**教育部科技发展中心高校产学研创新基金**

**-“新一代信息技术创新项目”申报书**

**项目名称：高考志愿填报及招生大数据实时分析与推荐算法研究**

**负 责 人：史广顺**

**学校名称：南开大学**

**所在院系：人工智能学院**

**项目类型：重点项目√ 一般项目□**

**填报日期：2019/1/10**

**教育部科技发展中心**

**2018年10月制**

**一、基本情况表**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **项目信息** | 申报项目名称 | **高考志愿填报及招生大数据实时分析与推荐算法研究** | | |
| 项目方向编号 | A02 | | |
| 申报项目类型 | 重点项目 √ 一般项目 □ | | |
| **负责人信息** | 姓 名 | 史广顺 | 性 别 | 男 |
| 出生日期 | 1978.01.20 | 最终学位 | 博士 |
| 技术职称 | 副教授 | 行政职务 | 硕士生导师 |
| 毕业学校 | 南开大学 | 毕业时间 | 2003.7 |
| 专业名称 | 人工智能 | 手 机 号 | 13920807749 |
| 电子信箱 | [904761615@qq.com](mailto:904761615@qq.com) | | |
| 通信地址 | 南开大学津南校区人工智能学院404 | | |
| **学校信息** | 学校名称 | 南开大学 | | |
| 主管部门 | 人工智能学院 | | |
| 学校类型 | √ 普通本科 （是否应用技术型： 是 □ 否 √ ） | | |
| □ 高职类 | | |
| **申请课题简介（不超过500字）：**  本项目研究目标为设计一个基于微信小程序的志愿填报系统。该系统能够综合分析根据当年其他考生的数据，有效反映用户高考志愿填报中的竞争激烈程度、并能够针对用户个人情况进行个性化的内容推荐。  该项目分为智能推荐模块和模拟志愿填报模块。用户填入基本信息，可以获取按风险大小排序的推荐学校列表。用户可以选择关注列表中的学校，以获取详细信息。  数据库里存储有当年的考生成绩分布和院校偏好，当用户进行模拟志愿填报时，算法根据当年其他考生的信息，动态返回用户的竞争力、成功率等。  智能推荐模块，当用户填入分数和排名之后，根据前五年学校线差，计算出各学校的风险值，根据预先划好的风险等级对学校进行分类，返回若干高风险的学校，若干低风险的学校，若干一般风险的学校。  模拟志愿填报模块，用户关注学校之后，填写模拟志愿。系统统计模拟报考了同一学校的考生人数，根据用户在这些考生中的排名位置，结合专业偏好，分析出用户各志愿的综合竞争力。 | | | | |

**二、项目组负责人及主要参加人员基本情况表**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **序号** | **姓名** | **性别** | **身份证号码** | | **学位** | **技术职称** | **项目分工** | | | | **签 字** |
| 1 | 史广顺 | 男 | 12010419780126819 | | 博士 | 副教授 | 项目总体指导 | | | |  |
| 2 | 张明昕 | 男 | 120105199708114851 | | 本科 | 无 | 前端设计 | | | |  |
| 3 | 刘逸伦 | 男 | 511002199808131519 | | 本科 | 无 | 接口设计 | | | |  |
| 4 | 万典 | 男 | 360111199806262139 | | 本科 | 无 | 后端算法 | | | |  |
| 5 | 骞晨旭 | 男 | 130582199909080017 | | 本科 | 无 | 数据处理 | | | |  |
| 6 |  |  |  | |  |  |  | | | |  |
| **总人数** | | **高级职称** | | **中级职称** | | **初级职称** | | **研究生** | **本科生** | **高职生** | |
| 5 | | 1 | |  | |  | |  | 4 |  | |

