暑期学校实验项目：高考志愿填报助手

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 小组名称 | AI算法C组 | | | | | | |
| 姓 名 | 黄一凡 | 专业 | 人工智能 | 班级 | 1 | 学号 | 09118119 |
| 实验时间 | 2020.8.31-2020.9.23 | | 指导教师 | 孔祥龙 | | 成绩 |  |
| **一、实验背景和目的**  当前社会，高考志愿的填报对每一名考生的影响都十分深远：不同层次的学校一定程度上决定着未来发展；而不同的专业又决定这今后所处的行业。与志愿填报的重要性相矛盾的是，许多考生或是考生家长并没有意识到其重要性。同样的，虽然随着信息化事业的飞速发展，但不可否认的是，如今仍然存在着难以想象的信息差，这也导致了许多考生因为信息的缺位而选择了错误的填报策略，对其今后的人生道路造成了不可估量的影响。  为此，在本次短学期的任务中，我们需要设计一个基于Web的高考志愿推荐系统。此系统主要基于各高校近三年的录取数据，从大数据等角度对考生的信息进行分析，从而较为合理的为其推荐高校以及专业。  在此系统的帮助下，我们可以在一定程度上帮助考生更加合理的选择专业，让更多的考生在填报志愿时不再茫然与彷徨。 | | | | | | | |
| **二、小组任务和个人任务**  小组任务：在与另外两组AI算法组充分交流后，确定了我们小组的任务为专业的推荐，即在AI算法A组推荐的学校的基础上，进一步进行专业的推荐。确定目标为专业的推荐后，我们小组首先进行了分析讨论，并得到了初步需求文档（详见码云仓库）。接着我们根据需求文档以及孔祥龙老师的建议，对初步需求文档作了适当修改，且进一步得到了小组任务的具体分工与安排，具体见下图：    图一：小组任务具体分工  个人任务：首先，作为本组的组长，我和组员一起讨论确定了初步设计文档以及小组的具体分工安排（见图一）。其次，我作为第三小组的组长，完成了基于用户能力的专业推荐，这里包含了两部分：用户能力的量化分析以及根据此来对形形色色的专业进行推荐。其中，我负责的是基于竞赛经历的能力分析、由能力的分析结果来对专业进行推荐以及最后的整合工作。 | | | | | | | |
| **三、个人任务需求分析**  我的个人任务大体可以分为三部分：基于用户的竞赛经历分析其能力、根据能力对专业进行推荐以及最后的整合工作。下面我也将分别从这三部分进行需求分析。  首先是基于用户的竞赛经历分析其能力。这里非常感谢孔老师为我们提供的宝贵数据，即各专业学群所需的能力，部分数据举例如下图所示：    图二：部分数据  如图所示，根据老师提供的数据，我们首先可以总结提取出能力的类型，从而对用户不同  维度的能力进行量化分析。在这一部分中，我们需要根据用户提供的高中时代的竞赛经历  以及所获奖项来对其能力进行分析，大概可以再细分为以下几个工作：   * 分析竞赛所对应的具体能力 * 分析不同竞赛不同奖项对应的针对能力的权重 * 分析得到用户的能力   其次是根据能力对专业的推荐，根据第一部分，我们现在已经有了对于用户能力的量化分析结果，接下来，我们需要根据此结果对用户推荐合适且匹配度高的专业。如某一位用户获得了数学竞赛的相关奖项，那么我们推荐的专业即为数学相关的，如计算机、数学、物理等。  最后是整合工作，我们从三个方面分析了用户的能力：单科成绩、竞赛经历以及调查问卷。接下来需要对这三方面的能力综合考量，再结合用户的总成绩，最终给出对用户的专业推荐。后续还需要和前端以及数据库的衔接整合。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  依据上述的需求分析，我接下来的实验也将从这三个方面展开，下面我将分别具体阐述我的实验过程。  