基于多源异构数据的“新高考”平行志愿智能填报系统的研发

张彦俊[1]　简卓然[2]　陈丽煌[3]　张继涛[4]
中山职业技术学院　广东　中山　528404

[摘　要]本文是在实行平行志愿投档录取模式的背景下，根据平行志愿投档录取模式的特点、规律等，结合“新高考”的政策、要求等，运用大数据技术，基于多源异构的高考大数据，开发出一套既适合平行志愿投档录取模式，又适应“新高考”政策下平行志愿智能填报综合服务系统，面向所有实行平行志愿及“新高考”省份参加高考的考生或家长，提供“新旧高考”政策解读、“新旧高考”平行志愿智能填报、大学学业规划、职业生涯规划等咨询服务活动。

[关键词]多源数据；高考志愿；系统开发与应用
【DOI】10.12252/j.issn.2096-6261.2021.08.1471

一、研究背景

新一轮高考制度的改革正带来高考志愿智能填报思维方式的转变。作为中央部署全面深化改革的重大举措之一，关于考试招生制度改革的实施意见。目前高考改革方案的亮点在于两方面，一是“2015年起推行自主招生安排在全国统一高考后进行”；二是“创造条件逐步取消高校招生录取批次。改进投档录取模式，推进并完善平行志愿投档方式，增加高校和学生的双向选择机会。2015年起在有条件的省份开展录取批次改革试点。”原因很简单，如果录取制度不进行深层次改革，在当前分批次按计划集中录取制度框架下，文理不分科、英语一年两次考，并不能改变基础教育的应试格局。因此，在当前分批次按计划集中录取制度框架下，研究各批次录取投档的规律，对于考生和家长来说，具有十分重要的现实意义。

目前，共有三批的省份实行了新高考。第一批新高考（2014-2017年），上海市：实行“3+3”和“6选3”，即语文、数学、外语三科必考；浙江省：实行“3+3”和“7选3”，即语文、数学、外语（英语、俄语、德语、法语、日语等）三科必考。第二批新高考（2017-2020年），北京、天津、海南、山东等4个省份，实行“3+3”和“6选3模式”，即语文、数学、外语三科必考，学生要从物理、化学、生物、政治、历史、地理等6个学科中，选出3个学科作为高考选考科目。第三批新高考（2018-2021年），河北、辽宁、湖北、湖南、重庆、江苏、福建、广东等8个省份作为第三批“新高考”试点省份，推行了 “3+1+2”的新高考模式。第四批新高考（2021-2024年），黑龙江、吉林、安徽、江西、广西、贵州、甘肃等7个省份，将成为第四批“新高考”试点省份，推行“3+1+2”的新高考模式。

二、研究内容

在实行平行志愿投档录取模式的背景下，根据平行志愿投档录取模式的特点、规律等，结合“新高考”的政策、要求等，运用大数据技术，基于多源异构的高考大数据，开发出一套既适合平行志愿投档录取模式，又适应“新高考”政策下平行志愿智能填报综合服务平台---“大学导航”，面向所有实行平行志愿及“新高考”省份参加高考的考生或家长，提供“新旧高考”政策解读、“新旧高考”平行志愿智能填报、大学学业规划、职业生涯规划等咨询服务活动。

1.面对多源异构的高考大数据采集、处理

项目采用数据爬取的方法对各省往年和今年高考大数据进行采集、处理与建库，在数据爬取中，多种数据获取渠道与多种数据渲染形式为高考大数据自动采集带来挑战，项目拟通过构建“中间件”，将这些多源异构数据进行自动采集并统一为同一类型的数据，具体包括以下几种格式：

（1）PDF 表格提取

采用技术为 PDFplumber：PDFplumber 找到可见或不可见的表格线；PDFplumber 根据表格边界线确定表格之间的交点；根据交点，确认包围的最小单元格；把单元格融合在一起，孵化检测出表格对象。

