。其设计理念包括教学方法、学法指导和教学设想，根据获取的数据研究者可以从高校创新创业教育改革、创新创业人才培养、基于产业发展需求的专业结构调整研究、学生学习行为分析、教师教授行为分析，以及个性化推荐等角度展开研究[1]。对教育大数据进行分析，需要从大量数据中进行提取与挖掘，在这个过程中包括数据的清洗、数据选择、数据变换、数据挖掘、模式

基金项目：本文受广东省科技计划项目(2017A040406001)，广东东软学院(基于大数据应用技术的智慧校园平台研究基金资助.

评估和知识表示等。这些分析环节的每个构成都应成为数据分析研究的重要内容，从而最大限度地保持与还原客观事实[2]。

在如今的学校教育中，数据已成为教学改进最为显著的指标，而更多科学决策也是基于数据而产生的。在学校的数据种类不仅仅指考试成绩，也包括入学率、出勤率、辍学率、升学率等。对于具体的课堂教学来说，数据应该是能说明教学效果的，比如学生考试成绩、作业正确率、上课出勤率、积极参与课堂科学的举手次数，回答问题次数、时长与正确率，师生互动的频率与时长等。

根据以上综合分析，基于大数据应用技术的学情分析系统平台研究和建设是有着重大的意义，广东东软学院作为全国应用型大学的典范，我们更加注重课程实践性和学生动手能力，根据大数据分析和挖掘技术可以更加合理调整学院的学科专业、教师教授方式和学生学习方式等。同时，该平台的建立会完善广东东软学院的大数据应用实践教学体系。因此，建立和完善学情分析平台可促进和深化学院学生的学习、教师教学实践，以及帮助相关部门提供更加合理的计划和建议。

3　研究现状(Current research situation)

国外对学习和学情分析的研究起步较早，从2011年起已经积累了很多的理论基础，近年来已经由纯粹的理论概念阶段发展到具体的实际应用阶段，涌现出很多已经处于实用阶段的学习和学情分析系统。与国外相比，国内到目前为止主要还是停留在理论探索和分析阶段，或者是在理论和概念上的拓展，对于具体的学习分析工具和系统的实际应用的研究和开发较少[1−3]。基于大数据应用技术的学情分析系统平台是广东东软学院在学习和学情分析领域的实际应用，利用通用大数据和互联网技术对教育数据进行多维分析。

通过大数据技术和数据挖掘技术结合能够更好为各阶段学生提供更有价值的信息，如“希维塔斯学习”就是一家专门聚焦于运用预测性分析、机器学习从而提高学生成绩的公司[1]。加拿大的一家教育科技公司“渴望学习”(Desire 2 Learn)已经面向高等教育领域的学生，推出了基于他们自己过去的学习成绩数据预测并改善其未来学习成绩的大数据服务项目[2]。“渴望学习”的产品通过监控学生阅读电子化的课程材料、提交电子版的作业、通过在线与同学交流、完成考试与测验，就能让其计算程序持续、系统地分析每个学生的教育数据。老师得到的不再是过去那种只展示学生 分数与作业的结果，而是像阅读材料的时间长短等这样更为详细的重要信息，如此老师就能及时诊断问题的所在，提出改进的建议，并预测学生的期末考试成绩。纽顿的创办人、首席执行官何塞 · 费雷拉和培生高等教育分公司的总裁格雷格 · 托宾合作研发将大学数学、大学统计学、大学一年级作文、经济学和科学等领域纳入教育产品中[3]。

在学习和学情分析系统建设方面，其研究成果主要涉及多个不同的教育系统。

基于Web的学习系统：Pardos等学者在基于Web的数学教学平台AssisTments上，花了两年时间，跟踪分析1393名8年级学生在该平台上的数学学习行为数据，研究学习中体现的情感如厌倦、专注、困惑、挫折等与最终的学习结果之间的关系[4]。Kizilcec等学者针对Mooc教学中低完成率的问题，提出根据学生与Mooc学习课程的交互，对不同的学习者进行分类，该研究对Mooc未来的教学指导设计有一定意义[5]。

