基于行为分析的学习资源个性化推荐

聂黎生
(江苏师范大学 计算机科学与技术学院,江苏 徐州 221116)

摘 要:随着数字化学习资源规模急剧扩张,“知识过载”和“学习迷航”等问题限制了在线学习资源推荐的性能,学习者从海量的学习资源中选择合适资源的难度随之增大。针对传统推荐算法中存在的数据稀疏和学习资源个性化推荐精度不高等问题,提出了基于行为分析的学习资源个性化推荐算法。首先,构建学习者-学习资源评分矩阵;其次,挖掘学习者行为数据并将行为数据格式化融入到协同过滤个性化推荐过程;最后,计算学习者相似度并为待推荐学习者生成学习资源推荐列表。为验证模型的有效性,以“Live Course 在线课程平台”数据为样本构建实验数据集,通过对比实验表明,该方法具有更高的推荐精度,能够更加精确和全面定位学习者的真实需求,实现学习资源个性化推荐。

关键词:行为分析;学习资源;个性化推荐;协同过滤;推荐精度
中图分类号:TP301;G434 **文献标识码**:A **文章编号**:1673-629X(2020)07-0034-04
doi:10.3969/j.issn.1673-629X.2020.07.008

Personalized Recommendation of Learning Resources Based on Behavior Analysis

NIE Li-sheng
(School of Computer Science and Technology, Jiangsu Normal University, Xuzhou 221116, China)

Abstract: With the rapid expansion of the scale of digital learning resources, “knowledge overload” and “learning maze” and other issues limit the performance of online learning resources recommendation, and it is more difficult for learners to select appropriate resources from a large number of learning resources. Aiming at the problems of sparse data and low accuracy of personalized recommendation of learning resources in traditional recommendation algorithms, a personalized recommendation algorithm of learning resources based on behavior analysis is proposed. First of all, the rating matrix of learner learning resources is constructed. Secondly, the behavior data of learners is mined and the behavior data format is integrated into the collaborative filtering personalized recommendation process. Finally, the similarity of learners is calculated and the learning resources recommendation list is generated for the learners to be recommended. In order to verify the validity of the model, an experimental data set is constructed based on the “Live Course online course platform” data. The comparative experiment shows that the proposed method has higher recommendation accuracy, can more accurately and comprehensively locate the real needs of learners and achieve personalized recommendation of learning resources.

Key words: behavior analysis; learning resources; personalized recommendation; collaborative filtering; recommendation accuracy

0 引 言

随着网络技术的飞速发展,良好的交互技术和丰富的在线资源使学习变得更加便捷、自由、开放,彻底改变了传统的学习方式,实现了教育领域的颠覆性创新。不同的学习者知识结构、知识能力、学习能力和兴趣偏好千差万别。通过挖掘学习者的学习偏好,在线学习系统可以准确推荐符合学习者学习需求的个性化学习资源,从而为其提供及时的资源推荐服务[1-2]。为了提高学习资源个性化推荐精度,众多学者进行了深入研究。文献[3]分析在线学习的行为特征,挖掘学习者的性格特征与学习效率的关系,实现个性化学习方法推荐。文献[4-5]认为用户之间的相似关系对于发现利益重叠的群体至关重要,可以产生多重相似关系和利益集群的形成。基于此开发了一种层次兴趣重叠检测方法,并提出了个性化推荐模式。文献[6-7]通过利用知识图谱构建知识点体系,提出了知识表示-协同过滤相结合的方式推荐有效资源,解决在线学习导航问题。文献[8]采用聚类算法将具有相同兴趣的用户聚集到同一个集群中为用户推荐可能喜欢的项目,从而提高推荐效率和精度。文献[9]基于本体

收稿日期:2019-08-21 **修回日期**:2019-12-23
基金项目:国家自然科学基金(21776119);教育部产学合作协同育人项目(201902172045);江苏省社科基金项目(15TQB002)
作者简介:聂黎生(1978-),男,硕士,讲师,研究方向为数据挖掘、移动学习等。

和顺序模式挖掘的混合知识对电子资源进行有效推荐。文献[10]则将地理位置近邻的用户具有更为相似的访问服务作为预测依据。文中基于学习者的学习行为和兴趣偏好,采用改进的协同过滤个性化推荐算法,从学习者自主学习的角度实现学习资源个性化推荐,有效缓解传统协同过滤推荐算法存在的冷启动和矩阵稀疏性等问题。

