文章编号：2096-1472(2020)-05-43-03
DOI:10.19644/j.cnki.issn2096-1472.2020.05.012

基于大数据的高校智慧校园学生综合测评系统设计与研究

茆灵铖，谢桂芳，邵周伟，时海茹，蒋秀莲

(徐州工程学院，江苏 徐州 221000)
✉3100786500@qq.com; 1872769089@qq.com;
2465532727@qq.com; 1277156627@qq.com; jxl@xzit.edu.cn

摘　要：当前，信息化正面临着一个全新的阶段，即以数据的深度挖掘和整合应用为核心的智慧化阶段，智慧校园已成为时下高校信息化建设的重要内容。分析高校信息化建设现状和Hadoop、Spark等大数据技术框架，并重点从数据存储层、核心业务层和信息展示层对智慧校园学生综合测评系统进行分析与设计，为大数据技术与智慧校园的深度融合提供方案。

关键词：智慧校园；数据挖掘；Hadoop和Spark
中图分类号：TP274　**文献标识码**：A

Design and Research of the Student Comprehensive Evaluation System for Smart Campus Based on Big Data

MAO Lingcheng, XIE Guifang, SHAO Zhouwei, SHI Hairu, JIANG Xiulian

(*Xuzhou University of Technology*, *Xuzhou* 221000, *China*)
✉3100786500@qq.com; 1872769089@qq.com;
2465532727@qq.com; 1277156627@qq.com; jxl@xzit.edu.cn

Abstract: At present, informatization is entering a new stage, that is, the intelligent stage with data deep mining and integrated application as the core. Smart campus has become an important part of university information construction. This paper analyzes the current situation of university informatization construction and big data technology framework such as Hadoop and Spark, analyzes and designs the student comprehensive evaluation system of smart campus from data storage layer, core business layer and information display layer, so as to provide a scheme for the deep integration of big data technology and smart campus.

Keywords: smart campus; data mining; Hadoop and Spark

1 引言(Introduction)

“智慧校园”源于IBM公司在2008年提出的“智慧”地球理念，其核心是感知、联通、智能。它是数字校园发展的高端形态，以物联网为基础，通过宽带移动、云计算、大数据等技术整合数字校园阶段规模巨大的多源异构数据[1]，以综合信息服务平台为载体，提供校园学习、工作、生活一体化的智能环境[2]。目前，各高校大都具有完备的信息系统和大量的学生个人数据，然而在信息化水平和应用上仍处于数字校园的阶段，没有充分探测全校师生认知行为和校园环境动态变化的信息支撑平台。

数据挖掘技术在企业运营中得到广泛应用，但高校数据挖掘意识不强。随着智慧校园的推进，研究者逐渐重视对有关学生教育大数据的分析与挖掘，因此针对学生信息测评方面的研究不是很多，且高校学生系统大都由不同部门运营和维护，学生测评方式单一，缺少统一支持海量数据处理的平台支撑智慧校园的建设。因此利用大数据技术建立分析挖掘学生信息的数据处理与应用平台，具有重要实际意义。

2 基于大数据技术的学生综合测评系统架构(Architecture of student comprehensive evaluation system based on big data technology)

大数据技术是指用一系列工具来对大量的结构化、半结构化和非结构化数据进行采集、存储，从而得到分析和预测结果的技术[3]。大数据萌芽于20世纪90年代，这一时期数据挖掘理论与数据库技术逐步成熟。21世纪以来，随着Web2.0应用迅猛发展，非结构化数据大量产生，大数据技术快速突破，形成了并行计算和分布式系统两大核心技术，Hadoop和

基金项目：江苏省大学生创新创业训练计划项目(XCX2019056).

www.rjgcz.com

Spark分布式计算框架也应运而生。

(1)Hadoop分布式计算框架

Hadoop由Java开发，是目前大数据技术的主流软件架构，具有良好的容错性和稳定性，以及强大的IDE支持。Hadoop生态圈以HDFS和MapReduce为核心，HDFS是分布式文件处理系统，它将大型文件拆分处理成多个小型文件单位，便于底层庞大数据的存储，而分布式并行编程模型MapReduce可对这些文件中的数据集进行并行运算。同时Hadoop生态圈还有Flume、Hive、HBase、Zookeeper、Sqoop、Mahout、Ambari、Pig等功能组件。

