基于内容和协同过滤加权融合的音乐推荐算法

彭余辉,张小雷,孙　刚*

(阜阳师范大学 计算机与信息工程学院,安徽 阜阳 236037)

摘　要:随着互联网的普及以及音乐库的高速更新换代,用户对音乐的需求变得越来越大,传统的推荐算法已经无法满足用户及时准确地寻找到所喜欢的音乐。因此,针对传统音乐推荐算法的不足,通过对协同过滤推荐算法的分析,提出基于内容和协同过滤加权融合的音乐推荐算法。与传统推荐算法及部分相关推荐算法比较,加权融合推荐算法计算出的推荐结果可以更高效快速地将用户感兴趣的音乐推荐出来。

关键词:内容;协同过滤;融合;音乐推荐算法　　**DOI:**10.13757/j.cnki.cn34-1328/n.2021.02.009

中图分类号:TP311　　**文献标志码:**A　　**文章编号:**1007-4260(2021)02-0044-05

Music Recommendation Algorithm Based on Weighted Fusion of Content and Collaborative Filtering

PENG Yuhui, ZHANG Xiaolei, SUN Gang

(School of Computer and Information Engineering, Fuyang Normal University, Fuyang 236037, China)

Abstract: With the popularity of the internet and the high-speed update of music libraries, users' demand for music has become greater and greater, and the traditional recommendation algorithms have been unable to satisfy users in finding their favorite music in time and accurately. Therefore, in view of the shortcomings of traditional music recommendation algorithms, we proposes a music recommendation algorithm based on content and collaborative filtering weighted fusion through the analysis of collaborative filtering recommendation algorithms. Compared with the traditional recommendation algorithm and some related recommendation algorithms, the recommendation results calculated by this weighted fusion recommendation algorithm can more efficiently and quickly recommend the music that users are interested in.

Key words: content; collaborative filtering; fusion; music recommendation algorithm

近年来,互联网信息和电子音乐网站爆炸式增长,音乐资源异常非富,但各种音乐无法有效地整合,造成信息过载,使人们无法快速找到他们所喜欢的音乐。一个好的音乐推荐算法对于音乐网站尤为重要,它既给用户带来方便,也给音乐网站带来了更多的利润和流量。

传统的音乐推荐算法是单一的基于内容的推荐或者是使用协同过滤推荐。根据文献[1],基于内容的推荐本质是对于信息的检索和过滤。协同过滤推荐算法分为基于用户(User-CF-Based)和基于物品(Item-CF-Based)的协同过滤算法。User-CF-based算法按照用户之前对物品所作的行为来分析用户的偏好并进行衡量和评分,计算用户之间的相似度,并根据相似度将物品推荐给有相似偏好的用户。Item-

收稿日期:2020-07-17

基金项目:安徽省教育厅自然科学研究重点项目(KJ2018A0328,KJ2019A0532,KJ2019A0542,KJ2020ZD48),阜阳师范大学大数据与智能计算创新团队(XDHXTD201703)和阜阳市人文社会科学研究专项项目(FYSK2019QD10)

作者简介:彭余辉(1996—),男,安徽阜阳人,阜阳师范大学计算机与信息工程学院硕士研究生,研究方向为推荐算法。
E-mail:15256867605@163.com

通信作者:孙刚(1978—),男,安徽阜阳人,博士,阜阳师范大学计算机与信息工程学院教授,研究方向为大数据、云计算、人工智能。
E-mail:ahfysungang@163.com