1　个性化资源推荐系统模型

数字化时代在线学习产生的行为数据凸显重要,通过挖掘其背后隐含的重要信息,能够得到更加丰富的内容甚至超出人们的期望。文中基于“学习者-资源”二元网络,依据学习者在线学习生成的学习行为,以协同过滤技术算法为核心,构建学习资源个性化推荐系统模型,如图 1 所示。该模型的关键是通过个性化主动推荐服务,实现推送符合学习者本身知识水平和学习偏好的学习资源,达到与原有知识主动、快速的衔接,提高学习者的学习效率。

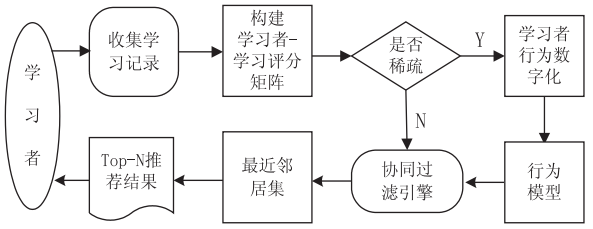

图 1　个性化资源推荐系统模型

1.1　学习者行为采集

学习行为是个性化推荐系统的依据。学习者在线学习过程中会产生大量的学习行为直接或间接地反映了学习者的学习偏好。通过收集和记录学习者的学习行为,进一步挖掘学习过程中产生的浏览、收藏、分享、评论等学习行为数据进行量化分析处理,并建立学习者行为模型,清楚地了解学习者的学习偏好。

1.2　学习资源库

学习资源是个性化推荐系统的基础。学习资源库支持文本、音频和视频等多种媒体类型,为学习者提供全面、完善且有助于提高认知水平的学习资源。为了方便对学习资源内容进行分类,实现资源的统一管理和高度共享,学习资源库将所有资源都加入了知识点属性标签。

1.3　协同过滤引擎

协同过滤是个性化推荐系统的核心。文中通过挖掘和分析学习者的历史学习行为,准确预测学习者潜在的学习偏好,进而向其推送适合的学习资源,实现个性化推荐服务,优化学习者的学习体验。传统的协同过滤推荐算法存在冷启动和矩阵稀疏性等问题,其过分依赖学习者对资源的评分导致推荐结果精度受到影响。文中将学习行为融入到协同过滤算法并对其做出改进,在矩阵初始化时,如果学习者对某学习资源评价较少,则挖掘学习者对资源的其他行为并且将学习者行为模型数字化为学习权重加入到相似性计算中,有效地缓解矩阵的稀疏性问题,使推荐精度大幅提高。

2　学习资源个性化推荐过程

学习者模型构建过程其实质就是学习者-学习资源评分矩阵的形成过程,在推荐过程中若计算出的矩阵过于稀疏,该算法通过挖掘学习者隐式学习行为并融入到推荐系统,避免矩阵稀疏对推荐结果造成的不利影响。通过充分利用与其相似学习者信息进行学习者聚类分析,基于相似学习者的学习偏好预测目标学习者的学习需求,实现学习资源个性化推荐,提高学习效率。

2.1　学习者偏好矩阵构建

系统采用知识结构对学习资源建立知识体系。首先将学习者对学习资源的评价转化为 $n * m$ 阶矩阵:

$$\boldsymbol{\beta} = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1m} \\ R_{21} & R_{22} & \cdots & R_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ R_{n1} & R_{n2} & \cdots & R_{nm} \end{bmatrix} \tag{1}$$

该矩阵由 n 个学习者参与对 m 个学习资源的评分构成,式中 $R_{ij}(i \in [1,n], j \in [1,m])$ 代表了学习者 i 对学习资源 j 的评分。

2.2　计算学习者-学习资源矩阵稀疏性

一方面由于学习者之间选择的差异性,导致学习者的评分差别非常大;另一方面学习资源和学习者数量的增长,必然存在有些学习资源没有经过学习者的评价,同时由于系统无法获取新进入学习者的学习偏好,从而导致新增的学习者和学习资源无法获得推荐。为了缓解上述数据稀疏性和冷启动带来的问题,可以为矩阵稀疏性设置一个临界阈值 x,并通过式(2)初步判别矩阵是否稀疏:

$$\text{Sparsity} = \frac{\text{Num}_{\text{Eval}}}{\text{Num}_{\text{Learner}} * \text{Num}_{\text{Res}}} \tag{2}$$

