基于协同过滤的高考志愿推荐系统①

徐兰静, 李 珊, 严 钊

(南京航空航天大学 经济与管理学院, 南京 211100)

摘 要: 近年来信息过载问题的出现使得个性化推荐技术应运而生, 其中协同过滤推荐技术通过在用户和信息之间建立联系, 被广泛应用于电子商务各个领域. 而在高考志愿填报领域考生也存在无法高效的从诸多高校中选取适合自己的高校这一“信息过载”问题. 为此, 可以将协同过滤思想应用到高考志愿填报这一新领域, 将考生看作是推荐系统中的用户, 高校看作是系统中的项目, 通过分析历年的考生志愿填报相关数据, 从构建用户属性矩阵, 查找邻居用户和产生推荐三个过程进行详细描述, 并对实验产生的推荐结果进行分析, 说明了推荐系统的有效性, 也为进一步的研究工作奠定基础.

关键词: 协同过滤; 高考志愿; 推荐系统

College Entrance Examination Voluntary Recommendation System Based on Collaborative Filtering

XU Lan-Jing, LI Shan, YAN Zhao

(College of Economic and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 211100, China)

Abstract: Information overload problem in recent years makes the personalized recommendation technology arise, the collaborative filtering recommendation technology by establishing contacted between the user and the information has been widely used in every field of e-commerce. And in the field of the college entrance examination voluntary students also have the “information overload” problem, which means they cannot choose the suitable college from many colleges efficiently. Therefore, the idea of collaborative filtering is applied to this new field, take the students as users and colleges as the items in the recommendation system. By analysing students’ voluntary reporting relevant data from the previous year, three processes of building user attributes matrix, finding the neighbor users and generating recommendation are described in detail. The recommendations results of the experiment show the effectiveness of recommendations systems, and it lays the foundation for further research work.

Key words: collaborative filtering; college entrance examination voluntary; recommendation system

随着高考招生工作信息化的不断深入, 积累了大量有用的高考志愿相关数据信息. 在大量的历史录取相关数据信息里面蕴含了丰富的决策信息, 如何有效地利用这些信息辅助考生填报志愿是考生、家长、学校及招生管理部门都关心的问题[1], 也是当前招生考试业务信息化研究的热点问题. 在电子商务领域, 为解决由于商品个数和种类快速增长导致的信息过载问题, 个性化推荐技术应运而生[2]. 其中协同过滤的推荐技术, 是利用用户以及项目的数据, 有效的帮助用户发现自己感兴趣的项目, 是个性化推荐中研究和应用最为成功的技术之一[3], 并被广泛地应用于电子商务的各个领域.

目前国内将协同过滤应用于高考志愿推荐的研究较少, 由于高考机制不同, 国外研究成果很难适应于我国的高考志愿领域. 王灵峰[4]基于协同过滤算法设计高考信息推荐引擎时, 利用用户对网页的浏览次数

① 基金项目:教育部人文社科基金(10YJCZH073);江苏省自然科学基金(BK2012385);博士点基金(20123218120034);南京航空航天大学基本科研业务费(NS2013083)

收稿时间:2014-11-17;收到修改稿时间:2015-01-03

和浏览时间作为用户的信息关注度, 并转化为用户对分类信息的评分, 构建用户评分矩阵, 这在一定程度上解决了用户冷启动问题, 但无法保证准确性. 王亚婧[5]为了提高应用基于用户的协同过滤算法在高考志愿推荐过程中的精确度, 提出采用信息增益率作为属性选择标准, 并对信息增益率较高的属性给予较高的权值, 但是在推荐志愿数为 5 时只有 50%准确率, 系统最终向考生显示 10 个推荐志愿才能保证较高的准确度. 由此发现, 在应用协同过滤方法时, 推荐准确度与志愿推荐个数两方面存在矛盾, 而且在用户相似度计算时的属性选取也是一个关键因素.

基于此, 本文采用文献调研的方法选取影响高考志愿填报的因素并建立用户属性矩阵, 以此作为计算用户相似度, 最后通过两个阶段产生推荐集. 并通过实验, 对志愿推荐个数和推荐准确度两方面进行调整分析, 使得在志愿推荐个数有限时也保证了较高的推荐准确度, 说明了推荐系统的有效性.