CF-Based算法根据用户喜欢的物品，将与该物品相似的物品推荐给用户。

为了改进传统的推荐算法，提高协同过滤推荐算法的准确率和效率，文献[2]提出了RC-DFM模型，经过评论和内容的加权融合，缓解了数据的稀疏性，增加了推荐的准确率，但是对于大型数据集来说，这种模型的效率会随所需时间的增加而变低。文献[3]针对冷启动问题进行研究，根据用户评分和项目属性进行评分预测，并将推荐结果推荐给用户，但未考虑新用户因素。文献[4]通过提取特征词，使用特征标签来代替物品本身，将多种标签结合进行分析与融合以提高精确率。以上推荐算法远远无法满足需求宽泛的用户，在“长尾理论”的支持下，很多不受欢迎商品的销售规模足以比拟受大众欢迎的商品。文献[5]在长尾理论的基础上提出了item-CF-IIF算法，通过对热门商品的惩罚以及对推荐物品排序优化来增加推荐质量、准确率以及用户的体验程度。文献[6]通过探索情绪标签来构建情绪模型，并结合协同过滤产生推荐列表。文献[7]利用分类和情境感知融合的方法对音乐偏好进行特权融合，降低推荐复杂度，提高了推荐的效率和质量。在文献[8]中，将用户–项目类别评分相似度和用户–项目类别兴趣相似度加权融合，有效地缓解了数据的稀疏性。文献[9]根据用户的满意度改进余弦相似度计算方法，进行协同过滤推荐以得到满意的推荐列表。文献[10]将基于内容的推荐算法与基于用户的协同过滤推荐算法进行混合，形成一种新的推荐算法，处理了物品在没有评价的情况下同样能被推荐给用户，并且很明显地增加了推荐结果的准确率，但缺少了一定的平衡。

基于传统推荐算法的闪光点和不足，本文提出基于内容和协同过滤加权融合（MR_CCFI）的推荐算法，来分别得到基于用户–内容和基于物品–内容融合的推荐算法，以提高推荐准确率。

1　基本框架

MR_CCFI推荐算法的基本框架如图1所示。通过基于内容和协同过滤加权融合的算法来进行内容–用户、内容–物品加权融合的推荐，在数据集的影响下，不同推荐算法给用户带来的推荐效果是不同的。

1.1　基于内容的推荐算法

根据音乐内容的文本信息，这里使用词频–反文档频率（TF-IDF）的方式得到用户偏好矩阵[11]，设给定音乐的集合为$L=\{L_1,L_2,L_3,\cdots,L_n\}$，特征（关键）词组为$I=\{i_1,i_2,i_3,\cdots,i_m\}$，$L_j$表示第$j$首音乐，词频公式为：

$$P_{\mathrm{TF}}(i,j)=\frac{f(i,j)}{\sum_{k\in j}f_{k,j}},$$

其中，$f(i,j)$是词i在音乐j中所出现的次数，$\sum_{k\in j}f_{k,j}$是音乐j中所有词出现次数的总和，$k\in j$表示词在音乐j中。反文档频率公式为：

$$P_{\mathrm{IDF}}(i)=\log\frac{N}{n(i)},$$

其中，N指所有音乐的数量，$n(i)$指N中特征词i出现过的音乐数量。

音乐j中特征词i的组合TF-IDF权值计算为：

$$P_{\mathrm{TF\text{-}IDF}}(i,j)=P_{\mathrm{TF}}(i,j)\times P_{\mathrm{IDF}}(i),\tag{1}$$

其中，$P_{\mathrm{TF\text{-}IDF}}(i,j)$表示第$j$首音乐中与特征词$i$对应的词。归一化处理：

$$W_{ji}=\frac{P_{\mathrm{TF\text{-}IDF}}(i,j)}{\sqrt{\sum_{i=1}^{|I|}P_{\mathrm{TF\text{-}IDF}}(i,j)^2}},\tag{2}$$

