融合项目热门惩罚因子改进协同过滤推荐方法

刘雯雯,汪皖燕,程树林
(安庆师范大学 计算机与信息学院,安徽 安庆 246133)

摘 要:推荐系统是大数据时代解决信息过载问题的一种重要工具,协同过滤是推荐系统中出现最早、应用最广泛的一种推荐算法。针对传统协同过滤推荐算法存在的项目热门度偏差问题,提出了一种融合项目热门惩罚因子改进协同过滤推荐方法。引入热门阈值,根据项目热门度将项目进行二分类,即热门项目(项目热门度较高的项目)和非热门项目(项目热门度较低的项目)。重点针对热门项目,融合项目热门惩罚因子改进协同过滤推荐方法,降低热门项目的贡献,从而提升推荐精度。在 MovieLens 100 K 数据集上对所提推荐方法进行实验验证。实验结果表明,在参数取最优值时,所提推荐方法较为有效地降低了评分预测的平均绝对误差和均方根误差,一定程度上验证了项目热门惩罚因子的有效性。

关键词:推荐系统;热门度偏差;协同过滤;二分类;评分预测
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2023)03-0015-05
doi:10.3969/j.issn.1673-629X.2023.03.003

Improved Collaborative Filtering Recommendation Method Integrating Item Popularity Punishment Factor

LIU Wen-wen, WANG Wan-yan, CHENG Shu-lin
(School of Computer and Information, Anqing Normal University, Anqing 246133, China)

Abstract: Recommendation system is an important tool to solve the problem of information overload in the era of big data. Collaborative filtering is the earliest and most widely used recommendation algorithm in recommendation system. Aiming at the problem of item popularity bias in traditional collaborative filtering recommendation algorithm, an improved collaborative filtering recommendation method integrating item popularity punishment factor is proposed. The popularity threshold is introduced, and then items are classified into two categories according to the item popularity (items with high popularity), namely popular items and non-popular items (items with low popularity). Focusing on popular items, the collaborative filtering recommendation method is improved by integrating the item popularity punishment factor to reduce the influence of popular items on neighbors, so as to improve the recommendation accuracy. The proposed method is experimentally verified on MovieLens 100K dataset. The experimental results show that the proposed recommendation method can effectively reduce the mean absolute error and root mean squared error of rating prediction when the parameters are taken the optimal value, which verifies the validity of the item popularity punishment to some extent.

Key words: recommendation system; popularity bias; collaborative filtering; binary classification; rating prediction

0 引 言

随着互联网技术的飞速发展和各种网络平台的日益兴盛,用户面临着日益严重的数据超载、检索困难等问题[1]。用户无法快速、精准地从海量的信息库中获取所需要的信息。为缓解这些问题,推荐系统[2]应运而生,它不需要用户提供明确的需求信息,而是通过主动跟踪用户的历史行为记录来发掘用户的偏好,进而为用户推荐合适的项目[3]。推荐算法是推荐系统的核心,好的推荐算法会为推荐系统带来更好的推荐效果。

但传统推荐算法大多存在着项目热门度偏差问题[4-5]:传统协同过滤推荐算法[6]通常偏向于推荐热门项目[7],忽视了对小众用户偏好的非热门项目的推荐。虽然推荐热门项目通常是一个很好的选择,但推荐列表中可能存在较为普遍的项目,因此,传统协同过滤推荐算法在促进新项目的发现上存在一定的不足,亦忽视了小众用户的偏好,从而导致对非热门项目或新项目的推荐存在缺陷,因为只有极少数用户甚至没有用户对这些项目进行过评分。研究表明,非热门项目的价值远高于热门项目,可以为用户带来更大的收获或利益[8-9]。推荐系统中非热门项目的数量远多于

收稿日期:2022-04-28 **修回日期**:2022-08-30
基金项目:安徽省自然科学基金项目(2008085MF193);安徽省质量工程项目(2019jyxm0285)
作者简介:刘雯雯(1997-),女,硕士研究生,研究方向为个性化推荐;通信作者:程树林,博士,教授,研究方向为智能信息处理。

热门项目(项目热门度较高的项目)的数量,随着推荐系统的不断发展,非热门项目的数量也在持续增长中,如果过多推荐热门项目而忽视对非热门项目的推荐,就有可能在一定程度上降低推荐的精度。