1　基于协同过滤的高考志愿推荐算法

在电子商务环境下, 协同过滤技术由于其良好的算法思想和优秀的推荐结果得到了广泛应用. 协同过滤技术在实际应用中主要分为两类[6]: 基于用户的协同过滤推荐和基于项目的协同过滤推荐, 其基本原理是将口碑效应的过程自动化, 系统提供的建议是基于其他口味相似的用户之喜好来决定的[7]. 本文采用基于用户的协同过滤算法, 并针对高考志愿填报的特点, 进行适应性修改使之适用于高考志愿填报这一新领域.

基于用户的协同过滤算法实现过程分为 3 步[8]: (1)建立用户-项目评分矩阵; (2)查找最近邻居; (3)产生推荐. 然而在高考志愿填报系统中不存在评分数据, 而且考生的属性是影响志愿填报的因素而非评分数据, 为此本文利用用户的属性作为相似性计算的基础, 算法实现过程如图 1 所示.

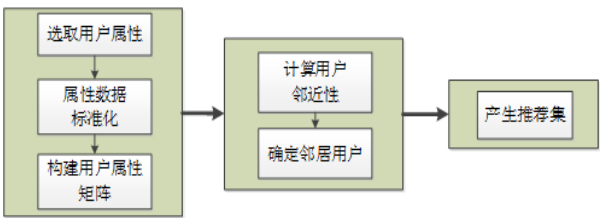

图 1　基于协同过滤的高考志愿推荐算法过程图

首先在相关文献的基础上, 综合分析影响考生填报志愿的因素, 从高考数据库的考生数据表以及高校数据表中提取相关属性构建用户属性矩阵; 然后根据欧几里得距离计算用户之间的距离, 根据距离最近原则确定邻居用户, 最后将邻居用户的录取院校作为推荐集推荐给考生.

1.1　用户属性的选取

在高考志愿推荐系统中, 由于不存在用户对项目的评分数据, 系统将利用用户属性数据来计算用户之间的相似性. 高考数据一般包括报名库、志愿库、成绩库、录取库等信息表, 每个表都包括许多不同的属性, 属性的选取将直接影响推荐结果的准确性. 李令青[9]等人在《高考专业填报决策的影响因素探析》一文通过问卷调查总结影响考生报考决策的因素主要有 4 个, 分别是发展前景、兴趣特长、他人意见、录取机率, 并认为高校招生录取主要根据考生的高考成绩, 考生实事求是评估自己的实力与特点, 合理定位, 是填报志愿的重要依据. 殷员分[10]在《高考考生志愿数据分析与挖掘研究》中利用决策树的方法对历史高考数据进行分析, 发现考生能否被录取与标准投档成绩、批次名称、志愿序号名称、标准投档成绩和标准批次分数线有着较强的相关性. 王毅杰[11]等人在《高考志愿填报中的行为策略: 户籍的影响》一文中通过分析不同户籍类别的某校、某特定专业学生的高考成绩差异, 基于弱势者的理性行动原则, 认为越是社会较低层的家庭子女在高考志愿的填报中越倾向于保守.

综合以上相关文献的结论, 本文认为影响考生填报志愿主要包括高考成绩以及考生社会属性两方面因素, 结合高考数据库, 最终选取性别、户口类型、经济条件、名次四个属性作为影响考试填报志愿的属性. 其中性别对于高考志愿的影响主要是考虑到兴趣爱好的不同, 即偏向于选择文史或理工类院校; 户口类型是指农村或城市户口, 这对高考志愿填报的影响主要在于学校的选择上, 社会较低层的家庭子女在高考志愿的填报中越倾向于保守; 经济条件是根据考生家庭所在地的 GDP 划分为富裕、一般、贫困三种, 这对高考志愿填报的影响体现在对高校所在地的考虑, 家庭经济条件不好的考生在其他条件一样的情况下会倾向于选择在非一线城市的高校; 而名次则是影响高考志愿填报的最关键因素, 相对于高考成绩而言, 各高校

每年录取的学生平均名次较稳定, 且与试卷难易程度无关. 这样四个因素就包括了考生填报志愿时对高校类型、高校所在地、高校档次等多方面的考虑.

一般高考数据库中包括报名库、志愿库、成绩库、录取库、高校计划库等信息表, 其中性别、户口类型数据可以直接从考生报名表中得到, 而经济条件则按照考生报名信息中的家庭所在地(区、县级)当年的GDP水平高低将考生的经济水平划分为富裕、一般、贫困, 名次则按照高考录取政策, 根据成绩表中的各项成绩对考生进行排名即可得到.

