文章编号:2096-1472(2023)10-0036-04
DOI:10.19644/j.cnki.issn2096-1472.2023.010.008

基于协同过滤的知识推荐模型研究

张 晴[1],喻 健[1],张 涛[1],李幼凤[2]

(1.湖北工程学院计算机与信息科学学院,湖北 孝感 432000;
2.湖北工程学院人工智能产业技术研究院,湖北 孝感 432000)
✉ zhangqing_lax@163.com; yuj@hbeu.edu.cn; htsg@yeah.net; feng.li@hbeu.edu.cn

摘 要:随着在线教育的普及,各类学习平台产生了海量的知识资源和用户数据。基于这些辅助数据和协同过滤算法,提出了一种混合的知识推荐模型,用于解决传统推荐系统中的冷启动和稀疏性问题,从而提高在线学习的效率。该模型首先利用用户的注册信息和对知识的隐式反馈构建初步的个人概貌,其次根据用户对知识的显式评分和其他反馈完善个人概貌,最后根据概貌和知识间的相似度进行推荐。实验表明,该模型使用不同邻居数量测试的*MAE*(Mean Absolute Error,平均绝对误差)均值约为0.775 5,低于修正余弦算法的0.790 1,且没有明显的噪点,同时*MAE*的标准差约为0.072 4,低于皮尔逊算法的0.083 7,相较于传统的修正余弦算法和皮尔逊算法,其在知识推荐上能兼顾良好的准确性和稳定性。

关键词:协同过滤;知识推荐;个性化学习;推荐算法
中图分类号:TP315 **文献标志码**:A

Research on Knowledge Recommendation Model Based on Collaborative Filtering

ZHANG Qing[1], YU Jian[1], ZHANG Tao[1], LI Youfeng[2]

(1.*School of Computer and Information Science, Hubei Engineering University, Xiaogan* 432000, *China*;
2.*Institute for AI Industrial Technology Research, Hubei Engineering University, Xiaogan* 432000, *China*)
✉ zhangqing_lax@163.com; yuj@hbeu.edu.cn; htsg@yeah.net; feng.li@hbeu.edu.cn

Abstract: With the popularization of online education, various learning platforms have generated a massive amount of knowledge resources and user data. Based on these auxiliary data and collaborative filtering algorithms, this paper proposes a hybrid knowledge recommendation model to solve the cold start and sparsity problems in traditional recommendation systems, thereby improving the efficiency of online learning. Firstly, in this model, the user's registration information and implicit feedback on knowledge are used to construct a preliminary personal profile. Secondly, the user's explicit rating of knowledge and other feedback are used to improve the personal profile. Finally, recommendations are made based on the similarity between the profile and knowledge. The experiment shows that the average *MAE* (Mean Absolute Error) of the model tested with different number of neighbors is about 0.775 5, which is lower than the 0.790 1 of the modified cosine algorithm and there is no significant noise. At the same time, the standard deviation of *MAE* is about 0.072 4, which is lower than the 0.083 7 of the Pearson algorithm. Compared with traditional modified cosine algorithm and Pearson algorithm, the proposed model can well balance accuracy and stability in knowledge recommendation.

Key words: collaborative filtering; knowledge recommendation; personalized learning; recommendation algorithm

0 引言(Introduction)

随着网络和信息技术的发展,在线学习平台的普及让用户可以低成本地获取知识和提高学习效率,与此同时,学习平台积累了海量的知识资源和用户数据,如何高效提取、分析和利用这些数据资源是个性化学习领域的一个重要研究方向。目前,传统的在线学习模式存在一些问题,例如部分用户对网络