另外,推荐系统中存在的项目热门度偏差问题可能会对社会造成不良影响[4]。比如,在微博、Facebook 等社交网站中,用户的订阅大多来源于推荐系统,其影响着用户的言行举止,若大肆为用户推荐不当的言论或行为,可能会对用户造成不好的影响。例如总是为用户推荐凶杀、猥亵、校园暴力等事件会造成用户的恐慌和不安,甚至造成社会动荡。针对该问题,该文提出一种融合项目热门惩罚因子改进协同过滤推荐方法。引入热门阈值,根据项目热门度将项目进行二分类,即热门项目和非热门项目。对于热门项目,融合项目热门惩罚因子改进协同过滤推荐方法,降低热门项目的贡献,从而提升推荐精度,提高用户满意度。

1 相关工作

传统推荐算法的推荐列表中热门项目更容易被推荐,这过于强调大众偏好,忽视了向小众用户推荐其偏好的非热门项目的重要性。以书籍推荐为例,绝大多数的研究者都购买过《机器学习》这本书,而《推荐系统》是仅研究推荐系统的小众用户才可能购买的一本书。相比较而言,推荐《推荐系统》比推荐《机器学习》更能满足这些小众用户的需求,提高其满意度。为此专家学者从不同层面对热门推荐进行了相关研究,主要的热门推荐分为以下三类。

(1)基于正则化的推荐。Abdollahpouri 等人[10]引入一个灵活的基于正则化的框架,以增强学习排名(Length-to-Rank,L2R)算法中推荐列表的非热门项目覆盖率,但这里的热门项目与非热门项目是按 50/50 划分的,未考虑其他情况,如 20/80。Zhu 等人[11]提出公平性问题中的一个新定义:popularity-opportunity bias。他们认为根据用户的偏好,热门项目比非热门项目更有可能被推荐给用户的这种偏差现象是不合理的,并实证了该偏差的存在,然后通过后处理和内处理两种算法来缓减这种偏差。后处理是按一定规则提高非热门项目的得分,从而使其获得与热门项目相似的排名,内处理是采用正用户项目对的预测偏好分数和相应项目热门度间皮尔逊相关系数的平方的正则化项来实现项目的公平推荐。

(2)基于交互关系的推荐。Sun 等人[12]提出了一种基于主题模型的用户行为记录、项目标签信息和社会关系联合建模方法,分别构建目标用户与一组项目、项目标签和社交用户的对应关系,从而间接构建用户与非热门项目的映射关系,提高非热门项目的推荐。Meghawat 等人[13]提出一种多模式方法,该方法利用视觉特征(即内容信息)、文本特征(即上下文信息)和社交特征(例如平均浏览量和群组数量)来预测社交媒体照片在浏览量方面的项目热门度。

(3)基于检测关键信息的推荐。Liu 等人[14]提出一种新的结合关键词驱动和项目热门度感知的基于无向引文图的论文推荐方法,该方法推荐的论文关联性强且项目热门度高,同时其推荐的结果支持用户对某个主题或领域进行深入、持续的研究,但建立的无向引文图可能存在严重的稀疏问题。Wang 等人[15]提出一种新的带有辅助语义学习的联合深度网络模型,引入文本分析到热门推荐中。首先,使用优化后的字符级卷积神经网络(Character-level Convolutional Neural Network,CharCNN)从用户评论中学习辅助语义向量;然后,使用因子分解机(Factorization Machine,FM)组件和深度组件学习项目属性特征的相应向量表示;同时,采用卷积来模拟隐藏的潜在向量间的相互作用。Piotrkowicz 等人[16]提出一个只使用标题来预测新闻文章项目热门度的新任务,侧面证明了标题对新闻文章项目热门度的影响。

2 融合项目热门惩罚因子改进协同过滤推荐方法

由于原始项目评分数据稀疏度较高,会降低推荐的精度,该文先按照文献[17]的方法对原始项目评分数据中的空缺评分项进行预填充,得到项目评分矩阵 $\boldsymbol{R}$。即根据用户的动态偏好阈值将项目划分成用户偏好项目和用户非偏好项目两部分,对于用户非偏好项目,采用原始项目评分数据中该用户的最低评分进行填充;对于用户偏好项目,融合用户平均评分和项目平均得分以合适的比例进行填充。然后,分析项目热门度,根据热门阈值将项目划分成热门项目和非热门项目。对于热门项目,采用项目热门惩罚因子降低它的贡献;对于非热门项目,则用传统方法进行处理。