|  |
| --- |
| **二、项目研究的意义、国内外研究现状和预期成果** |
| 大数据，指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。近年来，大数据逐渐成为工业界，学术界乃至各国政府高度关注的热点。《自然》和《科学》等杂志相继出版专刊来讨论大数据带来的挑战和机遇。  这其中，教育逐渐被认为是大数据可以大有作为的一个重要应用领域，有人大胆预测大数据将给教育带来的革命性变化。大数据技术允许中小学和大学分析从学生的学习行为、考试分数到职业规划等所有重要的信息。许多这样的数据已经被诸如美国国家教育统计中心之类的政府机构存储起来用于统计和分析。 教育领域的大数据分析最终目的是改善学生的学习成绩。  但是改善学生的成绩并不是教育的全部内容。升学时不同考生对有关信息的掌握是有差异的。掌握信息较为充分的学生往往较掌握信息匮乏的学生处于比较有利的地位。利用大数据和智能推荐系统，考生可以有效节约填报志愿的时间成本和信息索取成本，让考生选择一个兼具个人兴趣和职业竞争力的专业。  现在是个信息爆炸的时代，如何在多而杂乱的信息中筛选出对于解决问题有帮助的有效信息，将有效信息呈现在用户面前，这些都是推荐系统的主题。  “推荐系统”一词的定义早在1990年代由美国的ReSnick等人提出，即推荐系统在综合用户意见的基础上给出有价值的意见输出给用户[1]。在1992年Goldberg等人提出了协同过滤的概念，即利用用户和系统物品交互的历史数据，找出用户的近邻用户集，通过近邻用户集来推测用户在别的新的物品上的评分来产生推荐物品[2]。2000年NEC研究员的Kurt等人为搜索引擎CiteSeer增加了个性化推荐功能[3]。其后国内的清华大学的卢增详等人提出混合智能推荐系统，实现个性化推荐[4]。到了2002年之后，推荐系统在国内关注度上升，在诸多网络领域有所应用。比如有2010年周珊丹等人提出了一种基于用户的协同过滤推荐系统，用于虚拟博物馆中为用户提供个性化的文物推荐[5]。近年来有Lei Zhang等人在文本信息挖掘中提出了一种深层协同神经网络模型，缓解了该领域的数据稀疏问题[6]。  对于高考推荐系统国内也有一些研究文献。许飞在2008年结合数据挖掘技术和Web服务技术，设计并实现了一个基于Web的高考志愿分析系统。殷员分等人在2009年对高考历史数据建立数据仓库，在对考生进行需求分析的基础上利用OLAP技术对高考数据进行多维分析，并运用数据挖掘技术。2012年肖灿等人使用SSH框架建立招生数据挖掘系统，给予学生院校和专业填报指导。2015年徐兰静将考生、高校分为推荐系统中的用户和项目，在该领域引入协同过滤的思想。2018年吴玲利用灰色预测对国家和高校分数线进行预测，同时引入人职匹配理论为考生推荐院校和专业[7]。  微信是现如今发展最成熟、使用人数最多的社交应用软件。微信小程序是一款借助微信平台开发的应用软件，具有无需下载、搜索应用便利、不占据内存空间等诸多优势。微信小程序的出现和应用丰富了微信的应用功能，满足了人们的实际需要。  预期成果：   1. 建立覆盖全国1200所本科院校，符合大数据应用规范的历年高考招生数据，细化到学科、专业、最低分、最高分、省份分布等。形成第一个全国可用的高考招生大数据资源包。 2. 小程序成功上线，以免费服务的形式，为高考考生和各大学招生办提供大数据分析及智能推荐服务，预期在2019年高考结束后吸引全国不少于100万考生使用，不少于100所本科院校使用； 3. 发表学术论文3-5篇 。 4. 申请专利及软件著作权3项。   参考文献：  [1]Dan Cosley, Steve **Lawrence**, David M.**Pennock**, **REFEREE**: an **open framework for practical testing of recommender systems** using ResearchIndex, Proceedings of the 28th international conference on Very Large Data Bases, p.35-46, August 20-23, 2002, Hong Kong, China.  [2] Gaul W,Schmidt-Thieme L.Rencommender systems based on navigation path features[A].International Conference on Knowledge Discovery and Data Mining. San Francisco, 2001.  [3]邓婕.基于匿名用户使用挖掘的个性化信息推荐研究[D].西安电子科技大学,2009．  [4]卢增祥，李衍达，路海明.基于多Agent混合智能实现个性化信息推荐[J].高技术通讯，2001，（4）：28-31.  [5]周珊丹，周兴社，王海鹏，等.智能博物馆环境下的个性化推荐算法[J].计算机工程与应用,2010，(19)：224-226．  [6] Lei Zheng, Vahid Noroozi, and Philip S.Yu.2007. Joint Deep Modeling of Users and Items Using Reviews for Recommendation. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM’17). ACM, New York, NY, USA, 425-434.  [7]吴玲. 高考志愿智能填报系统的设计与实现[D].云南财经大学,2018. |
| **三、课题研究目标、内容、研究方案和科学问题**  详细说明项目的研究目标和主要内容；为达到研究目标，拟采取的措施和技术路线、进度安排、拟达到的技术指标、提交成果方式、拟解决的关键科学问题等。 |
| **1、研究目标**  本项目研究目标为设计一个基于微信小程序的志愿填报系统。该系统能够综合分析根据当年其他考生的数据，有效反映用户高考志愿填报中的竞争激烈程度、并能够针对用户个人情况进行个性化的内容推荐。  **2、研究内容**  该项目主要的研究内容其一在于海量数据的收集、清洗以及整理；因此选择阳光高考等网站爬取学校及专业录取信息。同时还要收集用户的志愿填报以及偏好信息，通过大量用户的关注信息得到志愿的真实竞争水平。其二是志愿填报系统的功能设计。系统提供给考生“选择关注院校—模拟填报—评估结果”的模拟填报流程，考生可根据评估结果明确志愿填报方案的优劣。同时，针对考生的关注院校以及偏好专业，还要有个性化的内容推荐。此外，对于专业选择，还能够通过选择“门类、一级学科”获取各个学校该一级学科下本科专业近年来分数线等招生情况以及教育部学科评估结果。  目前微信用户数量巨大，微信小程序有着较好的开发前景，因此该项目载体选择微信小程序，结合Python语言进行系统开发。  **3、推荐策略设计与分析**  **3.1 院校/专业推荐策略设计**  当用户输入考生信息后，首先根据考生成绩与一分一档表查出考生的省内排名，从院校数据库中检索，如果考生排名高于该院校五年录取排名，则为低风险结果；若考生排名高于该院校三至四年录取排名，则为一般风险结果；若考生排名只高于院校三年以下录取排名，则为高风险结果。同时考虑考生排名和院校近五年平均录取排名的比较。若考生排名高于院校近五年平均录取排名，则将风险下调一档。在考虑排名的同时也要综合考虑招生计划的变化带来的影响。  在输入考生信息后选择“偏好专业”或“偏好学校”。偏好专业需选择偏好的专业大类。偏好学校则在推荐结果中按学校排名择优推荐。偏好专业则按所偏好的专业学科评估结果推荐。  从低风险结果中择优选择前N所院校加入推荐结果；从高风险和一般风险院校中选择五年录取排名波动较小的结果中选择M所院校加入推荐结果。将推荐结果进行排序。此为院校初步推荐规则。    在完成初步推荐之后，根据用户在推荐列表中点选的“关注”确定志愿填报界面的下拉菜单院校。  **3.2 模拟志愿填报**  进入模拟志愿填报界面后，用户可选择下拉菜单中的关注院校，也可手动输入院校。选择院校之后，根据该院校的学科评估结果，综合专业录取排名（与推荐院校类似的策略，但院校推荐的策略相比，因为专业录取人数相对更少，因此要增加招生计划变化所造成影响的比重）给出专业推荐结果。用户也可手动修改志愿专业。  在进行模拟志愿填报之后，程序会给出对当前志愿填报情况的评估结果。如图所示，考生能否成功录取取决于该院校/专业的招生计划和填报该院校/专业的总人数以及考生在同志愿考生排名中的情况。程序统计当前所有用户对当前志愿的填报情况，根据比例关系预估该考生的实际同志愿中的位次，进而估算出报考成功率。  