计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(8):224−229 [doi: 10.15888/j.cnki.csa.007561]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

结合信任关系的用户聚类协同过滤推荐算法①

孟　晗[1,2], 高　岑[2], 王　嵩[2], 张琳琳[2], 刘　念[2]

[1](中国科学院大学 计算机控制与工程学院, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)
通讯作者: 孟　晗, E-mail: 1163827828@qq.com

摘　要: 在传统的协同过滤推荐算法中, 相似度计算是算法中的核心, 然而之前的计算方式过于依赖用户的评分, 没有考虑到用户本身的属性以及信任度, 并且没有对恶意用户进行区分, 为解决上诉问题, 本文将一种改进的新型信任关系度量方式融入到相似度计算中, 这种新型的方法不仅考虑了恶意用户的影响, 并且有效地结合用户本身的属性. 另外, 文章就热点问题对相似度计算也进行了改进. 算法最终利用初始用户聚类不断迭代得到相邻用户, 有效的消除了冷启动和数据稀疏的问题. 实验部分, 通过与其它几种推荐算法的比较可以证明, 提出的算法能够有效提升推荐准确度.

关键词: 协同过滤; 信任关系; 相似度算法; 用户聚类; 相邻用户

引用格式: 孟晗,高岑,王嵩,张琳琳,刘念.结合信任关系的用户聚类协同过滤推荐算法.计算机系统应用,2020,29(8):224–229. http://www.c-s-a.org.cn/1003-3254/7561.html

User Clustering Collaborative Filtering Recommendation Algorithm Combined with Trust Relationship

MENG Han[1,2], GAO Cen[2], WANG Song[2], ZHANG Lin-Lin[2], LIU Nian[2]

[1](School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

Abstract: In the traditional collaborative filtering recommendation algorithm, similarity calculation is the core of the algorithm. However, the previous calculation method is too dependent on the user's score, does not consider the user's own attributes and trust relationship, and does not distinguish malicious users. In order to solve the appeal problem, this study introduces an improved new trust relationship measurement method into similarity calculation. This new method not only considers the influence of malicious users, but also combines the properties of users effectively. In addition, the study also improves the similarity algorithm on the hot issues. The algorithm finally uses the initial user clustering to get the adjacent users, effectively eliminating the cold start and data sparsity. In the experimental part, it can be proved that the proposed algorithm can effectively improve the recommendation accuracy by comparing with other algorithms.

Key words: collaborative filtering; trust relationship; similarity algorithm; user clustering; adjacent users

引言

在大数据时代, 互联网给用户带来了大量信息, 但随着网络的不断发展和信息的迅速增长, 用户在浩瀚数据中想要找到所需信息将会愈发困难, 出现了信息超载问题. 目前, 解决信息超载一个非常高效的方法就是推荐系统. 推荐系统简单来说就是一种信息过滤系统, 用于预测用户对物品的喜爱程度. 在推荐系统中, 协同过滤 (Collaborative Filtering) 技术是推荐系统中应

① 收稿时间: 2020-01-11; 修改时间: 2020-03-08; 采用时间: 2020-03-11; csa 在线出版时间: 2020-07-29

用最早和最为成功的技术之一, 其核心是用户的评分数据, 与被推荐的内容无关, 即: 在阅读历史中对新闻评分相似的用户在将来也会相似, 从而把推荐转换为评分预测问题[1]. 同时, 由于协同过滤具备发现用户隐藏兴趣的能力, 并且可以有效的使用其他相似用户的反馈信息, 加快个性化学习速度, 因此逐渐成为最受欢迎的推荐算法.

虽然协同过滤推荐技术有着广泛的应用, 但是也存在很多问题, 比如冷启动、数据稀疏等, 因此研究人员不断的对其进行创新和改善. Lokhande 等[2] 提出一种附加的层次聚类技术, 并且将主成分分析用于数据降维, 提高了传统 CFRA 输出的准确性. Bobadilla 等[3] 研究了神经网络学习算法在冷启动问题中的应用. Zhou 等[4] 提出了功能矩阵分解模型 (functional matrix factorization), 利用决策树和矩阵分解的结合, 在冷启动过程中为用户选择合适的物品进行打分, 从而尽可能准确地理解用户的偏好.