（2）Excel 文档表格提取

采用 Python 中的依赖库pandas，基于read_excel（url）方法，读取页面表格，再采用建立csv文件存入，即可实现页面表格提取归一化。

（3）图片表格提取

选择特别的图像元素采用膨胀腐蚀方法提取表格结构，将表格结构继续细化，提取完全的表格图像；提取图像后，定位表格单元格，将其切分为小单元结构。

（4）HTML/XML 数据提取

运用正则化、JSON 串、DOM 树、Webmagic 等工具对HTML、XML 等格式网页源码进行提取。

2.建立核心基本库

（1）“批次控制线”

运用所学数据爬取方法，在老师的指导下，对各省（市）当年往年的高考“批次控制线”大数据进行采集、整理、分类、加工、处理、制表，从而建立起各省（市）高考“批次控制线”数据库。

（2）“一分一段表”

运用所学数据爬取方法，在老师的指导下，对各省（市）当年往年的高考“一分一段表”大数据进行采集与网络爬虫、清洗与存储挖掘、分析与应用处理，从而建立起各省（市）高考“一分一段表”数据库。

（3）“最低投档分”

运用所学专业知识，在老师的指导下，对各省（市）当年往年的高考“最低投档分”大数据进行采集与网络爬虫、清洗与存储挖掘、分析与应用处理，从而建立起各省（市）高考“最低投档分”数据库。

3.“大学导航”高考志愿智能填报系统研究开发与完善

（1）“大学导航”高考志愿智能填报系统数据结构建立

运用所学专业知识，在老师的指导下，以上述“批次控制线”、“一分一段表”、“最低投档分”为依据，利用高考数据运算数学模型，在前期开发的“大学导航”高考志愿智能填报系统1.0版的基础上，进一步完善系统数据结构，为2.0版系统开发升级搭建框架。

（2）“大学导航”高考志愿智能填报系统研究与开发

高考志愿智能填报系统运用所学的《WEB 前端应用程序开发》、《数据库设计与应用》、《程序设计与应用》、《Spring Boot 企业级开发》、《JavaScript 高级程序设计》等专业课程知识与技能，在老师的指导下，围绕高考志愿智能填报准备、高考志愿关键信息、高考志愿“海选”方

