王瑞阳, 章韵. 基于 Jaccard 系数和修正因子的协同过滤算法[J]. 智能计算机与应用,2025,15(10):175-180. DOI:10.20169/j.issn.2095-2163.251027

基于 Jaccard 系数和修正因子的协同过滤算法

王瑞阳, 章 韵
(南京邮电大学 计算机学院, 南京 210023)

摘 要: 随着信息技术不断发展,推荐算法成为处理海量信息的一种有效方式。然而,传统协同过滤算法存在一些弊端。首先,热门项目在计算相似度时被过度考虑,不能反映用户的真实需求。其次,评分矩阵存在数据稀疏问题,导致推荐准确性不高。因此本文提出一种改进用户相似度的协同过滤推荐算法,在余弦相似度和修正余弦相似度的基础上引入 Jaccard 系数与修正因子,来减少热门项目和稀疏矩阵的影响,并在 MovieLens 数据集上验证了本文算法的有效性。

关键词: 协同过滤算法; Jaccard 系数; 修正因子; 相似度计算
中图分类号: TP312 **文献标志码**: A **文章编号**: 2095-2163(2025)10-0175-06

Collaborative filtering algorithm based on Jaccard coefficient and correction factor

WANG Ruiyang, ZHANG Yun
(School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: With the continuous development of information technology, recommendation algorithms have become an effective way to process massive amounts of information. However, there are some drawbacks of traditional collaborative filtering recommendation algorithms. First of all, popular items are overconsidered when calculating similarity and do not reflect the real needs of users. Secondly, the scoring matrix has the problem of data sparseness, resulting in low recommendation accuracy. Therefore, this paper proposes a collaborative filtering recommendation algorithm to progress user similarity, introduces Jaccard coefficient and correction factor on the basis of cosine similarity and modified cosine similarity to reduce the influence of popular items and sparse matrix, and verifies the effectiveness of the proposed algorithm on the MovieLens dataset.

Key words: collaborative filtering algorithm; Jaccard coefficient; correction factor; similarity calculation

0 引 言

随着移动互联网的迅猛发展,人们正置身于一个信息过载的时代。在这个时代中,信息生产者很难确保所提供的内容能够有效地传达给潜在的信息消费者,而信息消费者则面临着从海量信息中找到自己感兴趣的内容的挑战。这就是为什么推荐系统变得如此重要,因其充当了信息生产者和信息消费者之间的桥梁。与传统的搜索引擎相比,推荐系统[1]更为智能和高效。根本就无需用户明确提出需求,而是主动根据用户的历史行为和兴趣,就能向其推荐可能感兴趣的内容[2],提高了信息获取的效率。

推荐系统通常以互联网平台为载体,通过分析用户行为日志等数据,了解用户兴趣和需求。然后,将利用这些信息来给不同用户个性化地展示不同的内容,以提高网站或移动应用的点击率、转化率、留存率等关键指标。总地来说,个性化推荐系统[3]在现代信息技术环境中发挥着至关重要的作用。不仅能够帮助信息生产者将内容有效传达给目标受众,还可使信息消费者更轻松地获取感兴趣信息,从而促进了互联网平台的发展和用户体验的提升。

1 相关工作

协同过滤推荐算法是最经典、最常用的推荐算

基金项目: 青年项目基金(62205156)。
作者简介: 王瑞阳(1999—),男,硕士研究生,主要研究方向:推荐算法。
通信作者: 章 韵(1963—),男,博士,教授,硕士生导师,主要研究方向:计算机网络技术与应用。Email:zhy@njupt.edu.cn。
收稿日期: 2023-12-25