其中，W_{ji}指第j首音乐第i个词的归一化处理，于是可得到用户音乐偏好矩阵

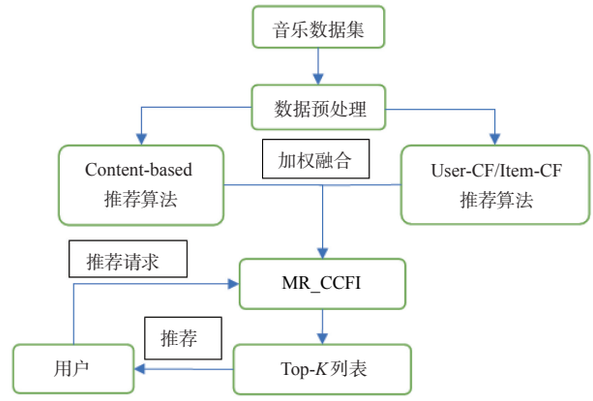

图1　MR_CCFI推荐算法流程

$$\boldsymbol{P}_u = \begin{pmatrix} W_{11} & W_{12} & \cdots & W_{1n} \\ W_{21} & W_{22} & \cdots & W_{2n} \\ \vdots & \vdots & & \vdots \\ W_{n1} & W_{n2} & \cdots & W_{nn} \end{pmatrix}。 \tag{3}$$

1.2 协同过滤推荐算法

1.2.1 基于User-CF的推荐算法

User-CF算法必须要先找到“相似的用户”,再寻找“相似用户所喜欢的物品”。这里首先使用余弦相似度:

$$W_{uv} = \frac{|N(u) \cap N(v)|}{\sqrt{|N(u)||N(v)|}}, \tag{4}$$

此计算方式的时间复杂度是$O(|u| \times |v|)$,u指用户个数[12]。我们对其进行优化,仅计算式(4)中分子不为0的情况,然后再计算用户之间的相似度。

第一步,构造一个歌曲到User的倒排表D,记录用户对哪些歌曲发生过动作。

第二步,依据倒排表D,构造一个用户相似矩阵$\boldsymbol{M}$,$\boldsymbol{M}$指式(4)中的分子部分,在倒排表D中,对于每首歌曲i,设其相应的用户为a、b。

第三步,在$\boldsymbol{M}$中,更新对应位置的元素值,$\boldsymbol{M}[a][b] = \boldsymbol{M}[a][b] + 1$。依次扫描便能够得到一切用户之间不为0的$\boldsymbol{M}[a][b]$值。

根据文献[13],对式(4)改进,有

$$W_{uv} = \frac{\sum_{x \in N(u) \cap N(v)} \frac{1}{\lg(1 + |N(x)|)}}{\sqrt{|N(u)||N(v)|}}, \tag{5}$$

式(5)中分子的倒数部分用来惩罚用户u和v的共同偏好列表中的受欢迎的音乐,减少热门歌曲对用户相似度的影响。用户对歌曲的偏好公式为:

$$\boldsymbol{P}(u,j) = \sum_{v \in S(u,K) \cap N(j)} W_{uv} \boldsymbol{R}_{vj}, \tag{6}$$

其中,$\boldsymbol{P}(u,j)$指用户u对歌曲j的喜欢程度,$S(u,K)$指与用户u兴趣最相近的前K个用户,$N(j)$指用户对歌曲j产生过行为历史的集合,W_{uv}指用户u、v之间的偏好相似度,$\boldsymbol{R}_{vj}$指用户对歌曲j的偏好评分矩阵(如果数据集为隐反馈数据集,那么当用户对歌曲产生了行为,可使$\boldsymbol{R}_{vj} = \boldsymbol{E}$,$\boldsymbol{E}$为单位矩阵)。

1.3 基于Item-CF的推荐算法

Item-CF推荐算法是通过兴趣物品来寻找相似物品,将相似物品推荐给用户。歌曲之间的相似度为:

$$W_{hj} = \frac{|N(h) \cap N(j)|}{|N(h)|}, \tag{7}$$

其中,$|N(h)|$表示多个用户都喜欢歌曲h的数目,分子则表示多个用户都喜欢歌曲h、j的数目。惩罚热门音乐后,相似度计算为:

$$W_{hj} = \frac{|N(h)| \cap |N(j)|}{\sqrt{|N(h)||N(j)|}}, \tag{8}$$

式(8)降低了歌曲j的权重,减小了任何歌曲和热门歌曲很相似的可能。

Item-CF推荐算法首先建立一个用户到歌曲的倒排表E,得到用户与歌曲之间的对应关系;其次通过倒排表E来构建同现矩阵,根据式(8)计算两音乐之间的相似度,得到各音乐之间的相似度矩阵;最后计算用户对于歌曲偏好程度。计算用户对于歌曲的偏好程度的公式为:

$$\boldsymbol{P}(u,j) = \sum_{j \in S(h,K) \cap N(u)} W_{hj} \boldsymbol{R}_{uj}, \tag{9}$$