2.1 项目热门度

项目热门度:在有限项目集合 $\boldsymbol{S}_I$ 和用户集合 $\boldsymbol{S}_U$ 中,对项目进行过评分的用户数与项目评分数据中总用户数的比例即为该项目的项目热门度,如式(1)所示。

$$H_i = \frac{\sum_{u \in S_U} \lambda_{u,i}}{|\boldsymbol{S}_U|} \tag{1}$$

其中,若用户 u 对项目 i 进行过评分,则 $\lambda_{u,i}=1$;否则 $\lambda_{u,i}=0$。$|\boldsymbol{S}_U|$ 表示项目评分数据中总用户数,H_i 表示项目 i 的项目热门度,其取值范围在[0,1]之间。H_i 的值越大,表明项目越热门。若直接采用对项目进行过

评分的用户数度量项目热门度,则忽视了用户基数总量对项目热门度的影响。假设在数据 A 和 B 中,数据 A 共有 10 位用户,数据 B 共有 50 位用户,其中,均有 5 位用户对项目 i 进行过评分。若直接用对项目进行过评分的用户数来度量项目 i 的项目热门度,则其在数据 A、B 中的项目热门度均为 5,很明显这是不准确的;若采用文中的度量方法,则项目 i 在数据 A、B 中的项目热门度分别为 0.5、0.1,结果更合理,也符合实际情况。

2.2　项目热门惩罚因子

在实际生活中,热门项目出现的频率远远高于非热门项目,更容易被用户接触和评分,但用户多是出于从众或猎奇的心理,因此热门项目较少甚至不能体现用户潜在的偏好,应采用项目热门惩罚因子降低这些项目的贡献。项目热门惩罚因子的具体设置如式(2)所示。

$$W_i = \begin{cases} 1, & H_i \leqslant \mathrm{PT} \\ \mathrm{e}^{-\gamma * H_i}, & H_i > \mathrm{PT} \end{cases} \tag{2}$$

其中,W_i 表示项目 i 的项目热门惩罚因子,PT 表示热门阈值,γ 为平衡参数,其取值范围为 $\gamma>0$,表示项目热门度对项目热门惩罚因子的影响系数。若 $H_i \leqslant$PT,则认为项目 i 为非热门项目,将其对应的项目惩罚因子设置为 1;若 $H_i >$PT,则认为项目 i 为热门项目,其对应的项目热门惩罚因子设置为 $\mathrm{e}^{-\gamma * H_i}$,热门项目的项目热门度越高,其对应的项目热门惩罚因子越小。

由于绝大多数用户都对热门项目进行过评分,导致用户相似性普遍偏高,基于此,该文将设计的项目热门惩罚因子引入用户相似度计算公式中,用来降低热门项目对用户相似度的贡献。改进后的余弦相似度计算公式如式(3)所示。

$$\mathrm{sim}(u,v) = \frac{\sum_{i \in I_{u,v}} R_{u,i} * R_{v,i} * W_i}{\sqrt{\sum_{i \in I_u} (R_{u,i} * W_i)^2} * \sqrt{\sum_{i \in I_v} (R_{v,i} * W_i)^2}} \tag{3}$$

其中,u、v 分别表示用户集合中的两个用户,$\mathrm{sim}(u,v)$ 表示用户 u 与用户 v 的用户相似度。$I_{u,v}$ 表示用户 u 与用户 v 共同进行过评分的项目集合,I_u、I_v 分别表示用户 u、用户 v 进行过评分的项目集合。$R_{u,i}$、$R_{v,i}$ 分别表示用户 u、用户 v 对项目 i 的评分。若用户进行过评分的项目集合中含有热门项目,则其对应的热门惩罚因子小于 1,$\mathrm{sim}(u,v)$ 值随之变小,弱化了用户 u 与用户 v 的用户相似度,降低了热门项目对用户相似度的贡献;若用户进行过评分的项目集合中全是非热门项目,则其对应的热门惩罚因子为 1,$\mathrm{sim}(u,v)$ 值保持不变,即保留用户 u 与用户 v 的原始用户相似度。

2.3　评分预测与评估指标

通过式(3)计算出用户间的用户相似度,选取与目标用户最相似的 K 个近邻,根据这些近邻对目标用户未评分项目的评分预测目标用户的评分,然后将预测评分最高的前 M 个项目推荐给目标用户。

$$P_{u,i} = \frac{\sum_{v \in N_u} \mathrm{sim}(u,v) * R_{v,i}}{\sum_{v \in N_u} \mathrm{sim}(u,v)} \tag{4}$$