第一部分是根据用户的竞赛经历量化分析其能力。在这一部分中，首先需要收集竞赛的种类，如果收集的竞赛种类不够全面或者不够有代表性，那么会对能力的量化分析造成很大的影响。考虑到这一点，我选取了教育部在2020年认证面向高中学生的28项竞赛，并一一分析其对应能力，部分结果如下图所示：    图三：部分竞赛数据  如图所示，根据不同的竞赛，我收集了其奖项设置以及分析其对应能力。据此，我可以根据用户的竞赛经历，分析其能力。这里我遇到了另一个问题：如何评估一个奖项的含金量。举例来说，全国中学生数学竞赛的国家一等奖和省三等奖明显反应出能力的不同，国家一等奖表示数学能力及其强大，而省三等奖只能说明其数学能力较强。为此，我引入了能力的等级，该等级的范围为1~5，等级越高，也就说明该能力越强，我们用ri来表示某个能力i的等级。由于不同竞赛的奖项设置不同，我不能简单的根据奖项名称来对能力等级进行定义，因此，我使用了一种归一化的思想，即对于一个特定的竞赛，其最高奖项对应的能力等级为5，最低奖项对应的能力等级为1，其余奖项也相应对应一定的能力等级，相关代码如下：    在这个函数中，输入竞赛名称和所获奖项，返回其对应的能力等级r。  有了各竞赛各奖项对应的能力以及能力等级后，接下来我需要统计其与各学群的匹配度。在这一步骤中，我采用了加权累计的想法：第一步将各学群的初始分值均设为0；第二步考虑用户提供的竞赛以及所获奖项，将竞赛对应的能力加权累加到各学群中，具体定义如下：  其中，表示某个学群j所对应的分值，表示与其能力相匹配的竞赛奖项的能力等级。举例来说，若某用户竞赛经历只有一项为“丝路国家青少年国际摄影大赛”一等奖，一等奖为该竞赛的最高奖，则其对应的能力：科学能力和操作能力的分值均为5，因此，对于各学群，若其所需能力中含有科学能力或是操作能力，其分值S也就相应累加。具体代码如下：    在该函数中，输入用户的竞赛经历，通过调用上述的计算能力分值的函数，最后返回各学群的得分。  得到各学群的得分之后，由于不同学群的得分差异巨大，不便于后续的整合工作，因此我加入了一个归一化函数，将所有学群的得分归一化至0~1之间，并将此结果作为最终的学群推荐度，具体代码如下：    至此，我实现了根据某一用户的竞赛经历，得到各学群的推荐度。  最终我需要实现的是专业推荐，而现在只得到了对于各学群的推荐度，还需要实现学群到专业的转换，实现思路如下：    图四：实现思路  如图四所示，前文我已经实现了由竞赛对能力的分析，与此同时，我们小组内的其他成员分别实现了由调查问卷以及单科成绩来对能力进行分析，也就是说，我们已经完成了得到了用户的能力，并可以据此来推荐相关学群。  然而，最终我们需要推荐的是相关的专业，所以现在需要将这三部分整合，并加入总分的影响，最终得到推荐专业，这也就是我的第三部分任务。在这一部分中，我首先确定了各部分输入的数据格式，以便于和前端组进行对接，数据格式举例如下：    其中，gradelist为用户的各单科成绩列表；exp为用户的竞赛经历以及所获奖项列表；answer为用户填写的调查问卷的结果列表。目前，已经得到了分别通过竞赛经历、调查问卷和单科成绩三个角度的学群推荐度，且均通过0-1归一化至0~1之间，我赋予其相同的权重，累加便可得到总的对学群的推荐度，具体代码如下所示：    通过该函数，可以得到总的学群推荐度。  有了学群推荐度还远远不够，我们需要的是专业的推荐度。这里就需要用到我们组内其他成员语义消歧的工作：通过语义消歧，我可以了解每一个专业对应的学群（首先对应各个专业的一级学科，接着用一级学科对应各学群），这样，便可以实现由学群的推荐得到专业的推荐。  最后，我们小组的其他成员将基于成绩的专业推荐与基于能力的专业推荐进行了融合，得到了最终的项目。 | | | | | | | |
