

基于智慧校园数据的高职学生行为画像研究

杜立

(汉中职业技术学院 陕西 汉中 723000)

【摘要】随着时代的发展及技术的进步,互联网技术在各个领域深度应用,很多高校基于信息化技术,在智慧校园建设取得了长足进步,围绕教育信息化2.0行动计划和教育信息化“十四五规划”,以“消除信息孤岛,建设智慧校园,提升互联网+应用能力,增长管理效能”的总体目标,提升学校信息化建设与应用水平,支撑教育高质量发展,本研究提出了基于智慧校园数据的告知学生行为画像研究,本研究依托于智慧校园建设各业务系统的数据,对数据进行整合、清洗、加工,使用数据分析手段对数据进行分析,对学生各个维度的行为进行刻画,最终的到学生在学校生活学习的实时综合画像,对于学生的异常情况及时进行预警,及时采取措施,提高管理能力,提高教学质量,为决策、管理、和规划提供可靠依据。

【关键词】大数据;行为画像;学生管理

1 引言

近些年来,信息化在各个行业落地生根并取得了长足进步,物联网、云计算和大数据等技术助力智慧校园建设,随着校园一卡通的使用和各大业务系统使用,学生以及教师在学校的一举一动都产生了海量的数据,这些海量数据,这些海量数据在以惊人的速度累计,简单的存储已经不能满足学校决策者对于新形势下信息化发展的需求。我们不能紧紧满足于构建系统,产生数据,而应该渐渐将重心转移到使用数据,挖掘数据背后的潜在规律和价值上。

对于学生群体来说,校园一卡通数据最能反映出学生在校园中的所有行为习惯,除此之外,还有上网情况、学生在校园中的轨迹以及基于新建智慧教室的上课情况。

通过对学生成绩、智慧教室考勤、图书馆借阅情况,上网数据等的科学分析,可以有效提升学校管理决策能力。除此之外,对于智慧校园数据的分析还可以促进教学进步,让决策者更了解学生,也让学生更了解自己,同时对于学生管理工作也是大有裨益。

2 学生行为画像的构建

2.1 研究数据

本研究所使用的数据为汉中职业技术学院的各类系统的数据库,包括来自教务系统、学生管理系统、一卡通数据、图书管理系统等业务系统的数据。传统模式下的各业务系统数据分别存储于各业务系统中,使用分散,获取不方便,学校近年来已建成基础平台,打通各业务系统数据获取,但是这些数据获取到的还是原始数据。后期共享数据平台的建设正在进行,预期建成后将大大减少数据获取的难度。目前情况下,只能通过基础平台打通的数据通道对数据进行获取,整合。对于其他暂未信息化的数据,我们使用OCR对于这部分数据进行自动录入,可以填补一些数据上的空白,使得校本数据更加完善丰满。

2.2 数据预处理

基础数据平台获取数据可将数据保存为各种形式,各种数据库也支持将数据保存为各种形式,对数据进行清洗与集成有多种方式,常见的有使用数据仓库技术(ETL)进行整合^[1]。但是ETL多适用于大型数据规模下的数据仓库构建,在这里我们选择使用Python的pandas库进行数据整合、清洗^[2]。pandas是一个快速、有效、易用的开源数据分析处理工具,广泛应用于数据分析领域。pandas提供功能强大的类库,不管数据处于什么状态,他可以帮助我们清洗数据,排序数据,最后得到清晰明了的数据。本研究使用pandas删除重复数据,对于缺失值异常值进行处理,最终整合得到想要的数据库,这是数据分析的第一步,好的数据可以大大降低处理难度,得到更好的效果。

2.3 数据可视化

行为画像很重要的一点是数据可视化,可视化可以帮助我们更直观的审视数据。在这里选择python的matplotlib包进行可视化,Matplotlib是Python绘图领域使用最广泛的套件。它能让使用者很轻松地将数据图形化,并且提供多样化的输出格式^[3]。

2.4 聚类分析

聚类分析是对数据进行分组的技术,它是一种寻找数据之间内在关系的技术,聚类在数据中寻找相似的组,这些组被称为簇。聚类分析是一种无监督的学习方式,相比于监督学习,

它不需要先验的知识,无需训练即可实现分组。聚类可以很好的发现数据之间的关系,对学生进行分类,比如根据一卡通中的消费情况对学生经济状况进行分组,已有通过一卡通数据对学生进行资助的案例,获得很好的效果,这也是大数据技术的一项良好应用。同时,对于学生上网日志记录可以将不同上网行为的学生进行分类,同时结合学生成绩及上课出勤情况对于学生整体情况有了掌握。

在这里使用的是K-means算法对数据进行聚类,K-means算法也叫K均值算法,是最常见的聚类算法之一。K-means算法的计算过程如下:

(1) 从样本中随机选择K个点作为初始质心。

(2) 计算每一个样本到步骤1选择的质心的距离,将样本划分到距离最近的质心对应的簇中。

(3) 计算每个簇的所有样本的新的质心。

(4) 重复步骤2、3指导达到截止条件。

2.5 学生画像

通过数据清洗、可视化、聚类步骤,最终将形成学生的综合画像,它包含学生在校学习生活的方方面面,将前几个步骤得到的结果进行有机的整合、分析。分析出学生的几个方面的画像:

(1) 学生个人基本情况画像:学生个人基本情况画像通过聚类分析对学生的生源地信息、年龄、性别、民族、专业等情况进行分析,全面对学生的基本信息进行刻画。

(2) 学生学业画像:学习情况是学生的一项重要信息刻画,它包括对学生的学业情况进行预警和等级划分。学习是学生在学校的重要环节,首先对于每一名学生的学习成绩、出勤情况进行分析;其次对于成绩异常的学生,例如对于存在补考、缺考以及重修情况的学生进行重点关注,最终得到学生的综合等级绩点。最终设定阈值对学生血液状况进行预警,这个预警方式是阶段性的,随着数据的更新实时进行更新,实时进行预警。系统将对学业预警的情况反馈给辅导员以及班主任,提醒他们留意学生的学业情况^[4]。

(3) 图书阅读画像:图书借阅和图书馆进出情况也是学生在校学习情况的另一种形式,通过对学生进出图书馆和图书借阅本书,图书类别进行刻画,反应学生的阅读学习情况。

(4) 综合画像:通过以上画像的刻画,已经得到了学生在校期间各个方面的画像,在此基础上将形成学生的完整行为画像,最终我们将评估学生的综合情况,做出预警。

3 结论及展望

本研究结合汉中职业技术学院各业务系统数据,对数据进行清洗整合等预处理,再对得到的数据使用数据分析的手段进行研究,构建了学生行为画像的构建方法及流程。随着技术的进步,数据的增多,我们还将探索更精细的数据分析方式和行为画像刻画研究,结合其他方面的研究,探究职业院校学生和其他高等院校学生行为的异同等。

参考文献:

[1] 王军,何旭.基于数据挖掘的高职学生行为分析模型构建与研究[J].现代信息科技,2020,4(18):119-121.

[2] 刘新鹏,高斌.利用Python和Pandas进行学生成绩处理[J].信息与电脑(理论版),2020,32(07):41-43.