法[4]。其基本思想是:根据每一个使用者的喜好以及喜好相近的使用者的选择来给用户推荐物品。基于对用户历史行为数据的挖掘发现用户的喜好偏向,并预测用户可能喜好的产品进行推荐[5]。一般是仅仅基于用户的行为数据,而不依赖于项目或者用户的任何附加信息。目前,得到广泛应用的协同过滤算法主要基于邻域的方法,包括基于用户的协同过滤算法(UserCF)和基于物品的协同过滤算法(ItemCF)。前者给使用者推荐与其喜好相近的用户喜爱的产品。后者给使用者推荐喜欢的物品相似的物品,或者是喜欢物品的配套产品。在实际应用中,协同过滤推荐算法已经被广泛应用于各个领域,如电子商务、社交网络、新闻媒体、音乐电影、医疗领域[6]等。随着信息量的增加和用户需求的变化,推荐系统也在不断地发展和完善。但是在互联网实际应用中,协同过滤推荐算法依旧存在着诸多问题,如推荐拓展性差、冷启动、数据矩阵较为稀疏[7]等,影响了推荐系统的速度和精度。为了降低这些因素的影响,提高推荐算法的准确性,已有诸多学者就此进行了更深层次的研究。

吴锦昆等学者[8]提出了一种方法,通过引入差异因子来解决用户在不同评价体系中存在偏差的问题,以改进相似度计算。曾安等学者[9]的研究考虑了时间长短对用户评分矩阵的影响以及用户之间的非对称影响度,对相似度计算方法进行了改进,以提供更精确的推荐。Liao 等学者[10]的研究利用信息素来反映使用者喜好的实时变化,从而提高了推荐算法的准确性。Wangwatcharakul 等学者[11]采用联合分解方法,结合潜在因素与评论文字的相关关键词,以反映使用者评分矩阵中的用户喜好动态,从而提高推荐推荐算法的精度。Joorabloo 等学者[12]提出一种新型推荐算法,考虑未来的相似性发展规律,重新排列用户或项目邻域集,以提升协同过滤算法的精确度。李亚欣等学者[13]提出了一种基于改进蚁群算法的动态协同过滤推荐算法,旨在更准确地描述用户特性。任永功等学者[14]主要探讨了完全和非完全冷启动问题,通过构建关系网络来解决这些问题,并挖掘近邻用户的选取方法,然后融入到协同过滤算法中。

上述研究虽在一定程度上增加了推荐算法的精确度,但未能充分解决项目热门度和评分矩阵稀疏性对计算用户相似度的双重影响。因此,本文创新性地提出一种基于融合 Jaccard 系数[15]和修正因子的改进用户相似度的协同过滤算法(Jaccard Factor and Correction Factor Collaborative Filtering, JFCFCF),以提升用户相似度的准确率与推荐算法的性能。

2 协同过滤算法推荐过程

协同过滤算法在 1992 年被提出,是最经典的推荐算法之一。协同过滤算法是在获取到用户信息后,寻找目标用户的近邻集合[16],选择近邻集合中感兴趣的对象推荐给目标用户,其具体步骤如图 1 所示。

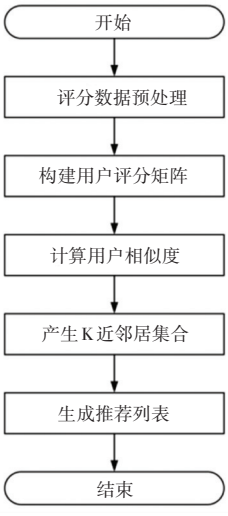

图 1 算法推荐过程

Fig. 1 Recommendation process of the algorithm

2.1 构建用户–评分矩阵

上述研究介绍了用户行为数据的多样性和详细性,并指出第一步是将用户反馈的语言文字进行处理,以将其转换为数字化信息。这个预处理的结果是建立了一个 $m \times n$ 的矩阵 $\boldsymbol{S}(m,n)$,用于表示用户对项目的喜好信息。在这个矩阵中,m 表示用户的数量,n 表示项目的数量,而矩阵元素 S_{ij} 表示用户 i 对项目 j 的评分值。这个评分值的高低反映了用户对项目的喜好程度,用户–评分矩阵[17]如下所示:

$$\boldsymbol{S} = \begin{bmatrix} S_{11} & \cdots & S_{1n} \\ \vdots & \ddots & \vdots \\ S_{m1} & \cdots & S_{mn} \end{bmatrix} \tag{1}$$