其中，$N(u)$表示歌曲被用户u喜欢的集合，$S(h,K)$指和歌曲h最相似的前K首歌曲的集合，W_{hj}表示音乐h和音乐j的相似度，$\boldsymbol{R}_{uj}$指用户u对音乐j的偏好评分（如果数据集为隐反馈数据集，那么当用户对歌曲产生了行为，可使$\boldsymbol{R}_{uj}=\boldsymbol{E}$，$\boldsymbol{E}$为单位矩阵）。

1.4　MR_CCFI推荐算法

MR_CCFI推荐算法偏好公式为

$$P=\beta P_u+(1-\beta)P(u,j),\beta\in \mathrm{R},0\leqslant\beta\leqslant 1, \tag{10}$$

其中，β指用户偏好矩阵的权重，$(1-\beta)$指协同过滤算法中用户对歌曲的偏好权重。β值越小，说明用户或物品之间偏好的相似对推荐的影响程度越大，随着β值的增加，音乐内容对推荐的影响程度也在增加。取前K个值，得到Top-K列表，将其推荐给相应的用户u。在用户量不变的情况下：

（1）当音乐数据集较小时，使用Content-User的推荐算法更容易，且更加准确地使目标用户获得相应的Top-K推荐列表。

（2）当音乐数据集较大时，使用Content-Item的推荐算法可以更精确地将Top-K推荐列表推荐给目标用户。

2　实验结果与分析

通过设置数据集的大小来测试不同算法在给用户进行推荐时的准确率。实验数据所采用的数据源是网易云音乐网站2020年上半年的部分数据，选择了一千多个歌单进行相应的数据获取，包含用户、音乐、歌单等信息，使用用户的已听歌曲记录和评分记录等进行实验。为了提高数据挖掘的质量，需要对信息文本进行预处理，再以其中的80%作为训练集（取其中的40%作为小数据集），另外20%作为最后的测试集。

2.1　数据预处理

（1）进行分类标注，即将爬取到的文本打标签，生成标签列表。

（2）分词，去停用词。通过jieba分词库将音乐内容文本进行分词，并对其去停用词，使用词频-逆文档频率（TF-IDF）的方式提取关键词，进而得到用户偏好矩阵$\boldsymbol{P}_u$。

（3）实验中主要选取数据维度信息为：音乐编号、音乐名、音乐发表时间、音乐类型等。

2.2　实验结果分析

将MR_CCFI算法与文献[2]中用户评分和物品属性相融合的推荐算法（RC-DFM）、文献[5]中item-CF-IIF算法以及Content-Based算法进行对比实验，观察几种算法在不同数据集上的推荐效果。

参数设置。在MR_CCFI推荐算法中参数β的设置会影响到推荐结果的准确性，本次实验中，设置K为10、15、20、25来分别检测不同β值对应的平均绝对误差（MAE），参数$\beta\in(0,1)$，如图2所示。实验结果表明，在数据集相同的情况下，MR_CCFI算法中的权重因子β更偏重于协同过滤算法。由图2可知，K固定时，当$\beta=0.7$时，MAE最小，故以下实验参数β的值均设置为0.7。