其中,Num_{Eval} 为学习者对学习资源的评价数量,$\text{Num}_{\text{Learner}}$、$\text{Num}_{\text{Res}}$ 分别为学习者和学习资源数量。当 $\text{Sparsity} < x$ 时,说明学习资源评价矩阵过于稀疏,评分向量会造成活跃学习者与其他学习者之间较高的相似度。为了避免这一问题,通过挖掘学习者学习行为,不同的行为赋予不同的得分规则,然后将学习行为格式化为权重 φ 融入到评价矩阵中。本研究主要基于学习者浏览(B)、收藏(F)、分享(S)、评论(C)这四种学习行为赋予不同的分数,得到学习者的实时行为分值:

$$S(\text{Learner}) = 1 * B + 2 * F + 3 * S + 5 * C \quad (3)$$

其中,对不同行为赋予的分数为 1,2,3,5,但这个值应该不断调整。当学习者数量少的时候,各项事件都小,此时需要提高每个事件的行为分值来提升学习者行为的影响力[11];当学习者规模变大时,行为分值也应该逐渐降低。考虑到学习者数量的动态变化,采用自适应调整行为权重得分 φ:

$$\varphi = \frac{\sum_{i=1}^{n} S(\text{Learner})_i}{n} \quad (4)$$

其中,$S(\text{Learner})_i$ 表示第 i 个学习者行为得分,n 表示学习者总数。这样就保证了在学习者规模的动态变化情况下仍能产生基本稳定的行为得分,然后将格式化学习者权重值 φ,添加到评价矩阵中。

2.3 学习者近邻集生成

在协同过滤算法中,最近邻居表示是最为关键的一步,决定着学习资源个性化推荐的精度。依据学习者之间相似度的计算值,发现相似度较高的目标学习者并且根据其学习行为信息,预测与学习者兴趣偏好相匹配的学习资源并推荐[12]。根据式 1,取出 n 个学习者对 m 个学习资源的评分,计算学习者之间的相似度。由于不同评价算法之间存在差异性,为了降低学习者主观性评分对研究结果的不利影响,通过对余弦相似度算法进行修正,在相似度计算时将每个资源的评分减去该学习者对所有资源的平均评分[13]。该算法将学习者对资源的评分看作是 m 维的向量,假设 i 和 j 分别代表两个不同的学习者,采用修正后余弦相似度算法计算两者间的相似度 $\text{Sim}(i,j)$。具体计算方法为:

$$\text{Sim}(i,j) = \frac{\sum_{k\in S_{i,j}} (E_{i,k} - \bar{E}_i) * (E_{j,k} - \bar{E}_j)}{\sqrt{\sum_{k\in S_{i,j}} (E_{i,k} - \bar{E}_i)^2} * \sqrt{\sum_{k\in S_{i,j}} (E_{j,k} - \bar{E}_j)^2}} \quad (5)$$

其中,$S_{i,j}$ 为学习者 i 和 j 均已评分的学习资源集合,S_i 和 S_j 分别表示学习者 i 和 j 对学习资源评分的集合,$E_{i,k}$ 和 $E_{j,k}$ 分别表示学习者 i 和 j 对学习资源 k 的评分,$\bar{E}_i$ 和 $\bar{E}_j$ 分别表示学习者 i 和 j 对已学学习资源的平均评分。

文中对式(5)相似性计算方法进行了改进,将计算的学习行为权重 φ 融入到相似性计算中。改进后的计算方法为:

$$\text{Sim}(i,j) = \frac{\sum_{k\in S_{i,j}} (\varphi_{i,k} - \bar{E}_i) * (\varphi_{j,k} - \bar{E}_j)}{\sqrt{\sum_{k\in S_{i,j}} (\varphi_{i,k} - \bar{E}_i)^2} * \sqrt{\sum_{k\in S_{i,j}} (\varphi_{j,k} - \bar{E}_j)^2}} \quad (6)$$

相似度计算完成后,按照目标学习者 a 和其他学习者的相似度,选择相似度最为接近的 n 个学习者构成待推荐近邻集 $Z = \{L_d, d \in [1,n]\}$。余弦值越接近 1,表明两个向量越相似;反之越接近 0,表明两个向量越不相似。