关注该院校的人数  该院校计划招收人数  人数/排名  考生位次（成功率高）  考生位次（成功率低）  **3.3 院校/专业推荐策略分析**  历年院校录取分数线可能会随着各种因素有较大波动，但是录取排名相对稳定。同时为了排除录取排名可能的波动，考虑近五年院校录取排名作为风险评估的依据。择优选择的推荐结果可以大大减少推荐范围，有利于考生选择。对高风险和一般风险院校的操作能一定范围内扩大推荐范围，这样的话推荐结果中“冲一冲”、“稳一稳”和“保一保”的结果就会相对均衡，有利于下一步模拟志愿填报的进一步细化评估。  专业推荐的步骤大致与院校推荐相同，进一步缩小选择范围，同时学科评估权重提高，有利于考生在当前分数内选择最合适、水平最高的专业。  此外，还应考虑因用户行为产生的误差。   1. 存在考生在实际填报时并没有按照模拟填报时的志愿填报的情况。 2. 存在因用户数量不足而程序所收集信息无法反映真实志愿填报的情况。   以上两点均会造成成功率预测的误差。  针对第一点，可在程序界面内加入提示，提示用户尽量按照实际填报情况进行模拟，才能有更为准确的结果；还可采用用户回访的方法，对预测方法进行调整。  针对第二点，应考虑近五年各院校/专业的报录比，根据报录比和当年招生情况预估出当年该院校/专业的关注度期望，用这个期望值对实际收集到的关注度信息进行校正。  院校推荐流程图  否  是  是  否  在一分一档表中检索考生排名  院校近五年录取排名均低于考生  院校三到四年录取排名低于考生线差  低风险  一般风险  高风险  选择近五年录取排名较稳定的院校  院校推荐列表  **4、系统架构设计**  **4.1 三层B/S结构**    三层B/S结构图  **4.2 系统功能设计**  对于软件使用者和软件开发者而言，整个系统必须实现以下三大功能：  1.初步推荐学校  （1）实现推荐算法。具体要做的，是实现输入考生的信息，根据往年的分数线等给出初步推荐关注的院校与专业。  2.数据库的建立（往年以及当年）  （1）明确需要收集的当年数据有哪些，建立当年的考生信息和学校招生信息数据库。  （2）建立各校往年分数线数据库。  3.智能志愿评估、个性化风险排查  （1）实现志愿智能评估功能，包括明确从当年数据库中需要抓取的信息，以及推算志愿成功率的算法，反馈报告应包含的内容。  （2）实现核心功能要经过各方研究，关键是搞清楚从学生和学校端都需要获取哪些数据，以达到最准确的评估。  **4.3数据库表结构设计（ER图）**    **4.4处理流程设计（流程图）**      **4.5数据模型设计（数据结构）**    **5、评估验证标准**  可利用已知数据进行验证，如利用五年已知数据（2012-2016）预估新一年已知结果（2017）。可虚拟若干考生信息（数据量应大，分布应广泛，考生成绩的分布应近似实际高考分布情况，才能更好地反映程序的真实使用情况），输入程序，根据推荐列表以及推荐专业直接进行填报，对比评估结果和新一年实际招生情况。对同一所院校来说，若某分数段考生评估结果为一较高成功率且这部分考生成绩确实大部分在实际招生范围内，则可初步认为有效。  在初步认定有效之后，统计所有被推荐到该院校的考生人数和实际被录取的考生人数，将比值与成功率比较，若两值相近则可认为结果符合标准。  **6、数据资源准备**    **6.1业务数据资源**  1. 省份历年信息表：省的一本和二本分数线  2. 学校信息表：ID，名称，所在省份  3. 高校招生总表：文理科类、历年招生总名额、历年各省招生总名额、历年各专业招生总名额、历年各省各专业招生总名额、各专业历年各省招生分数（最高、最低、平均）、各专业历年招生人数排名（最高、最低、平均）  **6.2过程数据资源**  1. 用户信息表： ID，省份，文理，总分，省内排名  2. 用户志愿表：ID、模拟志愿中的六个学校、每个学校的六个专业  3. 统计数据表: 各学校各专业各省关注人数、各学校各省各专业模拟志愿填报人数。 |
| **四、已有工作基础和课题研究支撑条件**  详细说明团队在项目研究方向中已有的研究成果；以及现有的项目研究支撑条件（请对照《申报指南》表2填写）。 |