1.2 查找邻居用户

1.2.1 建立用户属性矩阵

在利用距离度量不同用户之间的邻近性时, 为了避免不同的属性尺度对距离度量的影响, 必须先对属性数据进行标准化处理, 即无量纲化处理. 本文采用极差变换方法对数据进行标准化, 如下所示:

$$x_{ij}'=\frac{x_{ij}-\min\limits_{j}(x_{ij})}{\max\limits_{j}(x_{ij})-\min\limits_{j}(x_{ij})}$$

$$i=1,2,\cdots,\mathrm{n}\ ;\ j=1,2,\cdots,\mathrm{p} \qquad (1)$$

其中 n 表示用户数, p 表示属性个数, x_{ij} 表示原始数据, $\max\limits_{j}(x_{ij})$, $\min\limits_{j}(x_{ij})$ 分别表示第 j 个属性的最大值和最小值.

1.2.2 计算用户邻近性

查找邻居用户的基础是用户之间相似度或相异度, 在基于协同过滤的高考志愿推荐算法中, 邻居用户的查找是最关键的一步, 这直接决定了推荐项目的产生. 本文利用欧几里得距离法计算用户之间的距离, 并以此来表示用户之间的邻近性, 公式如下:

$$d_{ij}=\sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2}\ ,\ i,j=1,2,\cdots,\mathrm{n} \qquad (2)$$

其中 x_{ik}, x_{jk} 分别表示第 i 个和第 j 个用户的第 k 个属性值.

在欧式距离公式中, 所有属性在计算邻近度时被看作是同等重要, 然而在高考志愿推荐系统中, 考生成绩类属性显然比社会属性重要, 因此, 可以根据属性的贡献程度对每个属性加权修改邻近度公式, 公式如下:

$$d_{ij}=\sqrt{\sum_{k=1}^{p}((x_{ik}-x_{jk})w_k)^2}\ ,\ \sum_{k=1}^{p}w_k=1 \qquad (3)$$

其中 w_k 表示第 k 个属性的权重, 该权重一般由领域专家提供或是根据统计数据产生.

1.3 产生推荐集

在高考志愿推荐系统中, 高分段的考生分布较稀疏, 对应的录取院校少; 而低分段的考生分布较密集, 且对应的录取院校较多. 为了提高推荐结果的准确性, 对于不同分数段的考生需要应用不同的策略. 为了使考生能够高效的选取适合自己的高校, 推荐算法可以根据考生实际可填报的高校数来确定推荐集的高校数目R, 一般可以将R定位实际填报高校数的2倍, 这样既可以有效缩小范围, 又可以给考生提供一定的选择空间. R 个推荐高校的产生是基于邻居用户的录取高校而产生. 首先对于邻居用户的确定, 由于高分段的考生较稀疏, 对其而言具有“参考价值”的邻居用户较少, 相反, 低分段的考生较为密集, 具有“参考价值”的邻居用户较多, 所以应根据考生所处的分数段来确定其邻居用户. 另外低分段的考生较密集, 对应的录取院校较多, 本文采取以“投票”形式从邻居用户所有的录取院校中选取 R 个高校作为推荐集, 并认为距离为0的用户具有最高“参考价值”, 具有“一票决定权”.

因此, 在高考志愿推荐系统中, 推荐集的产生分为两个阶段: (1)将距离为 0 的用户作为目标用户的最近邻, 并将其对应的录取院校加入到推荐集中; (2)根据目标用户的所处分数段, 选取距离最小的 N 个用户作为邻居用户, 并将 N 个用户所录取的院校按照人数降序排列, 并依次将院校加入到推荐集中, 直至推荐集中的院校个数达到预先设定的推荐个数 R, 形成最终的推荐集. 推荐集的个数 R 可以根据高考志愿填报系统中志愿个数来相应确定.

2 实验分析

2.1 数据来源与实验环境

本文利用某省 2011 年高考数据中的文科第一批的考生数据作为实验数据, 原始数据包括报名库、志愿库、成绩库、录取库、高校计划库等信息表, 每个表中包括多个属性, 根据2.1中的分析, 首先从考生报名库中抽取考生的性别、户口类型以及家庭地址属性, 并从成绩库中抽取考生总分进行排序, 作为学生的名次, 然后从录取库中提取录取高校, 形成完整的数据集, 包括用户属性数据以及对应的项目数据, 包括9423个考生, 148个高校. 随机抽取其中的900条数据作为测试集, 即看作录取高校未知的目标用户, 其余的8523条数据作为训练集, 即录取院校已知的非目标

用户.

实验过程中的算法是在 VS 环境中采用 C#语言实现.

2.2 建立用户属性矩阵

在训练集和测试集的基础上, 利用公式(1)对用户属性数据进行标准化, 表 1 为用户属性的原始数据, 表 2 为标准化后的无量纲属性数据.