2.4 生成推荐结果

根据式(6),基于生成的目标学习者 a 的近邻集,在包含学习者 a 的全部学习者评分集合中除去目标学习者的所有已评分学习资源,可得目标学习者的待预测评分资源 S_a。计算目标学习者 a 对每一学习资源 $t \in S_a$ 的预测评分,降序排序选取评分最高的前 N 项作为 Top-N 推荐给目标学习者。由于不同学习者评价存在差异性,推荐结果采用以下方式:

$$\text{Pre}_{j,k} = \bar{R}_j + \frac{\sum_{a\in z} \text{sim}(j,a)(R_{a,k} - \bar{R}_a)}{\sum_{a\in z} \text{sim}(j,a)} \quad (7)$$

其中,$\text{Pre}_{j,k}$ 为采用改进算法预测的学习者 j 对资源 k 的评分,$\bar{R}_j$ 为学习者 j 已经完成评分的均值,Z 为利用式(6)计算得到的近邻集。推荐过程中,依据目标学习者 j 与数据库中每个学习者的相似性产生最近邻居集,筛选出相似度最高的前 N 项资源作为推荐结果,推送给学习者。

3 评价指标及结果分析

3.1 实验数据

为验证文中个性化推荐方法的有效性,实验数据集来源于“LiveCourse 在线课程平台”,利用 MySQL 数据库存储领域专家对课程学习资源标注了 90 个知识点以及知识点之间的关联关系和相应的学习资源。数据集由 65 名学习者在 4 个月内对 900 个学习资源,包含 138 个视频、287 个幻灯片、475 个文本资源的 21 738条学习行为数据构成。实验主要提取浏览(B)、收藏(F)、分享(S)、评论(C)这四种学习行为数据,按照1∶4分成训练集和测试集两部分。

3.2 评价指标

依据学习者在训练集中的学习行为,通过文中算法与基于矩阵分解的协同过滤算法(probabilistic matrix factorization,PMF)、基于卷积神经网络的推荐算法(convolutional neural networks,CNN)分别向学习者推荐学习资源,评估算法的性能。精确率和召回率通常用来反映推荐算法性能,精确率反映推荐的精度,召回率衡量推荐系统的查全率。但也有可能出现推荐系统具有较高的精确率而召回率却很低的矛盾状况,因此单一的指标不能较为全面地评价推荐算法的好坏[14]。为了平衡二者之间的影响,通过引入了综合评价指标 F-Measure 和 MAE 评价各算法性能。F-

measure 值越高表明实验结果越好,其计算公式如下:

$$F-measure=\frac{2PR}{P+R} \tag{8}$$

其中,P、R 分别表示推荐结果的准确率和召回率。具体计算公式为:$P=\sum_{u_i\in U} hit(u_i)/\sum_{u_i\in U} L(u_i)$、$R=\sum_{u_i\in U} \mathrm{hit}(u_i)/\sum_{u_i\in U} T(u_i)$。$U$ 为数据集 65 名学习者的集合,$\mathrm{hit}(u_i)$ 表示推荐给学习者 u_i 的学习资源真正在测试集中被该学习者学习的数量。$L(u_i)$ 表示提供给学习者 u_i 的学习资源数量,$T(u_i)$ 为测试集中学习者 u_i 真正学习的学习资源数量。在本次实验中 $T(u_i)=36$,$\sum_{u_i\in U} T(u_i)=900$。

平均绝对误差(MAE)用于计算预测评分和实际评分之间的差异,是评判推荐系统结果精准与否的重要指标。推荐算法中,设置预测推荐结果为二元值 1 或 0,分别代表推荐资源和学习者习知识点是否一致。其计算公式如下:

$$MAE=\frac{1}{N}\sum_{i\in N}|P_{u,i}-r_{u,i}| \tag{9}$$

其中,N 表示推荐的学习资源数量,$P_{u,i}$ 表示学习者已学习的资源,此处 $P_{u,i}$ 的值为 1,$r_{u,i}$ 表示推荐结果是否准确的指标值,如果推荐结果和学习者学习的知识点一致,则 $r_{u,i}$ 的值为 1,否则 $r_{u,i}$ 的值为 0。因此,MAE 值越小表示算法推荐精度越高,反之则表示推荐精度越低。