通过这个矩阵,能够量化用户与项目之间的关系,并基于这些信息进行个性化的推荐。用户的相似程度可以用相似度计算公式去衡量,从而找出相似性较高的用户组。

2.2 计算相似度

余弦相似度是通过计算 2 个向量的夹角的余弦值来评估两者的相似度[18],夹角的值越小意味着相似度程度越高。余弦相似度的定义公式为:

$$\mathrm{sim}_C(a,b)=\frac{\sum_{t_{i\in I_{ab}}} r_{a,i}r_{b,i}}{\sqrt{\sum_{t_{i\in I_{ab}}} r_{a,i}^{\ 2}}\sqrt{\sum_{t_{i\in I_{ab}}} r_{b,i}^{\ 2}}} \tag{2}$$

修正余弦相似度在计算中涉及到评分偏向因子，使得推荐算法精度更高。修正余弦相似度的计算公式如下：

$$\mathrm{sim}_{AC}(a,b)\ \frac{\sum_{i\in I_{ab}}(r_{ai}-\bar{r}_i)(r_{bi}-\bar{r}_i)}{\sqrt{\sum_{i\in I_{ab}}(r_{ai}-\bar{r}_a)^2}\sqrt{\sum_{i\in I_{ab}}(r_{bi}-\bar{r}_b)^2}} \tag{3}$$

其中，$\mathrm{sim}(a,b)$ 用来度量用户 a 与用户 b 之间的相似性程度，值越大则相似性程度越高；r_{ai} 表示用户 a 对项目 i 的评分值，r_{bi} 表示用户 b 对项目 i 的评分值；I_{ab} 表示用户 a 与用户 b 的共同评分项目组；$\bar{r}_i$ 表示所有评分者对项目 i 的评分值的均值；$\bar{r}_a$、$\bar{r}_b$ 分别表示评分者 a 与评分者 b 对所有存在评分的项目的评分均值。

2.3　生成邻居集

邻居集是一组与当前用户具有相似兴趣和偏好的用户的集合，这种技术被称为最近邻居选择。最近邻居选择的核心思想是根据用户之间的相似度来确定邻居，并且使用相似度作为权重来衡量用户的重要性。一般情况下，会选择 Top－N 相似用户作为目标用户的邻居集[19]。

2.4　生成推荐集

确定了目标用户的邻居集后，结合项目的所有邻居的评分信息以及用户之间的相似度来预测目标用户在某个特定项目中的评分，通常会从评分集中选择 Top－N 条记录作为最终的推荐结果。

假设有一个目标用户 u 和一个目标项目 i，那么可以使用预测分数来估算用户 u 对项目 i 的评分。预测分数可由下式来求得：

$$p_{u,i}=\bar{r}_u+\frac{\sum_{l\in N_u}\mathrm{sim}(u,l)\times(r_{l,i}-\bar{r}_l)}{\sum_{l\in N_u}|\mathrm{sim}(u,l)|} \tag{4}$$

其中，N_u 表示用户 u 的最近邻用户集合；$\bar{r}_u$、$\bar{r}_l$ 分别表示用户 u、l 对所有项目的平均评分；$r_{l,i}$ 表示用户 l 对项目 i 的实际评分；$p_{u,i}$ 表示用户 u 对项目 i 的预测评分。