收稿日期:2023-02-28
基金项目:湖北省教育科学规划重点课题(2019GA043);湖北工程学院校级教学改革研究项目(202139);湖北工程学院科学研究计划项目(200301061616)

www.rjgcz.com

和新技术不够了解,导致学习目的性不强;一些学习平台存在特色不鲜明、更新速度慢、资源分散的问题,用户很难在平台中找到自己所需要的内容。实现信息筛选的方法目前主要有搜索引擎[1]和推荐系统[2],搜索引擎是通过关键词定位相关的信息,推荐系统是根据用户的特点主动把合适的内容推送给用户。翁可立[3]探索了基于智能技术的"虚实双空间一体化学习环境"、"一主两翼"个性化学习平台、"二主三环"个性化学习模式和"324"个性化学习评价体系,以期促进学生个性得到全面和谐发展。杜志建等[4]围绕中国现状、个性化测评与学习的基本原理和关键技术等展开分析,研究了基于知识图谱的动态评测技术在个性化学习中的应用。汤彤等[5]以英语学习为例,将基于 MOOC(Massive Open Online Courses,大型开放式网络课程)的个性化学习方式与传统学习方式的效果进行了对比研究。上述研究工作关注了不同技术在个性化学习领域的应用,但对技术实现细节的探究稍有不足。为此,本文针对协同过滤推荐技术进行了研究,改进并设计实现了一种知识推荐模型,模型可以根据用户背景信息和学习情况自动选择学习内容,尝试将个性化推荐技术应用于教学领域,以期提高用户的学习效率。

1　协同过滤技术(Collaborative filtering technology)

协同过滤推荐技术[6]的算法流程一般包括收集用户信息、计算用户和项目的相似度、根据相似度寻找相似邻居、生成推荐结果等步骤。相比于传统推荐方法的基于文本过滤,协同过滤有以下优点。

(1)能够过滤机器难以自动分析或文本难以形容的信息,比如图片、声音、视频等多媒体信息。

(2)能够利用他人的经验,避免机器分析的片面性。

(3)能够利用相似使用者的反馈信息,对推荐的内容进行更新。

虽然协同过滤是一种有效且广泛使用的推荐算法,但是它存在以下典型问题[7]。①冷启动问题。冷启动问题也称第一评价问题,算法在运行初期会因缺乏数据而无法工作:新用户没有评价过项目,系统对新用户的兴趣缺乏了解而无法为新用户精准推荐;从未被评价过的新项目也不会被推荐。②稀疏性问题。在一些商业系统中,相似度矩阵的规模很大,用户评分只是零星分布在巨大的相似矩阵中,而基于稀疏的评分矩阵难以准确计算相似度,导致推荐质量大大降低。③扩展性问题。协同过滤算法中相似度矩阵的阶数和计算量都会随着用户和项目的增加而急剧上升,导致严重的扩展性问题。

2　算法改进(Algorithm improvement)

为了解决冷启动问题,让新用户和新知识项目快速融入系统,需要对系统中的新项目主动尝试性推荐,就像新商品上市时需要促销员请用户免费试用。在新用户注册时,可以适当地推送一些新项目,这种方式既能减少算法中新项目的数量,又可以根据用户对新项目的反馈结果获取其新的偏好。推送项目中也要穿插一些热门项目,以便提高用户的使用兴趣。

稀疏矩阵会影响相似度计算的效果,所以要在计算相似度前尽量降低评分矩阵的稀疏度。常用的降低矩阵稀疏度的方法有填充评分的中间值、平均分等,但是这些方法在降低矩阵稀疏度的同时,也改变了矩阵中向量的实际相似度。本文研究使用父类填充法降低评分矩阵的稀疏度。不同的知识项目有不同的所属学科,所属学科可能又属于更高级的父类,按照小类到大类的划分,总能给当前项目找到一个合适的类别。当一个项目没有被用户评分时,可以根据该项目所属分类其他项目的评分给它一个预估分数;如果所属分类没有其他已评分的项目,则追溯到项目的上级分类,根据上级类别给出预估评分;重复多次直到顶级分类,就能找到已评分项目。本文研究假设所有项目都属于一个顶级分类,尽可能地利用已有项目评分生成预估分数,降低矩阵的稀疏度。

解决扩展性问题就是要保证系统的响应速度,可以让大部分的相似度和预估评分计算离线进行,系统推荐时就可以直接读取相应得数据。此外,可以设置不同的计算频率,调节推荐系统的实时性,离线处理能极大地提高系统的响应速度。