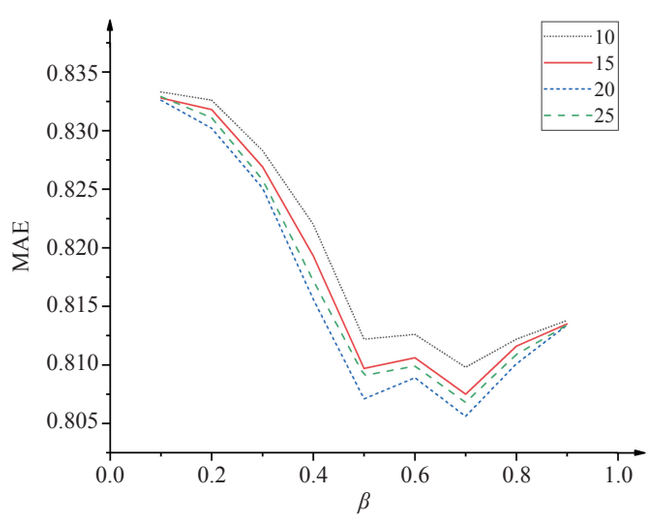

图2　相同K值下，不同β取值对应的MAE值

相似度计算。不一样的相似度计算方法会对推荐结果产生影响,设置K为5、10、15、…、35,比较Cosine、User-content-Cosine和Item-content-Cosine计算方法的MAE,当数据集较大时,如图3所示。由图3可知,K不变时,Item-content算法中余弦计算方法得到的MAE值是最小的,即该算法的推荐效果最好。数据集相对较小时,User-Content算法的结果误差较小,如图4所示。本文使用的余弦相似度计算方法为当前主流计算方法,通过实验表明,该相似度计算方法能够很好地满足用户对推荐结果的需求。

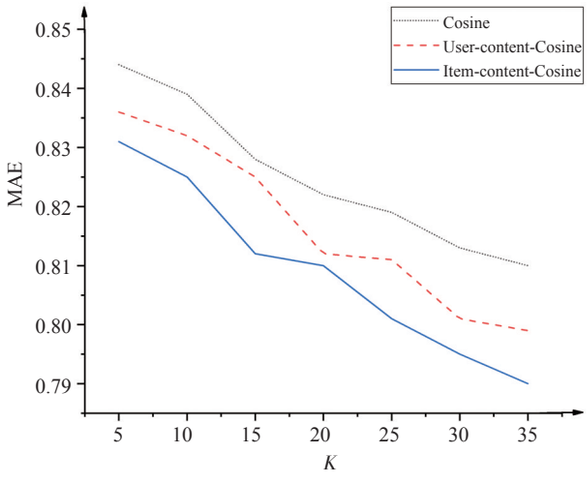

图3 大数据集下,相同K,不同算法的MAE值

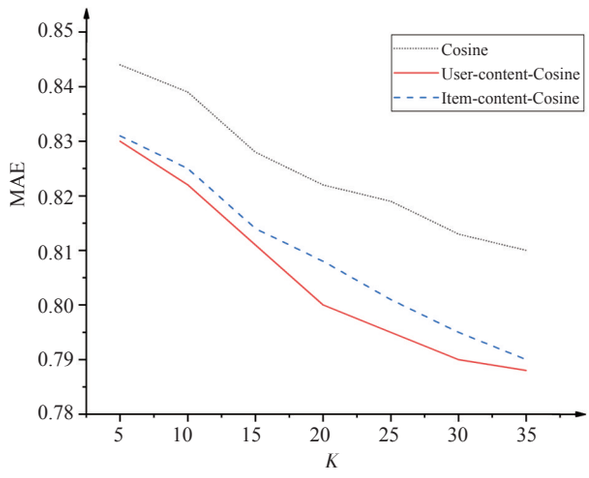

图4 小数据集下,相同K,不同算法的MAE值

MAE。在评分预测中,预测准确率大多使用MAE和均方根误差(RMSE)计算结果,这里运用MAE,将本文算法与文献[2]、[5]进行实验对比,分析本文加权融合算法的有效性。MAE值越小,最终所得到的推荐准确率就越高。当数据集较大时,对比几种推荐算法的MAE值,如图5所示。由图5可知,当K值相同时,基于User-Content的推荐算法和基于Item-Content的推荐算法这两种推荐算法MAE值都最小,且当K为20时,Item-Content算法的MAE是最小的。当数据集较小时,User-Content算法更适合对用户进行推荐,而RC-DFM算法在数据集较小的时候推荐效果会比较差,如图6所示。