3　本文改进算法

随着用户评分的项目不断增加，建立的用户－评分矩阵变得越来越稀疏。这意味着很多用户可能只对少数项目进行了评分，导致产生大规模但稀疏的评分矩阵[20]，即由于评分矩阵的稀疏性，传统协同过滤算法在处理缺乏数据的情况下可能效果不佳，难以找到足够的邻居用户或项目。而且一些传统算法在计算相似性时可能过于关注热门项目，而忽视了冷门项目或个性化需求。虽然传统的协同过滤算法在互联网推荐算法中应用广泛，但其寻找到的近邻集合准确度仍然较低。为此，本文将在 2 个方面对原始算法进行创新性改进。

3.1　修正因子

当热门物品频繁出现时，实际的相似度计算结果可能受到显著影响，导致推荐的物品主要集中在热门物品[20]上，无法真正满足用户的个性化需求。例如在亚运期间的体育网站主页里，女子排球、百米赛跑是经常被推荐的内容，而卡巴迪、壁球等运动不常出现，这是因为热门项目在很多用户行为数据中出现的频率很高，使其在相似度计算中权重过大，这种情况可能掩盖了用户对于其他类别冷门项目的真实兴趣。

为了解决这个问题，本文的改进算法提出引入一个修正因子 $f(ab)$ 作为加权系数，抑制热门物品对相似度计算产生的过大影响，对相似度计算值进行修正。具体而言，本文将项目出现次数的倒数作为修正因子的分子。这意味着项目出现次数越多、即越热门，其对于用户兴趣相似度的贡献越小。

本文提出的修正因子定义公式如下：

$$f(ab)=\frac{\sum_{i\in N(a)\cap N(b)}\frac{1}{N(i)}}{\sqrt{|N(a)|\cdot|N(b)|}} \tag{5}$$

其中，i 表示用户 a 与用户 b 的所有已评分项目中的共同项目集合，$N(i)$ 表示项目 i 出现的次数。通过引入这个修正因子，相似度计算可以更好地平衡热门物品和冷门物品之间的影响。热门物品的高频率出现不再过分强调物品的相似度，而是根据其频率进行适当的惩罚，以便更好地挖掘用户的实际需求。

3.2　Jaccard 系数

针对传统协同过滤算法在处理缺乏数据的情况下效果不佳问题，难以找到足够的邻居用户。例如用户 a 只对一个体育项目 Q 进行了评分，而用户 b 对包括 Q 在内的多个体育项目进行了评分，由于项目 Q 都出现在用户 a 和 b 的评分矩阵里，此时传统的协同过滤算法会计算得出用户 a 和 b 的用户相似度很高，但事实上并非如此，这是由于用户 a 的评分项目过少、评分矩阵稀疏造成的，因此本文提出引入

Jaccard 系数来解决用户-评分矩阵稀疏的问题。这里用到的公式具体如下：

$$s(a,b)=\frac{|I(a)\cap I(b)|}{|I(a)\cup I(b)|} \tag{6}$$

其中，$I(a)$、$I(b)$ 分别表示用户 a、b 各自所评分项目的集合；用户共同评分的项目和评分项目的总和通过 2 个集合的交集和并集来表示。Jaccard 系数的取值范围在 0~1，当 Jaccard 系数等于 0 时，表示 2 个集合没有任何共同元素，所以完全不相似。当 Jaccard 系数接近 1 时，表示 2 个集合具有较高的相似性，所以共享许多相同的元素。

3.3 改进的相似度计算公式

为了降低上述 2 种情况对推荐效果的影响，最终提出将上述修正因子和 Jaccard 系数结合相似度计算公式按照 α、$(1-\alpha)$ 的权重进行融合，最终得到的相似度计算公式：

$$\mathrm{sim}(a,b)=\alpha\,\mathrm{sim}_C(a,b)f(a,b)s(a,b)+(1-\alpha)\,\mathrm{sim}_{AC}(a,b)f(a,b)s(a,b) \tag{7}$$