基于学习管理系统(LMS)，许多学习分析研究是基于LMS记录的教育数据。Lonn等针对密歇根大学一二年级工程系学生，开发了M−STEM Academy作为早期学业预警系统，研究了如何挖掘LMS数据，以及将这些数据转化成警示数据每周提供给导师，以方便导师对学生进行有针对性的支持[6]。Garcia−Solorzano等指出LMS环境和面对面教学环境不同，教师在线监控学习较困难，许多LMS提供的学生跟踪数据难以理解。针对这个问题，他们设计了一个基于浏览的图像化教学监控工具，帮助教师洞察学生表现，并及时发现潜在问题[7]。

基于Web 2.0或社会学习系统：Gunnarsson和Alterman分析了班上107名学生的博客，使用学生之间互动的相关博客数据，特别是某学生对其他学生博客内容的推介，建立了一个模型来识别有价值的内容和对教师的意义[7]。Southavilay等学者研究了大学生使用云计算工具Google Docs进行合作写作的案例，提出三种可视化方法分析写作进程，包括校订版本演化、主题演化图、主题合作网络来探索学生的思维、能力表现，目的是让团队中的每一个学生的写作更有成效[7]。在社会学习分析方面，英国学者Ferguson和Buckingham做了全面的分析，提出了五种方法研究正式和非正式的教育环境数据源，注重从社会维度如情感、性格、学习网络所反映出的学习者的学习状态。

基于实时学习系统：传统教学环境中，教师通过分析学生表现如出勤率、考试、教室内的行为等传统数据来帮助学生。现在利用信息和通信技术(Information Communication Technology)，可以将教室中的交互情境数字化，从而使数据更加多源。Blikstenin提出多情态交互分析系统，数据包括视频、音频、文本、姿势、生物传感信息(如眼球跟踪)等[6,7]，研究者可以探究过去不可能获知的学生学习活动，进行更全面的分析。

通过上述分析，目前基于教育大数据的数据分析和数据挖掘仍处于发展的初期，特别是在国内的研究与实施仍处于起步阶段，在实际的应用中仍然不能依靠数据提供的有价值信息促进学生学习。因此，利用不同的数据源产生的分布式教育数据，建立一个集成和开放的学情分析系统是很有必要的。

4　基于大数据技术的学情分析系统框架(Framework of academic behavior analysis system based on big data technology)

基于大数据应用技术的学情分析系统平台研建是大数据技术与教育行业结合的一种实际应用的体现，通过平台提供的功能来改善学生的学习行为，为教师提供更好的教学方案，为职能部门提供合理的管理方案等。基于主流的Hadoop技术搭建大数据平台，提供了数据的清洗、过滤及汇总操作，根据业务需求选取合适的大数据框架进行大数据分析。在大数据平台之上构建了数据挖掘并行算法处理平台，挖掘更加有价值的数据信息，为学生推荐更加科学合理有用的学习资源或其他资源。

4.1　系统开发的目标

基于大数据应用技术的学情分析系统研建是将大数据技术、数据挖掘技术和机器学习技术等计算机技术应用于教育行业数字化和信息化的重要方面，可以通过平台帮助学生更好的学习、帮助老师更好的教学，为学校管理层和决策层提供更加科学的决策依据。区别于其他行业，教育行业逐渐被认为是大数据可以大有作为的一个重要领域，利用大数据技术促进和完善教育教学改革。此项目的建立将会更加加快高校信息化建设的速度和质量。

4.2 系统开发的可行性分析

根据教育行业业务需求，设计了合理的大数据处理与分析平台和数据挖掘并行算法处理平台，项目重点为利用Hadoop平台对大数据日志进行存储、分析、处理，对采集的数据进行分析，完成相应日志的入库、处理、分析、实时查询等主要功能。对经过处理后的数据进行数据挖掘，挖掘出有价值的信息，给用户推荐更好的资源。按照实施计划部署相应的大数据系统平台，根据平台的数据处理量，初步规划Hadoop集群的数量为5—10台。