(2)Spark分布式计算框架

Spark由基于静态编译的Scala语言开发[4]，速度快，在执行过程中注重函数本身而非数据和状态的处理，并将计算数据、中间结果都存储于内存中，大大减少了I/O开销，更适合数据挖掘中的运算。而Hadoop的MapReduce计算模型表达能力有限，磁盘I/O开销大，延迟高，难以胜任实时快速的计算需求，故可将Spark作为一种计算框架通过JVM取代MapReduce融入Hadoop生态圈中。并且Spark具有良好的API，能够给开发人员带来良好的用户体验。

(3)大数据技术与智慧校园的深度融合

通过Hadoop和Spark这两个大数据框架对高校信息化应用水平进行改善，即以Hadoop的分布式文件系统HDFS为主，存储数字校园阶段各管理信息系统的数据，再以Spark的计算处理功能为主，实现这些数据的深度挖掘。进而通过智能分析，为用户提供智能预测、预警并辅助决策，推动大数据技术与智慧校园的深度融合。同时结合数字校园阶段学生系统的建设特点，可构建一个基于大数据技术的高校智慧校园学生综合测评系统。

(4)学生综合测评系统总体架构

高校智慧校园学生管理系统遵循高内聚低耦合的设计原则，采用流行的Hadoop和Spark开源软件构建平台[5]，使系统可便利地实现平滑升级，并保证系统风格统一、美观、易于用户操作。在充分共享信息资源的同时对各种访问权限进行严格限制，保持高可靠性和高安全性[6]。测评系统分为三层，如图1所示。

图1中，数据存储层是系统的最底层，为上层提供数据源，如存储学生的学业成绩、消费、一卡通等校园大数据。核心业务层处于系统中间层，进行数据的整合和运用数据挖掘模型分析数据信息。信息展示层位于系统最上层，作为用户与系统之间的交互界面。

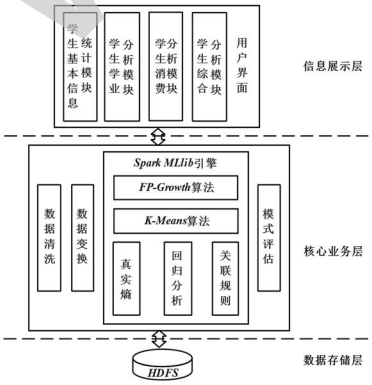

图1 系统总体架构图

Fig.1 Overall system architecture diagram

3 数据存储层(Data storage layer)

由于高校的学生信息数据大都存储在不同的管理信息系统中，故构建高校智慧校园学生综合测评系统的首要任务是对这些数据进行整合，其处理流程如图2所示。图2中，ETL是指将数据从源端处经过抽取、转换、加载至目的端处的过程；Sqoop是可实现Hadoop系统与关系数据库进行数据迁移的专门工具；HBase是具有高性能、高可靠性、可伸缩、实时读写等特点的列式数据库，一般采用HDFS作为其底层数据存储；Hive是基于Hadoop的数据仓库工具，可对Hadoop文件中的数据集进行数据整理、特殊查询和分析存储。数据存储层先通过ETL数据预处理工具[7]，将分布在各部门管理信息系统中的学生数据抽取到临时中间层，然后进行清洗、转换、集成、装载，最后结合Sqoop工具，将处理后的数据导入到基于Hadoop系统的中心数据库中，从而利用Hadoop中的HDFS分布式文件系统将学生日积月累产生的大量数据存储到数据仓库中。

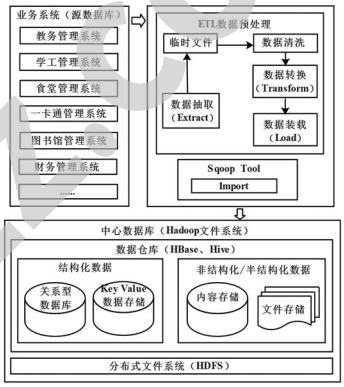

图2 数据存储层构造流程图

Fig.2 Flow chart of data storage layer construction

4 核心业务层(Core business layer)