其中，研究推得的公式为：

$$\mathrm{sim}_C(a,b)=\frac{\sum\limits_{t_{i\in I_{a,b}}} r_{a,i}r_{b,i}}{\sqrt{\sum\limits_{t_{i\in I_{a,b}}} r_{a,i}^{\,2}}\sqrt{\sum\limits_{t_{i\in I_{a,b}}} r_{b,i}^{\,2}}} \tag{8}$$

$$\mathrm{sim}_{AC}(a,b)=\frac{\sum\limits_{i\in I_{ab}}(r_{ai}-\bar{r}_i)(r_{bi}-\bar{r}_i)}{\sqrt{\sum\limits_{i\in I_{ab}}(r_{ai}-\bar{r}_a)^2}\sqrt{\sum\limits_{i\in I_{ab}}(r_{bi}-\bar{r}_b)^2}} \tag{9}$$

3.4 生成邻居集和推荐集

以相似度作为权重，选择 Top - N 用户作为目标用户的邻居集。将上述本文改进的相似度算法带入到式(4) 中，最终选择 Top - N 记录作为推荐结果。图 2 是本文改进的协同过滤推荐算法的具体步骤。

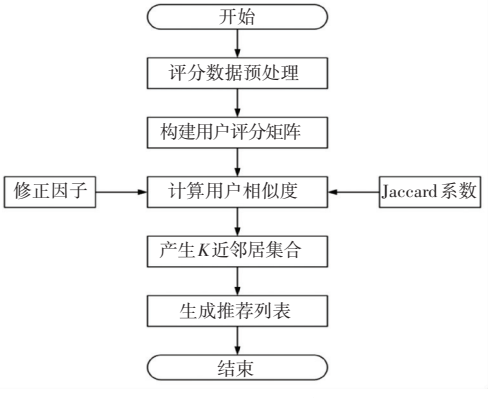

图 2 改进的推荐算法

Fig. 2 Improved recommendation algorithm

4 实验结果与分析

4.1 所用数据集与评价指标

本研究选择 MovieLens 数据集作为评估改进推荐算法的实验平台。该数据集由 3 个主要组成部分构成：用户信息数据、电影目录以及评分日志。该数据集涵盖了 1 682 部电影以及 943 名用户所提供的 100 000 条评分数据。实验设计中，数据被分割为两个部分：80% 的数据用作训练集以构建推荐模型，20% 则用作测试集，用以评估模型的性能。

本实验采用平均绝对误差(Mean Absolute Error，MAE) 来衡量预测值与真实值之间差异。MAE 值计算预测值和真实值之间的平均差值，其值越大，误差越大。MAE 定义计算公式如下：

$$\mathrm{MAE}=\frac{\sum\limits_{i=1}^{N}|r_{ui}-\hat{r}_{ui}|}{N} \tag{10}$$

其中，r_{ui}、$\hat{r}_{ui}$ 分别表示用户 u 对项目 i 的预测评分和实际评分；N 表示测试集合的项目评分的数量。MAE 值越低、表明预测结果与真实结果越接近，算法准确度越高。

4.2 实验过程

先将引入修正因子的余弦相似度(ICOS)、引入修正因子的修正余弦相似度(IACOS) 的协同过滤推荐算法在 MovieLens 数据集上进行实验，设置最近邻用户数量 N 变化取值区间为[10，90]，步长设置为 10，从而计算出的 MAE 值如图 3 所示。