4.3 系统开发数据来源

数据来源于学院学生信息管理系统、招生就业系统、校园考勤系统、图书管理系统平台、教务等真实数据，同时从辅助教学平台上抓取有价值的可信度高的数据，如发帖数据(贴吧等)、学习者调查、用户资料、网络社交媒体等获取相关数据，从而形成学情分析系统大数据平台的数据集。

4.4 系统开发过程及关键技术

首先根据数据集的数量级(PB或TB)确定集群数量，确定选择在线大数据平台还是本地建立服务器集群搭建大数据处理与分析的分布式平台。

对数据源进行初步整理和分析，学校相关信息系统需要与相关职能部分沟通数据中有价值或权重较高的字段或描述，从其他网站采集的数据需要经过讨论分析后确定技术可行性和评估数据源价值。

将采集到的各数据源通过大数据技术提供的Sqoop(主要用于在Hadoop(Hive)与传统的数据库(Mysql、Oracle等)间进行数据的传递)和Flume(日志采集工具)技术将数据源导入或推送到HDFS分布式文件系统中，对未来可能开发并投入使用的管理信息系统通过Log4G日志的形式记录，每天或每周定点通过大数据日志收集工具Flume向大数据平台的HDFS分布式文件系统推送记录数据。

对存储在HDFS中的数据进行数据的ETL(清洗、过滤、汇总)，大数据分析部分采用Hive与Impala结合方式，对查询速度要求较高的采用基于内存的迭代式框架Spark技术框架，此时经过大数据分析后的数据可直接通过Web系统作统计数据的页面展示。

处理之后的数据可以作为数据挖掘平台进行聚类、分类、关联和回归等数据挖掘算法的并行化处理媒介，从而得到学生行为分析的重要信息，最后通过推荐系统为学生推荐合理的资源信息。

4.5 集群环境搭建方案

根据业务需求搭建集群10台左右的大数据处理和分析平台， 项目中需要Hadoop集群能够商用，并且要求稳定，性能没有瓶颈。所以针对于Hadoop服务器，需要做一些操作系统级别优化(CentOS6.4)，以使得集群获得最优的性能和稳定性能[8]。

当Hbase提供服务速度难以保证情况下，使用Impala替换HBase、Impala StateStore和Impala Catalog Server安装到HBase master所在机器，HBase Region所在机器安装Impala Daemon。JobTracker机器变为ResourceManager，TaskTracker变为NodeManager。

以上集群安排是根据数据和业务进行预估暂定集群数量在10台以下，如果后期集群数量增加应该重新调整各节点的配置。

管理服务器是平台的主节点，负责管理计算和任务分配等，节点1—4和剩余节点机都属于从节点，从节点负责执行主节点分配的存储和计算的任务。要求数据节点尽可能放在一起利于数据的本地化，加快数据查询速率，这里的HRegion由HRegionServer存放和管理本地节点，主要用于读写HDFS，管理Table中的数据，因为应该将HRegion与HDFS中的DataNode安装在同一个从节点服务器中。HA采用管理服务器1和服务器2互备。

项目根据预期的数据和业务需求搭建集群在10台以下的大数据处理和分析平台，项目中采用稳定、性能瓶颈小的Hadoop集群。同时针对于Hadoop服务器，需要做一些操作系统级别优化(CentOS6.4)，以使得集群获得最优的性能和稳定性能。

4.6 架构设计

基于大数据应用技术的学情分析系统平台架构分为大数据处理与分析平台和数据挖掘并行算法分析平台组成。其中大数据处理与分析平台主要对数据源进行ETL过程，满足一部分的数据查询需求，以及图形化展示需求。数据挖掘并行算法分析平台主要对经过大数据处理后的数据挖掘出潜在有价值的信息，为学生的学习、生活等方面提供个性化的推荐和意见等[9,10]。