基于用户的协同过滤和基于项目的协同过滤分别利用用户相似度和项目相似度计算用户对项目的预测评分,并据此获得推荐结果。基于用户的协同过滤和基于项目的协同过滤在同样的数据集上计算出的推荐结果有较高的重合率,而且它们理论上有相同的精度。单一算法是把用户评分矩阵从行方向或者列方向划分成若干向量用于相似度计算,如果把基于用户的协同过滤和基于项目的协同过滤结合起来从两个方向同时划分向量,分别计算出推荐并取其交集,可以在一定程度上提高推荐的精度[8]。

3　模型实现(Model implementation)

3.1　收集信息

收集信息用于收集能反映用户兴趣偏好的各类信息并量化为数值,存入用户对项目的评分矩阵,本系统中的项目主要是指知识或课程。除了注册阶段获得的用户背景信息,信息采集还要收集用户对课程的评分、对课程的操作记录等行为。在用户注册后首次登录系统时,根据用户的背景信息,从相关学科选取热门和有代表性的课程、最近点击量最多和最少的课程同时推荐给用户浏览,并要求用户对课程做出反馈评分。推荐点击量多的课程能引起用户的兴趣,而推荐点击量少的课程可以减少未评分的课程,解决冷启动的问题。在用户登录访问课程页面时,后台会记录用户访问的课程类别、访问时间、所属学科、访问频率、收藏情况等信息,并量化为用户的兴趣数据;用户也可以在学习页面对当前课程进行评分,直接表达自己的兴趣。

3.2　预测评分

预测评分是推荐系统的核心部分,根据信息收集阶段得到的评分矩阵计算用户或课程之间的相似度。用户对课程的评分矩阵如表1所示。

表1　用户对课程的评分矩阵

Tab.1 User's scoring matrix for courses

单位:分

用户名	课程 C1	课程 C2	课程 C3	课程 C4	课程 C5	课程 C6
U1	5	0	0	1	2	1
U2	2	1	3	4	0	0
U3	0	1	5	2	5	2
U4	1	2	0	4	0	0
U5	5	0	1	0	5	1

www.rjgcz.com

该矩阵是5个用户(U1～U5)对6门课程(C1～C6)的评分。这里设用户对课程的评分为1～5分,分数越高,代表评价越好,0分为没有评分的课程。根据该评分矩阵可以计算出用户或课程之间的相似度,本文以课程相似度为例,说明计算过程。

计算相似度有多种方法,本文选用Pearson相关系数[9]和余弦相似度[10]度量课程相似度。Pearson相关系数可以用来计算两个项目之间的相关性,公式如下:

$$sim(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\times(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\times\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

计算出课程之间的皮尔逊相关系数,如表2所示。

表2 课程之间的皮尔逊相似度

Tab.2 Pearson similarity between courses

课程名	课程C1	课程C2	课程C3	课程C4	课程C5	课程C6
课程C1	1	−1	−1	−0.97	0	0
课程C2	−1	1	0	0.5	0	0
课程C3	−1	0	1	−1	0	1
课程C4	−0.97	0.5	−1	1	1	1
课程C5	0	0	0	1	1	0.5
课程C6	0	0	1	1	0.5	1

使用余弦相似度计算项目之间的相似度,在协同过滤系统中使用余弦相似度公式计算时,为了减小用户评分尺度对项目相似度的影响,通常先对每项评分值减去每行评分的均值,例如对表1的评分矩阵进行处理后得到的结果如表3所示。

表3 处理后的评分矩阵

Tab.3 Processed scoring matrix

用户名	课程C1	课程C2	课程C3	课程C4	课程C5	课程C6
U1	3.5	−1.5	−1.5	−0.5	0.5	−0.5
U2	0.33	−0.67	1.33	2.33	−1.67	−1.67
U3	−2.5	−1.5	2.5	−0.5	2.5	−0.5
U4	−0.17	0.83	−1.17	2.83	−1.17	−1.17
U5	3	−2	−1	−2	3	−1

根据表3的矩阵计算课程的相似度,余弦相似度的计算公式如下:

$$sim(x,y)=\cos(\boldsymbol{x},\boldsymbol{y})=\frac{\boldsymbol{x}\times\boldsymbol{y}}{\|\boldsymbol{x}\|\times\|\boldsymbol{y}\|}=\frac{\sum_{k=1}^{n}R_{x,k}\times R_{y,k}}{\sqrt{\sum_{k=1}^{n}R_{x,k}^2}\times\sqrt{\sum_{k=1}^{n}R_{y,k}^2}} \quad (2)$$