其中,$P_{u,i}$ 表示用户 u 对项目 i 的预测评分,N_u 表示用户 u 的近邻集合。该文采用平均绝对误差(Mean Absolute Error, MAE)和均方根误差(Root Mean Squared Error,RMSE)作为所提方法的评估指标,表示用户评分与预测评分之间的偏差。MAE 和 RMSE 的具体计算方式如式(5)、式(6)所示。

$$\mathrm{MAE} = \frac{\sum_{(u,i) \in N} |R_{u,i} - P_{u,i}|}{|N|} \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{(u,i) \in N} (R_{u,i} - P_{u,i})^2}{|N|}} \tag{6}$$

其中,$|N|$ 表示预测评分的数目。

2.4　方法描述

首先,依次根据式(1)、式(2)计算每个项目的项目热门度和对应的项目热门惩罚因子;然后,基于改进的余弦相似度计算用户相似度并选取与目标用户最相似的 K 个近邻,通过这些近邻对目标用户未评分项目的评分来预测目标用户的评分;最后,对预测评分进行降序排序,并获取对应的项目推荐列表。融合项目惩罚因子改进协同过滤推荐方法的算法如下:

算法:融入项目热门惩罚因子的协同过滤推荐方法
输入:项目评分矩阵 $\boldsymbol{R}$,最近邻个数 K
输出:用户对应的项目推荐列表
1:设置平衡参数 γ
2:设置热门阈值 PT
3:　for $i = 1,2,\cdots,n$ do
4:根据式(1)计算项目热门度 H_i
5:根据式(2)计算项目热门惩罚因子 W_i
6:　endfor
7:　for $u = 1,2,\cdots,m$ do
8:　　根据式(3)计算用户之间的相似度
9:　　挑选最相似的 K 个近邻用户
10:　endfor
11:根据式(4)预测用户对项目的评分 $P_{u,i}$
12:对预测评分进行降序排序,并输出推荐列表

3　实验结果及分析

3.1　数据集描述

实验采用美国明尼苏达大学 GroupLens 项目组提

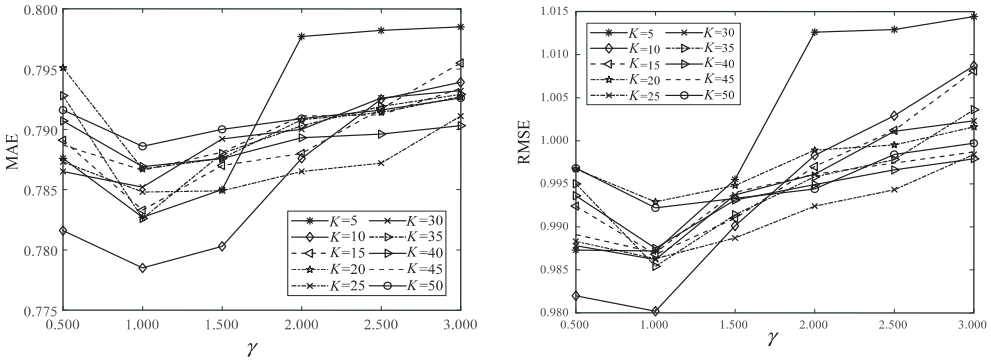
供的 MovieLens 100 K 数据集,每位用户至少对 20 部电影进行过评分,最高评分为 5 分,最低评分为 1 分。该数据包含 943 位用户对 1 682 部电影项目的 100 000 个评分记录,稀疏性为 93.7%。随机抽取 80% 数据作为训练集,剩下的 20% 数据作为测试集。

3.2 参数设置及分析

前期,在对原始项目评分数据中的空缺项进行填充时,偏好阈值参数和权重参数会影响最终的填充效果。由于该文与文献[17]均采用 MovieLens 100 K 数据集进行实验验证,故分别将偏好阈值参数和权重参数设置为 1/60 和 0.1。在后期融合项目惩罚因子降低热门项目的贡献时,平衡参数 γ 和热门阈值 PT 会影响最终的预测结果。所以,先针对这两个参数进行相应的优化实验,再进行对比实验。

3.2.1 平衡参数 γ 对实验结果的影响

该文采用平衡参数 γ 来调节项目热门度对项目热门惩罚因子的影响。将 γ 的取值范围设置为[0.5,3],步长为 0.5,在 PT=0,近邻数 K 的取值范围为[5,50],步长为 5 时进行相关实验,结果如图 1 所示。从图 1 可以看出,γ 值对实验结果有一定的影响。以近邻数 $K=10$ 为例来进行具体分析。MAE 值随着 γ 值增大呈现先减小后增大的增长趋势。当 γ 值较小时,热门项目与非热门项目对相似度的贡献相差不大,对热门项目的惩罚力度不够;当 γ 值较大时,则过分惩罚了热门项目对相似度的贡献。因此要对平衡参数 γ 进行调参,寻找最合适的 γ 值。图 1 表明 γ 最优值取 1。