4.6.1 学情分析系统的大数据平台架构设计

学情分析系统的大数据平台架构设计如图1所示。

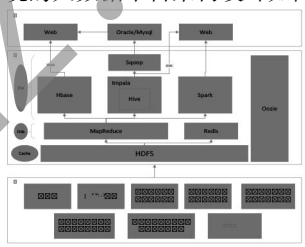

图1 学情分析系统的大数据平台架构设计

Fig.1 Framework design of big data platform for academic behavior analysis system

学院信息网站数据和由外部网站采集的数据均以压缩形式上传HDFS对应的目录，相应的Mapreduce直接从HDFS上获取原始数据进行数据处理和分析。Mapreduce主要包含三种类型：汇总部分(过滤、清洗、汇总)。使用分布式存储系统HBase存储一些数据量级较大的数据和进行一些简单的统计分析，同时，将Mapreduce处理后的数据存储在Hbase中，之后使用Thrift服务与Web进行交互显示。Spark分析部分主要利用SparkSql、SparkMLlib、Graphx三大组件进行复杂的批量处理、基于响应速度要求高的交互查询、基于实时数据流的查询。Mapreduce汇总部分的结果加载到Hive中并且使用Impala提供Web端的查询。需要做进一步分析和关联的部分使用Sqoop导出到Oracle或Mysql中，由Oracle或Mysql来完成Web端复杂图形的展现[11,12]。

4.6.2 学情分析系统数据挖掘并行算法分析平台设计

数据挖掘并行算法分析平台如图2所示。

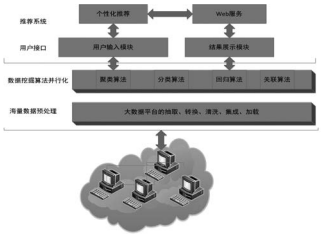

图2 数据挖掘并行算法分析平台

Fig.2 Analysis platform for data mining parallel algorithm

数据经过预处理后，需要考虑如何能让数据发挥作用。这就需要采用数据挖掘平台提供的数据挖掘和分析工具、算法进行有价值信息的抽取，从而实现从数据到信息的高效转化。对受教育者的学习数据、行为数据等进行深入分析和挖掘，查找可能存在的问题等重要信息，并利用这些数据为改善受教育者的成绩或学习行为提供个性化的服务。同时，借助数据中一位受教育者的各个维度数据来综合评判学生表现，利用大数据挖掘技术，针对学生存在的问题提供合理的建议与意见[13,14]。

通过大数据和数据挖掘进行学习分析能够为每一位受教育者创设一个量身定做的学习环境和个性化的课程，还能创建一个早期预警系统以便发现开除和辍学等潜在的风险，为受教育者的多年学习提供一个富有挑战性而非逐渐厌倦的学习计划。因此，学习可以依靠大数据驱动。通过分析和挖掘，进一步改善教学的方式与方法，进一步促进学生学习成绩的提高。

根据平台需求主要使用以下五种数据挖掘技术从大数据分析后的数据中提取有价值数据信息：

(1)预测(Prediction)——基于对历史数据的分析，预测新数据的特征或数据的未来发展趋势。例如，要具备知道一个学生在什么情况下尽管事实上有能力但却有意回答错误的能力。

(2)聚类(Clustering)——发现数据的内在结构。这对于把有相同学习兴趣的学生分在一组很有用。

(3)相关性挖掘(Relationship Mining)——发现各种变量或因素之间的关系，并对其进行解码以便今后使用它们。这对探知学生在寻求帮助后是否能够正确回答问题的可靠性很有帮助[14]。

(4)升华人的判断(Distillation for Human Judgment)——建立可视的机器学习的模式。

(5)用模式进行发现(Discovery with Models)——使用通过大数据分析开发出的模式进行“元学习”(Meta-Study)[14]。

5 结论(Conclusion)

本文从大数据视角提出，利用Hadoop生态圈构建基于大数据应用技术的智能化学情分析服务架构，该方案主要目标是解决海量教育信息的汇聚、存储和存取及分析和挖掘等，从而为智能化教育教学服务提供技术支撑。