在Hadoop和Spark的基础上，系统在本层可通过FP-Growth算法、真实熵、K均值聚类、回归分析等数据挖掘模型，对学生的个人信息、学业成绩、学业状态、一卡通消费、进出图书馆次数等数据进行分析，从而得到学生的测评结果[8]。

通过增加最小模式长度来优化FP-growth算法[9]，可生成描述能力更好的频繁模式，学校食堂和超市可以根据这些模式来调整菜品供应以及超市商品的摆放，同时还可调整物品的供应量。K-means聚类收敛速度快、易于理解，以学生消费的次数、金额和用途等数据进行聚类[10]，可对学生的消费水平进行分类。回归模型能够对图书馆、食堂的人员流动进行预测，相关管理人员可据此合理安排工作人员值班。真实熵用于解决人类移动行为的可预测性问题[11]，借助学生在校园各个地方的出入、消费数据可以了解学生的性格特征。将学生的日常行为数据和学业成绩作为训练集，可得到分类规则[12]，预估学生考试不及格、学业障碍等的可能性，提前预警，督促其完成学习任务。

校方通过特定的算法，以数据挖掘结果为依据和支撑，可以制定更合理的教学管理政策，如根据学生消费水平确定贫困生补助资格、等级，根据学业成绩和行为特征进行个性化教育、制定更加人性化的奖学金政策等。

5 信息展示层(Information display layer)

信息展示层是平台与用户交互的可视化窗口，本系统

www.rjgczz.com

在核心业务层的基础上，对数据挖掘得到的有价值的信息进行整合并分模块展示，将其分为学生基本信息统计、学业分析、消费分析和综合分析等四个模块，主要功能如表1所示。本系统将采用数据挖掘算法从校园大数据中得到的有价值的信息以可视化、模块化的方式呈现给用户，旨在方便快捷地为用户提供学情分析、消费分析、综合对比等服务[13]。

表1 信息展示层模块功能表

Tab.1 Information display layer module function sheet

模块名称	模块数据来源	模块主要功能
学生基本信息统计模块	以招生管理系统、学工管理系统的数据为主	按整体、校区、年级等分析学生的性别、民族、地区、家庭收入、学籍等信息
学生学业分析模块	以教务管理系统、图书馆管理系统的数据为主	分析学生的学业情况、学习习惯、学习状态、阅读偏好、奖惩和奖学金分布
学生消费分析模块	以一卡通管理系统、食堂管理系统的数据为主	分析学生的经济状况、消费偏好、消费趋势和消费结构
学生综合分析模块	综合各管理系统的数据进行多样化的分析	为校园管理者进行奖学金评比、贫困生补助、教学区开放时间提供决策支持

(1)基本信息统计模块

本模块整合学生的性别、民族、地区、家庭收入等基本的个人信息，由数据仓库提供的类似于关系数据库SQL语言的Hive QL即可对学生的个人信息进行特征分析，通过饼图、柱状图等在网页进行可视化展示。在本模块中，每个用户都可以查看全校学生整体统计分布情况，并且校园管理者用户在自己的权限范围内可以查看每个学生的详细情况，而每个学生用户仅能够查看自己的详细信息。

(2)学生学业分析模块

本模块整合学生的学业成绩、进出图书馆次数、借阅记录和奖惩情况等信息，经核心业务层处理得到学生学业的统计数据，如学生的学业情况、学习状态、奖惩分布、阅读偏好等。在本模块中，每个用户都可以查看学生总体的学业分布情况，并且学生用户可以查看自己的学业数据和学习记录，教师用户可以查看自己所教授班级学生的学业数据和学习记录。同时系统管理员可以根据阅读偏好来提醒图书馆管理者优化图书馆购书类别，根据学业情况对学生进行挂科预警等。