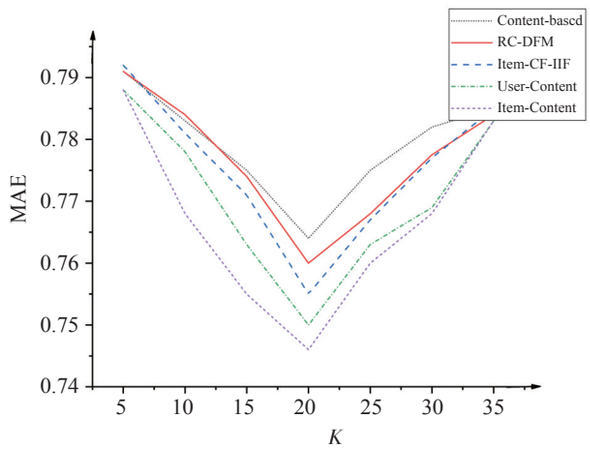

图5 大数据集下不同算法的MAE值

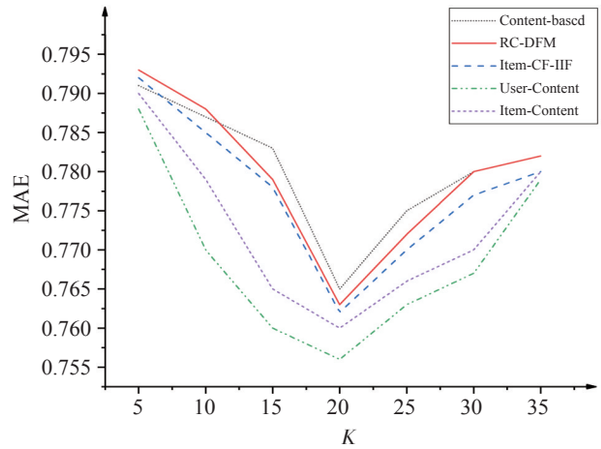

图6 小数据集下不同算法的MAE值

3 结束语

通过对音乐各方面内容和所用主流推荐算法进行分析,提出MR_CCFI推荐算法缓解数据的稀疏性并提高用户对于其偏好音乐获取的效率,经过对数据集的划分来测试相同用户在不同大小数据集下更适合于用户的推荐算法。实验结果表明,数据集较小的情况下,User-Content算法更适合对用户进行推荐;数据集较大的情况下,Item-Content算法更适合对用户进行推荐。在实际应用中,随着大量音乐数据的上传,这种MR_CCFI算法同样适用。

(下转第53页)