3.3　结果分析

实验分别选取推荐资源数量 12,24,36,48,60 验证不同算法的性能,通过图 2 可以看出文中算法 F-measure 值高于其他两种算法,具有明显的优势,表明推荐结果较好;在推荐资源数量 M 为 36 左右时,可以得到较高的推荐精度,学习资源个性化推荐结果更加符合学生的实际需求。M 值的选取对于推荐系统精度比较重要,但是推荐结果的精度对 M 值也不是非常敏感,二者之间不成线性关系,只要选择合适的范围就可以获得较高的推荐精度。

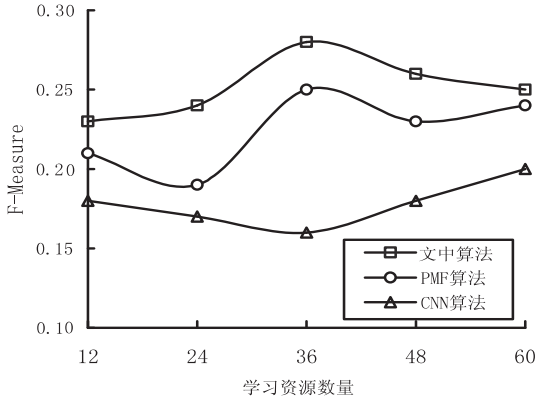

图 2　不同算法 F-Measure 值对比

图 3 显示了近邻集数量分别为 10,20,30,40,50,推荐学习资源数量为 36 的情况下不同算法的 MAE 值,测试结果表明文中算法的 MAE 值在不同近邻集数量下都明显低于其他算法,说明文中算法推荐质量最高,推荐结果符合目标学习者的学习偏好。随着近邻集数量的增加、数据的稀疏性降低,算法收敛的速度加快[15],MAE 值逐渐降低最后趋于稳定。实验结果中 Top-N 的 N 值为选取的学习者相似度较大的 N 个学习者作为近邻集,非最终推荐列表的 Top-N。

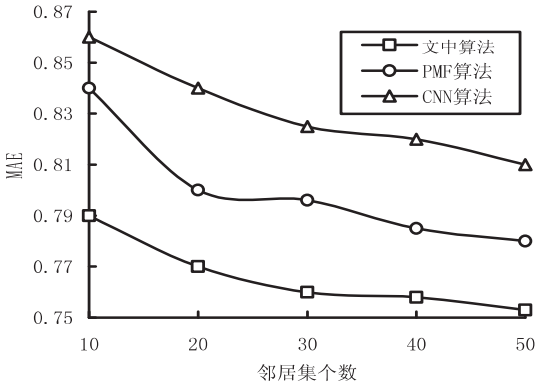

图 3　不同算法 MAE 值对比

4　结束语

针对如何提高学习资源个性化推荐的精度与效率问题,通过构建学习者-学习资源的评分矩阵,综合考虑学习者的学习行为,采用改进的相似度算法实现学习资源的个性化推荐。实验结果表明该方法优化了学习资源个性化推荐过程,推荐结果精度更高,效果更好。未来将挖掘更多能反映学习偏好的行为数据,以改进和完善推荐模型,促进学习系统提供更加精准的个性化服务,并将其推广应用到其他资源推荐领域。

参考文献:

[1] 李浩君,张　征,张鹏威. 基于阶段衍变双向自均衡的个性化学习资源推荐方法[J]. 模式识别与人工智能,2018,31(10):921-932.

[2] 李文欣,文勇军,唐立军. 教育资源个性化推荐方法研究与实现[J]. 计算机技术与发展,2019,29(6):18-22.

[3] 陈晋音,方　航,林　翔,等. 基于在线学习行为分析的个性化学习推荐[J]. 计算机科学,2018,45(11A):422-426.

[4] ZHENG Jianxing,WANG Suge,LI Deyu,et al. Personalized recommendation based on hierarchical interest[J]. Information Sciences,2019,479:55-75.

[5] LEE D H,BRUSILOVSKY P. Improving personalized recommendations using community membership information[J]. Information Processing and Management,2017,53(5):1201-1214.

[6] 王晓东,时俊雅,李　淳,等. 学习资源精准推荐模型及应

(下转第 41 页)