(3)学生消费分析模块

本模块整合学生的一卡通消费数据、食堂及商店消费数据等信息，经核心业务层处理得到学生的消费统计数据，如学生的平均消费情况、饮食偏好、消费结构等。在本模块中，学生用户可以查看学生总体的消费分布情况和自己的消费数据。同时系统管理员用户可以根据学生消费的偏好和频繁模式来提醒食堂和商店管理人员优化商品的供应，根据学生消费情况衡量学生家庭条件，为学校精准关爱贫困生提供数据支撑。

(4)学生综合分析模块

本模块是信息展示层的核心模块，基于前三个模块的分析数据，由系统管理员自定义设置，在核心业务层中进行更深层次的处理，可以得到不同指标的统计数据。如根据学生的消费数据加权得到经济富裕指数，根据学生的学业数据得到成就性指数，根据学生行为数据结合真实熵算法得到严谨性指数等[14]。在本模块中，每个用户都可以查看学生总体的指标分布情况，并且学生用户可以查看自己的详细分析情况。管理员用户可以根据这些指标数据结合相关规定进行奖学金评比、贫困生补助、教学区开放时间等活动。

6 结论(Conclusion)

通过对大数据技术和高校教育教学工作深度融合的研究，在数字校园的基础上，引入大数据计算框架Hadoop和Spark以及经典的数据挖掘模型，构建以大数据、物联网、云计算等技术为核心的学生综合测评系统，对学生的基本信息、学业信息、消费信息、综合信息进行分析挖掘，从而为高校进行精准的教育教学管理提供科学合理的有效支撑。

参考文献(References)

[1] Fang Dong,Xiaolin Guo,Pengcheng Zhou,et al.Task-Aware Flow Scheduling with Heterogeneous Utility Characteristics for Data Center Networks[J].Tsinghua Science and Technology,2019,24(04):400-411.

[2] Tongya ZHENG,Gang CHEN,Xinyu WANG,et al.Real-time intelligent big data processing:technology, platform,and applications[J].Science China(Information Sciences),2019,62(08):102-113.

[3] Hira Zahid,Tariq Mahmood,Ahsan Morshed,et al.Big Data Analytics in Telecommunications:Literature Review and Architecture Recommendations[J].IEEE/CAA Journal of Automatica Sinica,2020,7(01):18-38.

[4] Xiaoming Ye,Xingshu Chen,Dunhu Liu,et al.Efficient Feature Extraction Using Apache Spark for Network Behavior Anomaly Detection[J].Tsinghua Science and Technology,2018,23(05):561-573.

[5] 范振东,陈晖,王海涛,等.基于大数据的智慧校园学生综合测评系统[J].电信快报,2018(11):25-27;32.

[6] 常镜洳.基于大数据的智能工厂数据平台架构设计与研究[J].软件工程,2019,22((12)):34-36.

[7] 王继鹏,金云智,李伟.勘探开发数据整合之ETL系统的研究与实现[J].中国矿业,2019,28(S2):191-194;199.

[8] 段玉婷.基于校园卡的学生消费信息数据挖掘与应用研究[D].西南科技大学,2018.

[9] 黄婕.基于Spark平台的FP-Growth算法优化与实现[J].湖南工业大学学报,2020,34(01):77-84.

[10] 许家钰.基于k-means算法的WiFi用户行为分析系统设计与实现[D].北京:北京邮电大学,2019.

[11] 吴一帆.eduExplorer:基于校园行为数据的可视分析系统[D].成都:电子科技大学,2018.

[12] 周庆,王卫芳,葛亮,等.基于一卡通数据与课程分类的学生成绩预测[J].电脑知识与技术,2018,14(24):236-239.

[13] 申华.基于大数据的高校学生综合测评系统设计与实现[D].北京:北京工业大学,2017.

[14] 李蒙.基于校园大数据的学生行为挖掘方法应用研究[D].西安:西安电子科技大学,2019.

作者简介：

茆灵铖(1998-)，男，本科生.研究领域：数据挖掘.

谢桂芳(1998-)，女，本科生.研究领域：智慧校园.

邵周伟(1998-)，男，本科生.研究领域：数据分析.

时海茹(1999-)，女，本科生.研究领域：智能算法.

蒋秀莲(1968-)，女，硕士，研究员级高级工程师.研究领域：智慧校园，大数据.

www.jgcz.com