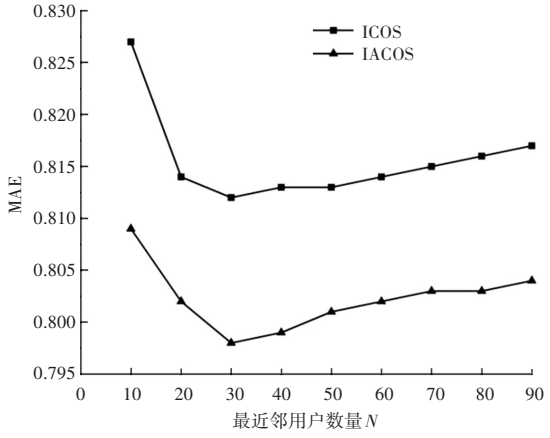

图 3 MAE 值的变化

Fig. 3 Changes in MAE value

由图 3 可知，随着最近邻用户数量 N 的增大，MAE 值呈先减小后增大趋势，并且在 N 为 30 时取最小值，因此接下来的实验中最近邻用户数量定为 30。

本次改进的协同过滤算法中引入了 α 和 $(1-\alpha)$ 两个参数，其中 α 取值范围为[-1,1]。为了确定 α 的取值，对 $N=30$ 情况下 MAE 的取值进行计算，设置 α 的值从 0.1 变化到 0.9、间隔为 0.1，如图 4 所示。

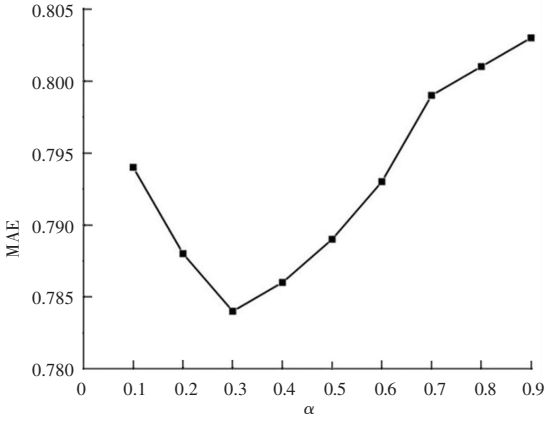

图 4　不同 α 取值下的 MAE 值

Fig. 4　The MAE value under different α values

在 α 取[0.1,0.3]时，最终改进的相似度计算方法的 MAE 值随着 α 值的增大而减小，此后随着 α 值的增大，MAE 值也不断增大。最后，当 α 为 0.3 时，MAE 降到最低，推荐算法精度最高。

在两端单一依赖余弦相似度算法和修正余弦相似度的推荐精确度取到较小值，因此取中间值能结合两者方法的优点，在参数 α 取 0.3 时预测精确度最好。

为了验证本文所提出推荐算法的精确度，将本文提出的 JFCFCF 算法和传统 CF 算法、文献[8]提出的算法进行比较，实验结果如图 5 所示。

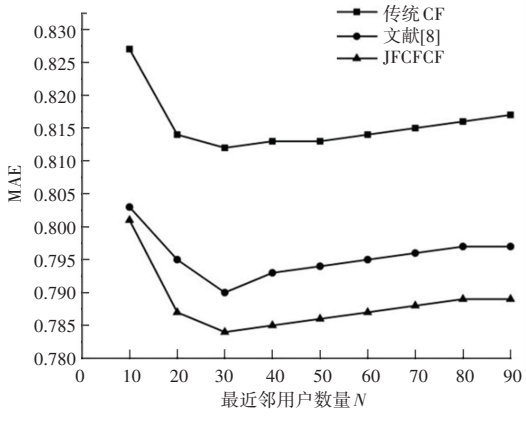

图 5　MAE 值变化对比

Fig. 5　Comparison of changes in MAE values

由图 5 可知，随着 N 值的增大，3 种算法的 MAE 值先减小后增大，变化的速度也逐渐变小，这是由于随着最近邻用户数量增加，计算过程中参考的用户数量也随着增加。当 N 为 30 时，JFCFCF 算法的 MAE 值达到最低，相比于传统 CF 算法，其 MAE 值降低了 3.45%，且 JFCFCF 算法的 MAE 值始终低于传统 CF 算法和文献[8]提出的算法。说明本文提出的 JFCFCF 算法的用户评分预测准确性更高，算法的性能更强，推荐效果更好。