在医疗行业及相关领域的数据分类和预测方面具有较好的辅助作用。

参考文献：

[1] 董跃华, 刘力. 基于相关系数的决策树优化算法[J]. 计算机工程与科学, 2015, 37(9): 1783-1793.
[2] 郭华平, 董亚东, 邬长安, 等. 面向类不平衡的逻辑回归方法[J]. 模式识别与人工智能, 2015, 28(8): 686-693.
[3] 周于皓, 张红玲, 李芳菲, 等. 局部关注支持向量机算法[J]. 计算机应用, 2018, 38(4): 945-948.
[4] ZHOU W, WANG H, YANG C, et al. Decision tree based medical image clustering algorithm in computer-aided diagnoses[J]. Journal of Computational Methods in Sciences and Engineering, 2015, 15(4): 645-651.
[5] 张晓惠, 林柏钢. 基于平衡二叉决策树SVM算法的物联网安全研究[J]. 信息网络安全, 2015(8): 26-31.
[6] 史宝鹏, 段迅, 孔广黔, 等. 应用分类模型研究迟发性颅脑损伤的影响因素[J]. 计算机技术与发展, 2018, 28(3): 201-204.
[7] 黄锦静, 陈岱, 李梦天. 基于粗糙集的决策树在医疗诊断中的应用[J]. 计算机技术与发展, 2017, 27(12): 148-152.
[8] 邹丽, 蒋芸, 陈娜, 等. 基于决策树对支持向量机的医学图像分类新方法[J]. 计算机工程与应用, 2016, 52(21): 76-80.
[9] 任仪. 基于决策树的海量医学图像数据挖掘方法研究[J]. 电子设计工程, 2019, 27(6): 33-36.
[10] 李玲, 刘华文, 徐晓丹, 等. 基于信息增益的多标签特征选择算法[J]. 计算机科学, 2015, 42(7): 52-56.
[11] NITHYA N, DURAISWAMY K. Correlated gain ratio based fuzzy weighted association rule mining classifier for diagnosis health care data [J]. Journal of Intelligent & Fuzzy Systems, 2015, 29(4): 1453-1464.
[12] PETERS J. Gini index-based digital image complementing in the study of medical images[J]. Intelligent Decision Technologies, 2015, 9 (2): 209-218.
[13] 董红斌, 滕旭阳, 杨雪. 一种基于关联信息熵度量的特征选择方法[J]. 计算机研究与发展, 2016(8): 1684-1695.
[14] BALDWIN J, LAWRY J, MARTIN T. A mass assignment based ID3 algorithm for decision tree induction[J]. International Journal of Intelligent Systems, 2015, 12(7): 523-552.
[15] SARKAR B, KUMAR A. A hybrid predictive model integrating C4.5 and decision table classifiers for medical data sets[J]. Journal of Information Technology Research, 2018, 11(2): 150-167.
[16] ZHU F. A classification algorithm of CART decision tree based on mapreduce attribute weights[J]. International Journal of Performability Engineering, 2018, 14(1): 17-25.

（上接第48页）

参考文献：

[1] LOPS P, GEMMIS M, SEMERARO G. Content-based recommender systems: state of the art and trends[J]. Recommender Systems Handbook, 2011, 75-105.
[2] 付文静. 基于评论和内容深度融合的跨域推荐问题研究[D]. 济南: 山东大学, 2019.
[3] 王辉, 姜丹, 徐海鸥. 基于用户评分和项目属性的稀疏矩阵预测研究[J]. 电脑知识与技术, 2019, 15(2): 273-275.
[4] 许璐璐. 基于标签的旅游景点个性化推荐研究[D]. 青岛: 山东科技大学, 2018.
[5] 袁煦聪. 基于长尾理论的物品协同过滤推荐算法研究[D]. 淮南: 安徽理工大学, 2019.
[6] 李卓远, 曾丹, 张之江. 基于协同过滤和音乐情绪的音乐推荐系统研究[J]. 工业控制计算机, 2018, 31(7): 127-128, 131.
[7] 吴海金, 陈俊. 融合分类与协同过滤的情境感知音乐推荐算法[J]. 福州大学学报(自然科学版), 2019, 47(4): 467-471.
[8] 于世彩, 谢颖华, 王巧. 协同过滤的相似度融合改进算法[J]. 计算机系统应用, 2017, 26(1): 135-140.
[9] 隋占丽. 基于协同过滤算法的音乐推荐系统[D]. 泉州: 华侨大学, 2013.
[10] 曹毅. 基于内容和协同过滤的混合模式推荐技术研究[D]. 长沙: 中南大学, 2007.
[11] ANKIT K, SAKSHI P, PALAK K, et al. PhishSKaPe: a content based approach to escape phishing attacks[J]. Procedia Computer Science, 2020, 171: 1102-1109.
[12] 陈昕宇, 杨帆. 基于时间衰减的物品相似度协同推荐算法改进[J]. 现代信息科技, 2020, 4(8): 90-92, 95.
[13] BREESE J, HECKERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering[J]. Uncertainty in Artificial Intelligence, 2013, 13(5): 3-5.