数据挖掘期末项目开题报告

基于改进的协同过滤算法的海外研究生院校推荐系统

小组成员：朱正尧 秦晋媛 李怡婷

指导老师：信息管理与工程学院 贺江宁

1 研究问题

基于国外留学学生的录取信息数据集，通过数据挖掘的方法，为目标尚不明确的学生推荐学校列表，并为他们分析被特定学校录取的可能性，帮助学生精准定位，有效解决信息不对称。

给定学生的成绩信息，输出大学推荐列表，其中分为三类：保底学校，主要申请学校和冲刺学校，划分的依据为录取难度，即录取率。因此，问题也被分成了两大块：如何推荐；如何学习录取率。

2 选题背景和意义

从上海财经大学发布的 2018 届本科毕业生就业质量报告中可以看出，我校的升学率逐步攀升的同时，出国留学比率也由 14 年度 19.5% 提高到了 18 年的 24.6%。同时，我们的邻居复旦大学的出国深造率达到了 32.4%。足以见得，随着社会经济水平的高速发展，家庭对教育高度关注，硕士学位已经成了很多岗位的敲门砖，海外深造也成为了很多人的选择。

2018 年我国出国留学人数首次突破 60 万大关，达到 60.84 万人，同比增长 11.74%，持续保持世界最大留学生生源国的地位。其中自费出国群体达到 90% 以上。留学市场持续带动着留学中介行业的发展，也不断涌入更多形式的留学服务机构。截止 2018 年，我国正常持有留学中介服务资质的公司有大约 600 家，实际从事留学中介服务提供的公司及机构远超这个数字。而对于渴望申请海外院校的同学来说，希望咨询的第一个问题就是，我该申请哪些学校? 我被录取的可能性有多大?

留学机构便应运而生，坐拥强大的师资经验和丰富的数据和案例资源，为学生们节省了择校所花费的时间。然而对于留学机构而言，一方面收取着较高的中介服务费，一方面也存在着信息的局限和不透明性。仅仅基于留学机构较为主观的推荐，那么自然有人会开始思考“出国留学是否需要留学机构”?

但是，对于学生们而言，广泛搜索会消耗大量时间，来自互联网的信息的准确性也有待商榷。于是我们希望设计一个基于数据挖掘的研究生择校系统，为学生们推荐海外学校，并估计他们的录取率，为莘莘学子的出国深造之路打下坚实基础。

3 数据描述和说明

表 1: 数据描述和说明

数据来源	申请学生数	被申请学校数	申请记录数
Graduate admission dataset from Gradcafe	19725	507	38318

3.1 数据集特征

从数据量的角度，不管是 user 还是 item 都有大量的数据基础。

表 2: 重要属性说明

重要属性预览	属性说明
university	申请大学的名称
major	申请的专业
degree	申请项目的学位
Decision	最终决定
Gpa	本科期间成绩
GRE($gre_{verbal}/gre_{quant}/gre_{writing}$)	GRE 语言成绩
Comments	申请人对于申请记录的评论

从特征属性的角度，涵盖了包括学校和学生的信息以及学生的一些文本评论，而非单向数据。

从数据质量的角度，该数据集详细显示了学校名称，专业和学位等重要指标，也包含了学生的基本信息和评论，美中不足的是，缺少一些必要的评判指标如托福成绩等。从学校列表这一块，我们发现学校的质量较高，存在探索价值。

4 拟采用的方法简介

基于目前的文献阅读和头脑风暴，我们将方法锁定在了普通分类算法和协同过滤两种算法上。根据问题本身的特征，我们作出了下列分析和改进。

4.1 传统 KNN，决策树等分类方法

将 user 的所有相关信息用一个向量表示，通过 KNN 或者决策树等算法进行分类；
给定一个新用户，套用学习好的分类模型学习标签；
将该标签下的其他用户所选择的学校推荐给该用户

4.1.1 可行性分析

- 优点: 在分类时直接考虑了成绩信息

- 可能的问题: 训练时抛弃了被拒绝的数据

4.2 协同过滤

学生信息作为 user，学校信息作为 item；
构建交互矩阵，利用录取信息去填补矩阵元素；
进行相似度计算，根据相似度生成 Top-N recommendation 的列表；
学生的成绩信息用来学习录取率，作为最后分类冲刺/保底的依据

4.2.1 可行性分析

- 优点: 交互性强，解释了用户决策的主观性

- 可能的问题: 矩阵过于稀疏而导致相似度计算困难加大；存在冷启动问题

- 可能的改进方向: Matrix Factorization

4.3 改进的基于用户的协同过滤

在传统的协同过滤方法中，用户之间的相似度是通过用户和物品的交互矩阵来判定的，但是交互矩阵中包含的用户信息较少，矩阵有过于稀疏的问题。

我们将交互矩阵定义为学生是否申请了这个学校，如果申请了，则 $R(u,j)=1$, 否则 $R(u,j)=0$，含义为学生对学校是否有兴趣。

	Harvard	Stanford	MIT
Alex	1	1	0
George	0	0	1

对于每个学生，有一个特征矩阵，矩阵中记录学生的信息。

	GPA	gre_{verbal}	gre_{quant}	$gre_{writing}$	Degree
Alex	4.0	170	170	6	Ph.D
George	1.0	120	120	1	Master

根据上述矩阵，我们通过下式进行计算：

$$R(u,j)=E(R(u))+\Sigma_{v\epsilon Nu}S(u,v)(R(v,j)-E(R(v))$$

我们根据特征矩阵计算出相似度 $S(u,v)=Cos(u,v)$，其中 u，v 是经过标准化和数值化的信息值，暂时认为各个信息之间是等权重的。

得到 $R(u,j)$ 之后，我们按照 $R(u,j)$ 从大到小，挑选 N 个推荐给学生。

之后基于原有变量进行逻辑回归，其中需要将学校设置为哑变量，估计出被每个学校录取的可能性，被将逻辑回归的结果，作为录取的概率。将概率分档，分为保底学校，主要申请学校和冲刺学校，为学生提供择校规划。