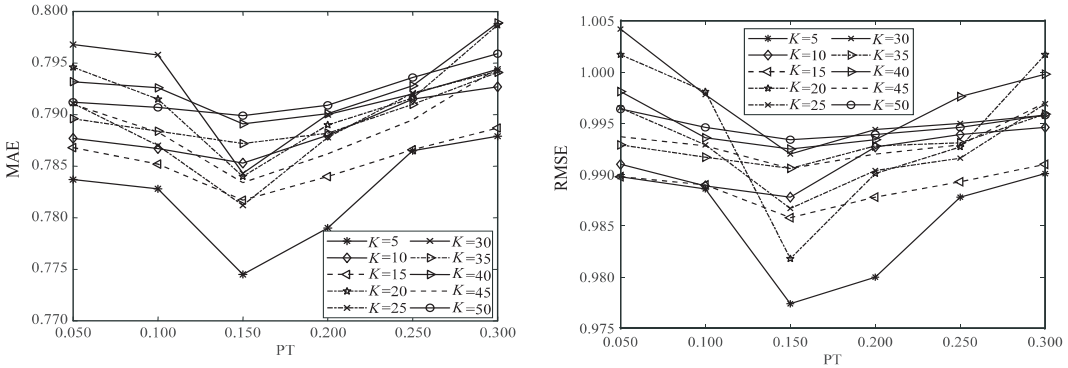


(a)平衡参数 γ 对 MAE 值的影响 (b)平衡参数 γ 对 RMSE 值的影响

图 1 平衡参数 γ 对实验结果的影响

3.2.2 热门阈值 PT 对实验结果的影响

该文用热门阈值 PT 来将项目划分为热门项目和非热门项目,即大于该阈值的项目则为热门项目,反之则为非热门项目。不同的 PT 值划分出的非热门项目集和热门项目集不同。为验证 PT 值对实验结果的影响,将 PT 的取值范围设置为[0.05,0.3],在 $\gamma=1$,近邻数 K 取值范围为[5,50],步长为 5 时进行相关实验,结果如图 2 所示。从图 2 可以看出,PT 值对实验的结果有一定的影响。以近邻数 $K=15$ 为例来具体分析热门阈值 PT 对实验结果的影响,当 PT 值较小时,其对热门项目区分的力度过大,误将非热门项目归为热门项目,导致 MAE 值偏大;当 PT 值较大时,其对热门项目的区分力度过小,误将热门项目归为非热门项目,亦导致 MAE 值偏大。因此,需要对 PT 的值进行调整以获得最合适的值。图 2 表明 PT 的最优值取 0.15。



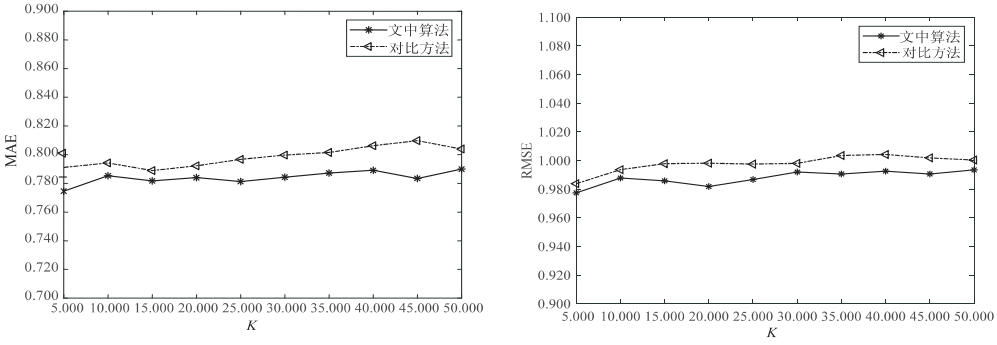
(a)热门阈值 PT 对 MAE 值的影响 (b)热门阈值 PT 对 RMSE 值的影响

图 2 热门阈值 PT 对实验结果的影响

3.3 对照实验及结果分析

为进一步验证所提方法的性能,将所提方法即融合项目热门惩罚因子改进协同过滤推荐方法与文献[17]的方法(用户偏好二分策略和平均评分融合的预评分方法,即对比方法)进行比较。在 MovieLens 100 K数据集上进行测试,不同近邻数(近邻数 K 为