表 1　用户属性矩阵(标准化前)

院校代号	性别	户口类型	经济条件	名次
3129	2	1	1	3545
3129	1	1	0	5389
3129	2	1	1	4810
3129	2	2	0	8058
3129	1	1	2	5675
3129	2	1	2	3586
3129	2	1	1	2172
3129	1	1	2	3691
3129	1	1	1	7126
3118	2	1	0	4449
5201	1	1	0	587
5201	1	1	0	726
5201	2	1	0	769

表 2　用户属性矩阵(标准化后)

院校代号	性别	户口类型	经济条件	名次
3129	1	0	0.5	0.376180872
3129	0	0	0	0.571913809
3129	1	0	0.5	0.510455365
3129	1	1	0	0.855217068
3129	0	0	1	0.602271521
3129	1	0	1	0.380532852
3129	1	0	0.5	0.230442628
3129	1	0	1	0.391678165
3129	0	0	0.5	0.756289141
3118	1	0	0	0.472136715
5201	0	0	0	0.062201464
5201	0	0	0	0.076955737
5201	1	0	0	0.081520008

2.3 计算用户邻近性

在标准化的用户属性数据基础上, 利用公式(3)就可以计算目标用户与非目标用户之间的距离. 考生各个属性的权重一般由专家给出, 或根据统计数据产生, 本文将考生的四个属性: 性别、户口类型、经济条件以及名次权重分别设为 0.1, 0.1, 0.1, 07, 据此, 就可以计算目标用户与非目标用户之间的距离, 从而确定邻居用户.

2.4 产生推荐集

对于测试集中的目标用户, 首先根据公式(3)计算其与非目标用户的距离, 根据 1.3 中的两阶段法产生推荐集时, 需要事先确定邻居用户数 N 和推荐集的院校个数 R. 本文首先对 2011 年一本文科生的成绩进行统计, 算出各分数段平均每个分数的考生数. 从文科一本分数线开始, 每 10 分为一档, 并将分数高于 393 的归位一档, 统计结果如表 3 所示. 以最低分数段(343-352)为例进行说明, 平均每个分数有近 400 考生, 因此可以在某种程度上认为对于该分数段的考生, 其具有参考意义的邻居用户有 400 人, 故对该分数段将 N 设置为 400, 其它分数段以此类推, 分别确定 N 取值. 由于江苏省高考志愿填报时, 对于一本文科生可以填报 3 个一本院校以及 3 个二本院校, 因此本文将推荐集中的院校个数 R 设定为 6.

表 3　各分数段平均每分值对应考生数

分数段	平均人数	N
393-414	3.6	10
383-392	16.2	20
373-382	61.8	70
363-372	150.2	200
353-362	319.4	400
343-352	389.7	400

在确定了 N 和 R 的取值后, 就可以依据 1.3 的两个阶段产生推荐集. 首先选取距离为 0 的最近邻用户的录取院校加入到推荐集中. 判断推荐集中院校个数是否小于 R, 若是, 则根据第二步, 依次加入不同的院校, 直至推荐个数等于 R, 形成最终的推荐集. 图 2 即为推荐集部分截图, 其中第一列为用户 ID 和其实际录取的院校, 后面以制表符隔开的即为推荐集中的院校.

```
8523:1106	1108	1271	1106	5303	9001	1101	1201
8524:5404	1119	1113	1271	2105	1108	1116	1107
8525:1119	1107	1116	1271	7101	1251	1201	1119
8526:9001	1271	1108	1116	1119	1113	1107	1251
8527:5202	2601	1108	1119	2112	1271	5206	6101
8528:1105	1101	1108	2105	3132	1119	1271	5206
8529:1271	1110	4301	1119	1271	1116	1251	1107
8530:1202	9001	1115	1108	1106	1361	1201	1202
8531:1107	1108	1119	1271	6203	1115	2661	2112
8532:2114	5206	4133	1101	6203	1108	1102	1103
8533:3112	1108	3119	2107	1103	1102	5201	3120
8534:1116	1271	1251	1361	1116	1107	1119	1108
8535:1361	1321	1201	1361	6202	1106	1251	1115
8536:3102	3102	3119	1101	3116	2201
8537:1101	3103	1101	3116	3102	2201	2109	2101
8538:7102	1291	5601	1361	1202	1322	1115	1201
8539:1108	1108	3142	4102	3307	1322	9001	1361
8540:5203	1271	1108	1107	1119	1116	1251	1201
8541:5202	2117	1108	2106	6103	1271	1119	1116
8542:1271	1116	1321	1119	1251	1271	1108	1361
```

图 2　推荐集部分截图

2.5 实验结果分析

根据算法输出的推荐集, 按照不同分数段进行统计分析, 统计结果如图 3 所示, 其中总人数表示测试集中 900 条数据处于该分数段的人数, 正确人数指考生实际录取院校在推荐集中, 正确率指考生实际录取院校在推荐集中的人数占其所处分数段总人数的比例.