同时，随着国家信息战略的实施，网络带宽及其他相关硬件设施的发展，这为大数据技术应用提供了较为广阔的空间。针对教育大数据对教育、教学及学生学习的方方面面影响，基于大数据技术的教育改革势在必行，高效创新创业教育改革、基于产业需求的高效专业调整及学生的学习行为分析和教师教授行为分析都将从基于知识或经验的改革转移到基于数据的教育教学改革。根据大数据平台分析结果我们可以更加合理的调整学科专业，教师教授方式和学生的学习方式等。该平台的建立促进和深化了学校学生学习、教师教学实践及帮助相关部门提供合理的计划和建议。

从教育行业整体上看，目前教育行业数据的采集仍处于布局和建构的初级阶段，大数据在教育决策、教学过程中的运用还处于摸索和起步阶段，大数据人才培养的完善体系还没有建立起来。目前大数据在教育领域的应用总体上也呈现出“产业应用的成熟度大于学校应用的成熟度”的态势。本文提出的学情分析平台正是在弥补这方面的空缺，同时如何将信息技术更加高效地运用与课堂和学生发展的，学情分析及其对教育的积极作用的不断展现，相信这些难题会不断得以解决。

参考文献(References)

[1] SOUTHAVILAY V,YACEF K,REIMANN P,et al.Analysis of collaborative writing processes using revision maps and probabilistic topic model[C].Proceedings of the Third International Conference on Learning Analytics and Konwledge.ACM,2013:38-47.

[2] FERGUSON R,SHUM S B.Social learning analytics:five approaches[C].Proceedings of 2nd Internationnal Conference on Learning Analytics and Knowledge.ACM,2012:23-33.

[3] BLIKSTEIN P.Multimodal learning analytics[C].Proceedings of 3nd Internationnal Conference on Learning Analytics and Knowledge.ACM,2013:102-106.

[4] Anya Kamenetz.Big Data Comes To College[EB/OL].http://www.npr.org/blogs/ed/2014/07/04/327745863/big-data-comes-to-college,2014-07-04.

[5] PARDOS Z A,BAKER R S,SAN PEDRO M O,et al.Affective states and state tests:Investigating how affect throughout the school year predicts end of year learning outcomes[C].Poreedings of the Third International Confernce on Learing Analitics and Knowledge.ACM,2013:08-12.

[6] KIZILCEC R F,PIECH SCHNEIDER construcing disengagement:analyzing learner subpopulations in massive open online course[C].Proceeding of the Third International Conference on Learning Analytics and Knowledge ACM,2013:170-179.

[7] LONN S,KRUMM A E,WADDINGTON R J,et al.Bridging the gap from knowledge to action:Putting analytics in the hands of academic advisors[C].In Proceedings of the 2nd International Conference on Learing Analytics and Knowledge.ACM,2012:170-178.

[8] 王元卓,靳小龙,程学旗.网络大数据:现状与展望[J].计算机学报,2013(6):1125-1138.

[9] 马晓玲,邢万里,冯翔,等.学习分析系统构建研究[J].华东师范大学学报(自然科学版),2014(2):10-39.

[10] 夏晓峰.基于Hadoop的Mooc学习分析系统的构建[J].计算机时代,2016(7):45-49.

[11] 李春燕,何一舟,戴彬.Hadoop平台的多队列作业调度优化方案研究[J].计算机应用研究,2014,31(3):705-707;738.

[12] 李天目.云计算技术架构与实践[M].北京:清华大学出版社,2013.

[13] 李馨.高等教育大数据分析:机遇与挑战[J].开放教育研究,2016,22(4):51-55.

[14] 朝乐门.数据科学[M].北京:清华大学出版社,2016.

作者简介：

李　强(1976-)，女，硕士，副教授.研究领域：软件技术，数据库技术，大数据技术.

赵晨杰(1990-)，男，硕士生.研究领域：大数据技术，人工智能.

罗先录(1973-)，男，硕士，副教授.研究领域：软件技术，人工智能.