| 1. **前期研究和准备**（相关代码已放至https://github.com/lunyiliu/College-Entrance-Application-Helper-）   **1、数据准备**  （1）已搜集1600所学校近6年在各省份文理录取平均分、最低分和最高分  （2）已搜集1600所学校各专业近6年在各省份文理录取平均分和最高分  （3）已搜集1600所高校名称、所在地、隶属、类型、风貌图片等  （4）为了获得专业对应的一级学科编写了用于众包标注的小程序   1. 已搜集部分省份部分年份高考一分一档表   **2、推荐算法部分**  已经完成学校推荐功能的编写。考生输入基本信息和当年高考分数，可以实现按照高、中、低三档风险推荐学校。考生可以选择关注其中的学校，以了解自己在所有关注同一学校的人中的竞争力。  图1志愿推荐关注页面  C:\Users\lenovvo\Documents\Tencent Files\932866784\Image\Group\thumbnail\898899d4-14e9-11e9-ad48-5452007beaa8Original\`GQ_G[_E}B`7B(N[LFL%@%7.jpg  图2 高校详细信息  **3、前端界面**  前端界面设计工作展示：  图3 基本信息页面    图4 模拟志愿填报界面   1. **接口设计与编写** 2. 数据流向：微信前端——>PHP脚本——>python脚本——>数据库 3. 处理流程：     **二、项目研究支撑条件**   1. 本项目依托的实验室已有完整的大数据技术研究与实践的科研条件，围绕Hadoop构建了一体化的大数据应用服务基础技术体系，足够支撑本项目的研发工作； 2. 本项目需要采集并积累过去三年内全国所有本科院校的高考招生数据以及详细到各专业各学科的高考招生录取数据，经过过去半年的努力，本项目已经基本完成大部分历史数据的采集和清洗工作，正在进行更加完善的数据采集； 3. 本项目需要的云服务器环境、移动应用接入条件等均已具备，满足项目研发所需的 条件。 |
| **五、课题经费使用计划、科研条件投建计划**  详细说明项目资助经费的使用计划；以及未来两年内在云计算、大数据、人工智能和虚拟现实相关领域的科研条件投建计划。 |
| **一、课题经费使用计划**  本项目拟申请经费5万元，使用计划如下：   1. 数据采集、清洗及录入费用：2万元。将过去三年中国1200余所本科院校的高考录取数据及详细数据（学科、专业、最高分、最低分等）进行采集、清洗和录入，形成稳定可用的数据资源包； 2. 云服务器、小程序云服务及认证费用：5000元，用于项目的运行环境部署、ICP备案、云服务开通及认证费用； 3. 信息资料费：10000元，发表学术论文、申请软著及专利费用； 4. 软件开发及设计费用：1.5万元，包括界面设计、美术设计、付费技术服务等。   **二、科研条件投建计划** 该项目依托南开大学人工智能学院机器智能研究所的科研环境展开，在科研条件方面已有良好基础，针对本项目的进一步科研投建计划如下： 1）大数据核心算法及关键技术体系的研究和应用 项目团队将围绕基于协同过滤算法、基于深度学习算法的两套方法体系，对各类算法进行研究和实践，依托实验室的现有环境，对大数据算法在高考志愿分析和预测方面的应用技术进行研究；  2）基于数据可视化的高考招生数据服务技术的研究和实践 截至目前，各高校的高考招生过程依然使用传统的信息手段和数据统计方法，难以提供实时有效的报考志愿相关的数据服务，本项目将重点围绕“报考志愿的动态变化”这一数据变化过程，使用先进的数据可视化技术实现动态灵活的招生数据服务。  3）基于小程序的移动互联网应用技术研发 本项目可为高考考生、高校招生办、各类中学的校方管理者提供服务，因此需要使用最易于传播的小程序进行应用开发，诸多关键技术在小程序环境下的应用集成是需要解决的技术重点之一，本项目将对此进行深度研究和实践。 |

**六、项目负责人承诺书**

|  |
| --- |
| 本人承诺申报书填写的各项内容属实，没有知识产权争议。如获准立项，我遵守有关管理规定，按照申报书填报的研究内容和时间如期完成研究任务，自觉接受检查与监督管理。资助项目获得的知识产权由资助方和项目承担单位共同所有。  项目负责人（签字）：  年 月 日 |

**七、申报单位推荐意见**

|  |
| --- |
| （请填写学校对本课题的推荐意见。申报书内容是否属实；学校是否能够保证为该项目的研究提供条件与时间；是否同意对项目负责人在研究周期内完成预期工作提供信誉保证。）  校长(签章): 学校公章：    年 月 日 |