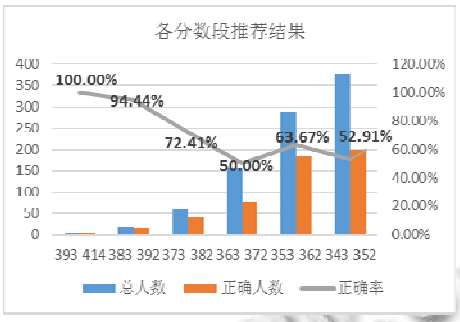

图 3 各分数段推荐结果数据分析图

从图中可以得出以下结论:

①高分段推荐效果好. 从图中可以看出, 在高分段正确率最高达到了 100%, 这一方面说明了算法的可行性, 另一方面也说明了高分段的考生对于院校的选择较为明确且具有主动权, 不存在“信息过载”的问题, 这部分考生更关注的是专业而非高校, 而这在本文的算法中并未考虑到.

②高分段与低分段的正确率差异大. 不同于高分段, 低分段的推荐效果则较差, 最低只有 50%, 这说明算法在设计过程中需要重点考虑考生分布较密集的中低分段考生, 比如结合院校的招生人数或者其他更多的属性数据进行算法优化.

③算法的可行性及改进方案. 各分数段的平均正确率为 72.23%, 这说明利用协同过滤思想进行高考志愿推荐是可行的, 但从推荐结果中也发现了很多不足之处, 比如对于高分段和低分段的推荐效果差异性说明应采取不一样的推荐方法, 对于高分段考生而言可填报的院校较为明确而且占据主动权, 他们更多的考虑专业等其他因素; 而对于低分段考生, 由于考生分布较为密集, 除了考虑到院校档次外, 还应结合院校对应的招生人数或其他因素进行分析, 以此提高推荐效果.

3 结论

本文将协同过滤的方法应用于高考志愿推荐的领域, 将需要填报志愿的高考考生看作是推荐系统中的用户, 高校看作是系统中的项目, 通过分析历年的考生填报数据信息, 为考生推荐其感兴趣的高校, 并对推荐结果进行分析, 说明了推荐系统的有效性, 但也从中发现了一些问题, 比如对于高分段和低分段的推荐效果差异性说明应采取不一样的推荐方法, 对此可以选取不一样的用户属性数据或者是结合其他院校属性(比如院校招生人数)数据来构建用户属性矩阵, 或者利用不同的距离公式来定义不同分数段考生间的距离等措施来提高推荐效果, 这也将是后期的研究内容.

参考文献

1 何小明,张自力,肖灿,夏大飞.基于 OLAP 与数据挖掘的高考招生数据分析.计算机科学,2012,6(39).
2 项亮.推荐系统实践.北京:人民邮电出版社,2012.
3 周丽娟,徐明升,张研研,张璋.基于协同过滤的课程推荐模型.计算机应用研究,2010,4(27).
4 王灵峰.高考信息推荐引擎的设计与实现[学位论文].广州:暨南大学,2011.
5 王亚婧.基于数据挖掘和协同过滤的成人高考志愿推荐系统研究[学位论文].北京:北京林业大学,2011.
6 黄裕洋,金远平.一种综合用户和项目因素的协同过滤推荐算法.东南大学学报(自然科学版),2010,5(40):917.
7 Basilico J, Hofmann T. Unifying collaborative and content-based filtering. Proc. of the Twenty-First International Conference on Machine Learning. Banff, Alta. 2004. 65–72.
8 陈志敏,李志强.基于用户特征和项目属性的协同过滤推荐算法.计算机应用,2011,7(31).
9 李令青,刘彦楼,建伟.高考专业填报决策的影响因素探析.中国健康心理学杂志,2008,8(16).
10 殷员分.高考考生志愿数据分析与挖掘研究[学位论文].重庆:西南大学,2010.
11 王毅杰,梁子浪,陆宏生.高考志愿填报中的行为策略:户籍的影响.天津师范大学学报(社会科学版),2008,3.