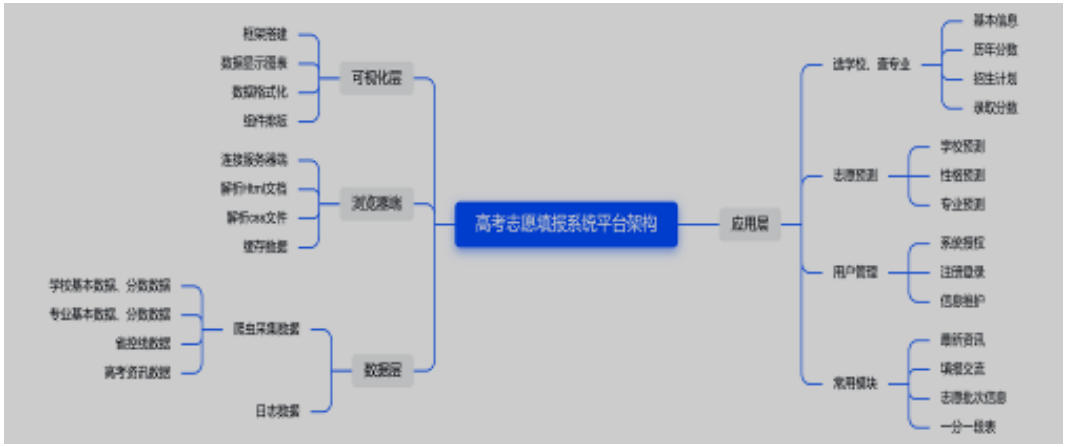

图1 高考志愿智能填报综合服务系统架构图

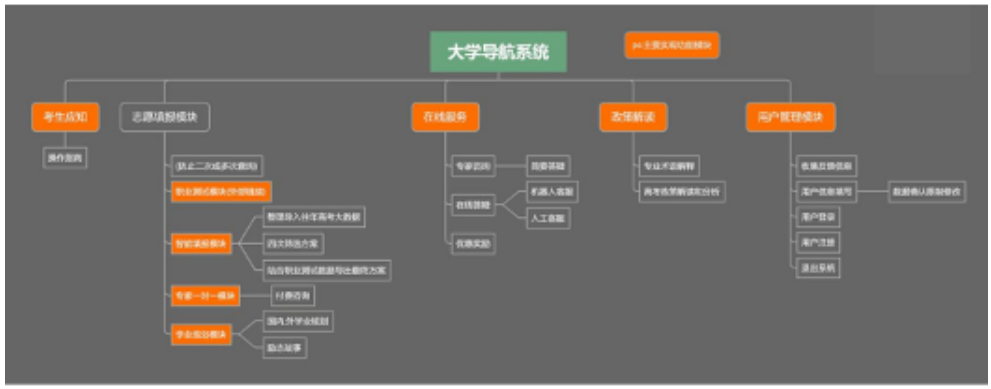

图2 高考志愿智能填报综合服务系统功能模块图

案、高考志愿“限选”方案、高考志愿“优选”方案、高考志愿“终选”方案等步骤，在前期开发的“大学导航”高考志愿智能填报系统1.0版的基础上，进一步研究、开发、完善、升级系统的2.0版各项功能。高考志愿智能填报综合服务系统架构，如图1所示，高考志愿智能填报服务系统功能模块设计，如图2所示。

三、系统功能

1.高考大数据查询

在“智能填报”模块中，系统为考生提供“批次控制线”、“一分一段表”、“最低投档分”等高考大数据查询服务。

2.系统登录

考生凭账号和密码登录进入本系统，可填写“批次”、“科类”、“参考年份”、“高考分数”等关键信息。

3.查看“海选方案”

系统以“基准志愿”为参考，按照“冲一冲”、“稳一稳”、“保一保”的平行志愿设置原则，为考生提供若干组院校志愿的“海选方案”，考生可根据需要从中筛选出符合自己高考志愿意向的院校志愿。

4.查看“限选方案”

考生可根据“职业意向”、“专业方向”、“院校性质”、“地理位置”、“办学条件”等个人意愿，按照院校代码在招生之友、招生手册等参考资料中查询、勾选符合个人意愿的院校志愿（数量可超过规定的数量），形成高考志愿的“限选”方案。

5.查看“优选方案”

考生依据招生之友或招生手册中各院校及其专业（组）的招生计划等信息，在每一个院校志愿组中填写1-5个自己中意的院校专业（组）（专业代码、专业名称等相关信息，并在前面进行勾选，形成高考志愿“优选”方案。

6.查看“终选方案”

考生对每一个院校志愿、专业志愿进行纵向、横向的比较，根据“职业兴趣”、“专业方向”、“就业方向”、“学科优势”、“师资优势”、“院校性质”、“办学地点”、“办学历史”、“办学条件”、“收费标准”、“奖助学金”、“晋升通道”、“转专业政策”等因素，经过再思考、再分析，对“优选”方案中的每一个院校志愿、专业进行再筛选、再选择，排除不满意的志愿，保留最心意的院校志愿、专业志愿，形成高考志愿的“终选”方案。

参考文献

[1]冯成.多源异构数据融合关键技术研究[D].北京邮电大学，2020.

[2]邓丽强.多源异构数据检索系统的研究与设计[D].桂林电子科技大学，2020.

[3]李少波，多源异构制造大数据融合分析与云平台关键技术研究及应用.贵州省，贵州大学，2020-05-03.

基金项目：2021年度中山职业技术学院大学生创新创业训练计划项目《基于多源异构数据采集的高考志愿填报综合服务平台研发与推广应用》。

作者简介：

1.张彦俊，1979，女，黑龙江哈尔滨人，原黑龙江大学数据科学与技术学院博士、副教授，现中山职业技术学院大数据技术专业主任、学科带头人，研究方向：大数据、人工智能、职业教育。