5　结束语

本文通过引入 Jaccard 系数和修正因子，对传统协同过滤算法中相似度计算的不足之处进行改进，创新性地将 Jaccard 系数和修正因子分配不同的权重，进一步整合到余弦相似度和修正余弦相似度计算过程中。在 MovieLens 数据集上对 JFCFCF 算法进行的实验验证表明，与传统的协同过滤算法及其它改进算法相比，JFCFCF 在提升推荐性能方面取得了显著成效，能够更准确地反映用户兴趣。然而，算法同样存在局限性，例如未考虑项目热度随时间变迁而引起的权重变动，即项目在不同时间段的重要性动态变化问题。因此，未来的研究将重点关注这一领域，以实现更为精确和动态的推荐效果。

参考文献

[1] 刘维超，杨有，余平. 基于内容的新闻推荐系统研究综述[J]. 福建电脑，2019，35(9)：71-74.

[2] 白莉婷. 用户画像构建与社群推荐技术研究[D]. 西安：西安电子科技大学，2021.

[3] 蒲彬. 基于社交信号的个性化新闻推荐系统的设计与实现[D]. 北京：中国科学院大学，2015.

[4] 张兰兰. 基于协同过滤的个性化新闻推荐系统的研究与实现[D]. 重庆：重庆大学，2016.

[5] 吴彦文，齐旻，杨锐. 一种基于改进型协同过滤算法的新闻推荐系统[J]. 计算机工程与科学，2017，39(6)：1179-1185.

[6] LV Yihang，KONG Jinjing. Application of collaborative filtering recommendation algorithm in pharmacy system[J]. Journal of Physics：Conference Series，2021，1865：042113.

[7] 刘昊东，王诚. 基于热门度修正因子和置信度的协同过滤算法[J]. 计算机技术与发展，2023，33(3)：127-132.

[8] 吴锦昆，单剑锋. 基于改进型相似度的协同过滤算法的研究[J]. 计算机技术与发展，2022，32(4)：39-43.

[9] 曾安，高成思，徐小强. 融合时间因素和用户评分特性的协同过滤算法[J]. 计算机科学，2017，44(9)：243-249.

[10] LIAO Xiaofeng，WU Hu，WANG Yongyi. Ant colaborative filtering addressing sparsity and temporal effects[J]. IEEE Access，2020，8：32783-32791.

[11] WANGWATCHARAKUL C，WONGTHANAVASU S. Dynamic collaborative filtering based on user preference drift and topic evolution[J]. IEEE Access，2020，8：86433-86447.

[12] JOORABLOO N，JALILI M，REN Y. Improved collaborative filtering recommendation through similarity prediction[J]. IEEE Access，2020，8：202122-202132.

[13] 李亚欣，蔡永香，张根. 结合实时推荐与离线推荐的推荐系统

[J]. 计算机系统应用,2019,28(10):45-52.
[14]任永功,石佳鑫,张志鹏. 融合关系挖掘与协同过滤的物品冷启动推荐算法[J]. 模式识别与人工智能,2020,33(1):75-85.
[15]董仕,马怀祥. 基于改进 Jaccard 系数的证据间相似性度量方法[J]. 石家庄铁道大学学报(自然科学版),2021,34(2):66-71.
[16]黄创光,印鉴,汪静,等. 不确定近邻的协同过滤推荐算法[J]. 计算机学报,2010,33(8):1369-1377.
[17]姜宇,张大方,刁祖龙. 基于点击流的用户矩阵模型相似度个性化推荐[J]. 计算机工程,2018,44 (1): 219-225.
[18] SINGH R, MAURYA S, TRIPATHI T, et al. Movie recommendation system using cosine similarity and KNN[J]. International Journal of Engineering and Advanced Technology, 2020,9(5):556-559.
[19]AIREN S, AGRAWAL J. Movie recommender system using K-Nearest Neighbors variants[J]. National Academy Science Letters, 2021,45:75-82.
[20]徐立民,李涵. 基于惩罚因子的协同过滤算法的改进与研究[J]. 物联网技术,2019,9(10):73-75.