5 研究现状

数据挖掘技术已经有效应用于推荐系统中。目前, 主要的推荐算法有协同过滤推荐, 基于内容的推荐技术, 基于用户统计信息的推荐, 基于效用的推荐, 基于知识的推荐, 基于关联规则的推荐, 组合推荐等等。对于基于 KNN 和协同过滤的推荐系统目前已经广泛应用在新闻、电商、音乐等平台中。在这些推荐模型和算法中, 协同过滤是使用最广泛的。P.Melville, R.J Mooney 和 R.Nagarajan(2002) [1] 提出了内容增强的协同过滤混合算法，提高了推荐系统的效率。针对协同过滤中的数据稀疏问题，张微，刘鲁，葛健 [2] 在 2005 年的小型计算机系统的研究中提出了一种基于粗集的协同过滤算法。首先通过自动填补评分降低数据稀疏性，然后采用分类近似质量计算用户之间的相似性，形成最近邻居，产生推荐预测，提高了推荐的质量。对于推荐系统在教育领域的应用，S.G. Martinez, A.H.Lahdj [3] 对教育类的推荐系统和一般的推荐系统的区别，以及推荐系统在教育应用中的限制进行了开创性的研究，另外，推荐系统已经被有效应用在学生选课机制中，基于学生成绩和之前学生选课为学生提供选课建议。A.G.Schulz, M.Hahsler, M.Jahn [4] 研究了推荐系统在真实大学中的教育和科研中的应用潜力。

对于研究生择校推荐系统而言，浏览新东方、啄木鸟、指南针这样的网站，都需要付费进行人工服务进行择校；对于已有的基于数据挖掘的研究生择校推荐系统方面的尝试，Mahamudul Hasan, Shibbir Ahmed [4] 等通过运用 KNN 算法，基于其他类似学生为用户推荐了 TOP-K 的学校，并且为学生获得更多奖学金提出了推荐。也有学者提出基于学生成绩等个人信息，运用 SVM、KNN 为用户推荐学校并进行交叉验证，得到了较好的效果。Rankishore, Joe [5] 等基于 Random Forest 和 KNN 提供的推荐的准确率达到了 50.6%。但是目前对于择校推荐系统，大多数才用的是 KNN 等

分类算法，由于交互信息相对较少，基于协同过滤等其他方法的推荐较少，我们将通过改进协同过滤算法，进行这方面的尝试，并进行录取率的预测。

区别于基于内容的推荐算法，协同过滤算法往往会面临冷启动的问题。对于我们希望实现的推荐系统而言，向新用户推荐现有的学校是不可避免的难关。目前，已经有一些针对冷启动问题的研究表明引入 Content-based 的信息能够较好的避免冷启动的发生。Yahoo! Labs 的研究者们在 Pairwise Preference Regression for Cold-start Recommendation [6] 中提出了一些解决方案。他们通过纳入 item/user 的内容信息去建立 profile，建立了一个回归模型去匹配 user/item 之间的相关性，构成 pairwise regression。Furong Peng, Xuan Lu, Chao Ma, Yuhua Qian, Jianfeng Lu, Jingyu Yang [7] 在 2018 年提出 Multi-level preference regression（MPR）。其重要贡献是将三部分因素纳入考虑：用户之间属性的相关性，用户对特定物品属性的偏好以及具有特定属性的物品在用户群中的受欢迎程度。他们将回归系数按照上述描述进行分解并引入正则项，其算法的效果在众多推荐算法中脱颖而出，也为我们改进推荐算法打开了新思路。

参考文献

[1] Prem Melville, Raymond J. Mooney, and Rajagopal Nagarajan. Content-boosted collaborative filtering for improved recommendations. In *AAAI/IAAI*, 2002.

[2] 张巍, 刘鲁, and 葛健. 一种基于粗集的协同过滤算法. 小型微型计算机系统, 26(11):1971–1974, 2005.

[3] Salvador Garcia-Martinez and Abdelwahab Hamou-Lhadj. Educational recommender systems: A pedagogical-focused perspective. In *Multimedia Services in Intelligent Environments*, pages 113–124. Springer, 2013.

[4] Mahamudul Hasan, Shibbir Ahmed, Deen Md Abdullah, and Md Shamimur Rahman. Graduate school recommender system: Assisting admission seekers to apply for graduate studies in appropriate graduate schools. *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 502–507, 2016.

[5] Ramkishore, Joe Manley, Swetha Krishnakumar, and Aditya Suresh. University recommender system for graduate studies in usa. 2015.

[6] Seung-Taek Park and Wei Chu. Pairwise preference regression for cold-start recommendation. In *RecSys*, 2009.

[7] Furong Peng, Xuan Lu, Chao Ma, Yuhua Qian, Jianfeng Lu, and Jingyu Yang. Multi-level preference regression for cold-start recommendations. 9:1117–1130, 2018.

6 附录


12:00
推荐系统
个人基本信息
请输入您的本科GPA
单行输入
请输入你的GRE成绩
单行输入
verbal
单行输入
quant
单行输入
writing
单行输入
专业偏好
请输入你的第一志愿
单行输入
请输入你的第二志愿
单行输入
确认提交
12:00
冲刺院校
Item
录取率
Item
录取率
主要申请院校
Item
录取率
Item
录取率
Item
录取率
保底院校
Item
录取率
Item
录取率