[5,50],步长为5)下的 MAE 和 RMSE 值如图3 所示。从图3 可以看出,随着近邻数 K 的不断增大,这两种方法的 MAE 和 RMSE 的波动比较稳定,即这两种方法受近邻数 K 的影响不大。所提方法在 MAE 和 RMSE 指标上优于对比方法,进一步验证了所提方法的有效性。



(a)不同方法的 MAE 值对比　　(b)不同方法的 RMSE 值对比

图 3　所提方法与文献[17]方法的对比

4　结束语

针对传统协同过滤算法存在的项目热门度偏差问题,即热门项目的推荐频率远远高于非热门项目的推荐频率,提出一种融合项目热门惩罚因子改进协同过滤推荐方法。项目热门惩罚因子有效地缓解了传统协同过滤推荐算法存在的项目热门度偏差问题。在真实数据集上的实验表明,在参数取最优值时,所提方法有效降低了评分预测的误差。但该方法也存在一定的不足,如度量项目热门度时未考虑时间因素,同一项目在不同时期对应的项目热门度可能有所不同等,未来将基于时间要素深入研究项目热门因子在推荐中的作用。

参考文献:

[1] 何佶星,陈汶滨,牟斌皓. 流行度划分结合平均偏好权重的协同过滤个性化推荐算法[J]. 计算机科学,2018,45(6A):493-496.

[2] 王鸿伟. 基于网络特征学习的个性化推荐系统[D]. 上海:上海交通大学,2018.

[3] 孙　红,韩　震. 融合物品热门因子的协同过滤改进算法[J]. 小型微型计算机系统,2018,39(4):638-643.

[4] ABDOLLAHPOURI H. Popularity bias in ranking and recommendation[C]//Proceedings of the 2019 AAAI/ACM conference on AI,ethics,and society. Honolulu:ACM,2019:529-530.

[5] ABDOLLAHPOURI H,MANSOURY M,BURKE R,et al. The unfairness of popularity bias in recommendation[J]. arXiv:1907. 13286,2019.

[6] 赵俊逸,庄福振,敖　翔,等. 协同过滤推荐系统综述[J]. 信息安全学报,2021,6(5):17-34.

[7] 梁贻乐. 面向长尾和冷启动物品的新颖性推荐方法研究[D]. 武汉:武汉大学,2021.

[8] 赵艳枝. 科学研究中的长尾数据及其监护[J]. 情报资料工作,2015,36(3):22-25.

[9] 冯晨娇,宋　鹏,王智强,等. 一种基于3 因素概率图模型的长尾推荐方法[J]. 计算机研究与发展,2021,58(9):1975-1986.

[10] ABDOLLAHPOURI H,BURKE R,MOBASHER B. Controlling popularity bias in learning-to-rank recommendation[C]//Proceedings of the eleventh ACM conference on recommender systems. Como:ACM,2017:42-46.

[11] ZHU Z,HE Y,ZHAO X,et al. Popularity-opportunity bias in collaborative filtering[C]//Proceedings of the 14th ACM international conference on web search and data mining. Virtual Event:ACM,2021:85-93.

[12] SUN C,XU Y. Topic model-based recommender system for longtailed products against popularity bias[C]//2019 IEEE fourth international conference on data science in cyberspace (DSC). Hangzhou:IEEE,2019:250-256.

[13] MEGHAWAT M,YADAV S,MAHATA D,et al. A multimodal approach to predict social media popularity[C]//2018 IEEE conference on multimedia information processing and retrieval (MIPR). Miami:IEEE,2018:190-195.

[14] LIU H,KOU H,YAN C,et al. Keywords-driven and popularity-aware paper recommendation based on undirected paper citation graph[J]. Complexity,2020,2020(1):1-15.

[15] WANG X,SHENG Y,DENG H. Joint deep network with auxiliary semantic learning for popular recommendation[J]. IEEE Access,2020,8:41254-41261.

[16] PIOTRKOWICZ A,DIMITROVA V,OTTERBACHER J,et al. Headlines matter:using headlines to predict the popularity of news articles on twitter and facebook[C]//Proceedings of the international AAAI conference on web and social media. Montreal:AAAI,2017:656-659.

[17] CHENG S,WANG W,YANG S,et al. Effective pre-rating method based on users′ dichotomous preferences and average ratings fusion for recommender systems[J]. Journal of Information Processing Systems,2021,17(3):462-472.