同理,利用公式(2)求出课程之间的余弦相似度,如表4所示。

表4 课程之间的余弦相似矩阵

Tab.4 Cosine similarity between courses

课程名	课程C1	课程C2	课程C3	课程C4	课程C5	课程C6
课程C1	1	−0.48	−0.74	−0.28	0.18	−0.31
课程C2	−0.48	1	−0.12	0.48	−0.75	0.49
课程C3	−0.74	−0.12	1	0.09	0.10	−0.04
课程C4	−0.28	0.48	0.09	1	−0.78	−0.47
课程C5	0.18	−0.75	0.10	−0.78	1	−0.03
课程C6	−0.31	0.49	−0.04	−0.47	−0.03	1

根据相似度可以找出课程的邻居,邻居是与当前课程相似度较大的一组课程集合。邻居查找算法包括固定数量的邻居和基于相似度门槛的邻居两种,系统中为了保证每个课程都能找到邻居集合,选用了固定数量的邻居算法。得到相似邻居集合后,既可以把集合中的课程推荐给当前课程的用户,也可以根据集合中课程的评分预测当前课程的评分,从而减少矩阵中未评分课程的数量。

3.3 产生推荐

系统从预测评分矩阵中找到该用户对应的行,然后把该行有评分的课程根据TOP-N原则选取N门评分最高的课程推荐给用户[11]。预测评分矩阵和用户评分矩阵类似,在用户评分矩阵中用户未评分的项目分值为0,但是在预测评分矩阵中用户未评分的项目有一个预测分值,而已经评分过的项目分值标记为0。例如,在表5的预测评分矩阵中,S1～S10是系统对用户未评分课程的预测评分,而分值为0的项目是用户已经评分过的课程。

表5 预测评分矩阵

Tab.5 Prediction scoring matrix

单位:分

用户名	课程C1	课程C2	课程C3	课程C4	课程C5	课程C6
U1	0	S1	S2	0	0	0
U2	0	0	0	0	S3	S4
U3	S5	0	0	0	0	0
U4	0	0	S6	0	S7	S8
U5	0	S9	0	S10	0	0

在用户接收到系统推荐的课程并学习后,用户可以对课程做出实际评分,系统在收到用户评分后,会把预测评分矩阵的对应项设置为0,同时把该课程评分存入用户评分矩阵。在系统下次推荐课程时,会根据最新的用户评分矩阵进行计算。

www.rjgcz.com

4　模型测试(Model test)

4.1　评价指标

评价推荐算法的一个常用方法是预测误差分析,最常用的指标是 *MAE*[12]。平均绝对误差是误差绝对值的平均值,计算公式如下:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

平均绝对误差能屏蔽预测误差的正负值,能更客观地测量预测的精度。在公式(3)中,n 表示用户已经实际评分的项目数,y_i 表示系统的预测评分,$\hat{y}_i$ 表示用户的实际评分,*MAE* 值越低,表示系统的预测精度越高。

4.2　测试结果

本文进行的测试基于 Item CF 算法分别用皮尔逊相关系数、修正余弦相似度、综合相似度三种方法计算了预测评分,并对比了不同邻居个数预测评分的 *MAE* 值。三种相似度算法产生的 *MAE* 值对比如表 6 和图 1 所示。

表6　*MAE* 实验结果

Tab.6 MAE experimental result

算法	邻居数量						
	1	2	3	4	5	6	7
皮尔逊算法	0.914 2	0.810 8	0.772 4	0.718 6	0.724 0	0.682 5	0.679 2
修正余弦算法	0.914 2	0.856 0	0.747 2	0.758 9	0.751 4	0.751 4	0.751 4
综合算法	0.914 2	0.823 1	0.760 3	0.754 3	0.748 4	0.735 8	0.692 2