| **五、实验结果与分析**  首先我对基于竞赛的专业推荐结果进行展示：    如图所示，假设某一用户获得了全国中学生数学竞赛的国家二等奖以及中国青少年机器人竞赛的二等奖，显然，这两项竞赛属于理科类竞赛，要求选手的科学思维和动手思维。相应的，其对应学群推荐度也展现出很强的倾向性，如工程学群、咨询学群和数理化学群优先推荐，而一些相对文科的学群的推荐度则较低。    若更换了竞赛经历，当某一用户的竞赛经历为全国中小学生绘画书法作品比赛的一等奖以及中国日报社“21世纪杯”全国英语演讲比赛的二等奖时，可以看出，此时的竞赛经历明显偏向文科和艺术。而相应的学群推荐也发生了改变，不同于以往，这次社会与心理学群、教育学群和管理学群的推荐度很高，而上次推荐度很高的工程学群、咨询学群等推荐度有了很明显的下降。显然，此时的学群推荐偏向的是文科与艺术方向。  从上面两次不同竞赛经历之间的对比可以看出，我实现的基于竞赛经历的专业推荐有着较高的合理性和科学性，实现了根据不同用户的不同竞赛经历量化分析其能力，进而对学群专业进行推荐。  然而这一部分还是存在许多不足之处，主要有下面两点：  第一，考虑的竞赛不够全面。由于时间原因，我初步只考虑了教育部认证的少许竞赛，虽然说确保了权威性和代表性，但不可避免地遗漏了许多竞赛，而这就导致了一些有相关竞赛经历的用户无法得到全面的分析与推荐结果。在初步需求文档中，我设想的是通过正则表达式以及自然语言处理自动提取用户所输入的竞赛经历，然而这也会导致一些缺乏权威的竞赛进入，从而导致了该系统的不准确，因此最后我还是采用了提前规定竞赛的方法，后续需要对竞赛内容做进一步的增添。  第二，在基于竞赛的专业推荐中，主观性较强，缺少客观的分析与验证。比如说在分析竞赛所需的能力时，采用的是根据生活经验以及主观判断的方法，难免会因为知识的缺乏与疏漏存在相应的错误。  问题的存在是难免的，但是从上面的结果展示中还是可以看出基本的科学性与合理性。后续，还需要考虑这两个重要的不足，对基于竞赛的专业推荐做进一步的修改与完善。 | | | | | | | | |
| **六、实验总结与心得体会**  经过这次实验，我有了许多的收获与感触：  首先，我们班级的全体同学第一次以一个公司的身份参与了一个项目。这种感受是十分新鲜的。不同于以往的代码编写，我们都是作为单独一个人在工作，写的代码只需自己理解就可，也很少需要与他人的沟通。而这次暑期短学期的课程则大不相同，每一个小组相当于一个小部门，有着自己的任务，还需要时刻保持与其他小组的沟通，以完成项目的衔接。而沟通则恰恰是最为关键的部分，颠覆了我过去的任务。我们小组在沟通方面就吃了很大的亏，由于缺乏相关经验，在项目前期，我们小组并没有及时与前端组以及数据库组进行沟通交流，而是想着先完成自己的任务，最后再连接数据库与前端。事实证明了我们的错误，由于前期的缺少沟通，在写代码时也没有考虑相关数据库和前端的接口。最后需要与前端数据库连接时，已经为时已晚，大量的代码需要修改，前期也做了许多无用工作。  其次，我作为组长，在完成自己任务的同时，也需要对全组的任务进行统筹与分工。在这次的项目中，考虑到我们小组需要完成的任务主要有三部分：基于成绩的专业推荐、基于能力的专业推荐和语义消歧，因此为了便于组内分工沟通，我在项目开始时便又在组内分出来三个小组，分别负责一块。这也在项目开展中提供了很大的便捷。  最后，我想感谢孔祥龙老师的辛苦付出。在项目初期，孔老师分析了我们的初步需求文档并建议我们先从几个方面入手，不需要追求多而全。而后期我们的实践也证明了该建议的正确：老师建议的几个方向都较为容易实现，且由于短学期时间较短，若一开始没有对初步需求进行精简，我们组将很难完成全部的任务。孔老师在项目进行中也给与了我们很多帮助，为我们提供了十分宝贵的数据，也为我们解决了许多困难，这里想再对孔老师说一声谢谢！ | | | | | | | | |

2020年9月制