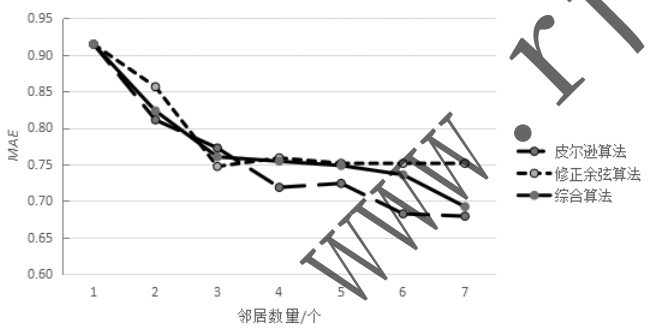

图 1　MAE 数据对比

Fig. 1 MAE data comparison

从测试结果可以看出,三种算法预测评分的 *MAE* 值都随着邻居数量的增加而有降低趋势。这是因为基于少数人评分计算出的预测精度会差一些,而选取邻居数量的个数越多,参考的信息越广泛,则预测精度也会越高。皮尔逊算法、修正余弦算法、综合算法的 *MAE* 值都基本符合预测精度的需求,三者的 *MAE* 均值分别约为 0.757 4、0.790 1、0.775 5,*MAE* 标准差分别约为 0.083 7、0.067 1、0.072 4,其中综合算法的稳定性好、噪点少,综合算法相比其他算法在 *MAE* 相近的情况下,提高了预测评分的稳定性。

5　结论(Conclusion)

本文基于对协同过滤推荐算法的研究改进,设计并实现了一种个性化的知识推荐模型,该模型首先利用用户的注册信息和对知识的隐式反馈构建初步的用户概貌,其次根据用户对知识的显式评分和其他反馈不断完善优化用户概貌,最后根据概貌和知识间的相似度为用户选择合适的知识资源,让用户的学习更有针对性。实验结果表明,相较于单纯的相似度推荐算法,该模型能有效预测用户对知识的需求和学习兴趣,预测效果良好且稳定。在后续的研究中,将继续改进推荐模型,把用户的反馈信息更高效地融合到算法的修正中,以期构建更精准的推荐模型。

参考文献(References)

[1] 陆俊. 搜索引擎优化技术在网站运营中存在的问题与对策研究[J]. 中国新通信,2021,23(16):62-63.

[2] 王国霞,刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用,2012,48(7):66-76.

[3] 翁可立. 基于智能技术的个性化学习研究[J]. 中国现代教育装备,2022(24):19-23.

[4] 杜志建,申华,管文英,等. 基于知识图谱的个性化学习系统[J]. 人工智能,2022(2):96-104.

[5] 汤彤,徐鲁强. 基于 MOOC 的个性化学习研究[J]. 软件导刊,2019,18(12):249-251,255.

[6] BEHERA G,NAIN N. Collaborative filtering with temporal features for movie recommendation system[J]. Procedia Computer Science,2023,218:1366-1373.

[7] 孙传明,周炎,涂燕. 基于混合协同过滤的个性化推荐方法研究[J]. 华中师范大学学报(自然科学版),2020,54(6):956-962.

[8] 翁小兰,王志坚. 协同过滤推荐算法研究进展[J]. 计算机工程与应用,2018,54(1):25-31.

[9] 夏景明,刘聪慧. 一种基于用户和商品属性挖掘的协同过滤算法[J]. 现代电子技术,2020,43(23):120-123.

[10] 严宇桥,张蔚坪. 信息茧房与准确率:基于复合型算法的个性化模拟推荐系统[J]. 电子技术与软件工程,2019(24):257-258.

[11] YU J,XIONG Z G,BAO Q,et al. Design of an algorithm for recommending elective courses based on collaborative filtering[J]. Journal of Computational Methods in Sciences and Engineering,2022,22(6):2173-2184.

[12] 孟晗,高岑,王嵩,等. 结合信任关系的用户聚类协同过滤推荐算法[J]. 计算机系统应用,2020,29(8):224-229.

作者简介:

张　晴(1981-),女,硕士,实验师。研究领域:计算机应用,实验室技术与管理。

喻　健(1989-),男,硕士,实验师。研究领域:智能信息处理与数据挖掘。

张　涛(1980-),男,硕士,高级实验师。研究领域:计算机应用技术,实验技术。

李幼凤(1978-),女,博士,讲师。研究领域:先进控制及其性能评估,复杂系统建模。

www.rigzz.com