上述研究都是依托于已有的评分数据进行分析, 但依托的这些数据并没有考虑用户间的信任关系, 而一个良好的信任关系模型能够有助于寻找相似度更高的近邻用户, 从而得到更准确的推荐结果. 基于此, Yang 等[5] 采用矩阵分解将用户映射到低维特征空间的信任关系, 更准确的反映了用户之间的相互影响. Xue 等[6] 提出将用户间的信任关系依据方差大小进行量化形成调节因子的思想, 算法将调节因子纳入用户相似性计算, 形成相似性用户聚类簇, 但是上述算法没有在真实应用场景中实验, 并且忽略了用户隐式信任关系导致实验结果不完整. Du 等[7] 创新了社会学领域的信任关系计算, 用其代替传统的相似度计算方法, 将信任度集成到最近邻居选择中. 信任网络是通过扩展不同的路径长度来构建的, 用户之间的信任值可以通过信任传输规则来获得.

综上所述, 很多研究学者为了解决协同过滤存在的问题, 都会考虑降维、矩阵分解、神经网络等方法. 而本文从描述用户关系入手, 结合用户属性和信任传递设计了一种改进的信任机制, 然后融合到相似度算法中, 使其能够更好的反映用户之间的关系, 最后结合聚类算法, 提高了推荐的精确度, 同时解决了协同过滤的冷启动以及数据稀疏问题.

1 协同过滤推荐算法

协同过滤是目前应用最广泛的推荐算法, 其基于系统中其他用户的评分或行为进行预测和推荐. 这种方法认为, 如果两个用户对某些项目有着相同的评价, 那么他们可能会对其他项目也达成一致意见.

基本的协同过滤算法通常经过以下几个步骤:

步骤 1. 收集数据并建立用户评分矩阵. 矩阵中行代表用户, 列代表项目, 元素值为具体的评分值, 一般在 0~5 之间, 评分越大喜爱程度越高.

步骤 2. 计算每个用户与其他用户的相似度. 皮尔逊相关系数是目前最常见的计算方式, 其公式如下:

$$sim(a,b)=\frac{\sum_{i\in I_{a,b}}(R_{a,i}-\overline{R_a})(R_{b,i}-\overline{R_b})}{\sqrt{\sum_{i\in I_{a,b}}(R_{a,i}-\overline{R_a})^2}\sqrt{\sum_{i\in I_{a,b}}(R_{b,i}-\overline{R_b})^2}} \tag{1}$$

式中, a,b代表两个用户, $R_{a,i}$表示用户 a 对项目 i 的评分, $I_{a,b}$表示用户a,b共同评分的项目, $\overline{R_a}$, $\overline{R_b}$分别表示用户 a 和 b 所有评分的平均值.

步骤 3. 寻找目标用户最近邻. 依据相似度选取与目标用户最相似的邻居集合作为目标用户的最近邻.

步骤 4. 生成推荐结果. 根据确定好的邻居集合, 通过式 (2) 计算出预测值, 最后依据预测评分的排序进行推荐, 得到最终的结果.

$$P_{a,j}=\overline{R_a}+\frac{\sum_{b\in UNN(a)}sim(a,b)(R_{b,j}-R_a)}{\sum_{b\in UNN(a)}|sim(a,b)|} \tag{2}$$

由上述算法可以看出, 在整个计算过程中, 最为核心的是相似度的计算以及最近邻的选取, 但是相似度计算公式中没有考虑用户间的隐性关系, 也没有考虑一个用户本身的可信任度问题, 更可预见的是, 在推荐系统初始化时, 由于数据稀疏还会产生冷启动问题. 同时, 最近邻的选取上也过于简单化, 容易出现推荐结果不准确的情况. 基于以上对传统协同过滤算法的分析, 本文将结合信任机制对相似度计算进行改进, 并且结合聚类算法对最近邻的选取合理优化.

2 信任关系

在推荐系统中, 信任关系指的是目标用户对其他用户评分可靠性的一种主观认可程度, 认可程度越高则信任关系越密切. 一般来说, 信任关系可以分为直接信任与间接信任.

2.1 直接信任

2.1.1 传统的信任关系模型

已有的描述信任关系[8]是由初始信任度与交互权重相乘得到的, 如式 (3) 所示. 其中$Init(a,b)$为初始信任度, 由式 (4) 表示, t 表示用户 a 和用户 b 达到信任的最小交互次数, 参考文献 [8] 的实验结果可设置 t 为 80. $I_a \cap I_b$ 表示两个用户的交互次数. C_s 表示成功交互次数, 所谓成功交互指两个用户在交互项目的评分差值小于 2. C_a 指的是交互总数.

$$DTrust(a,b) = Init(a,b)\frac{C_s}{C_a} \tag{3}$$

$$Init(a,b) = \frac{\min(I_a \cap I_b, t)}{t} \tag{4}$$

2.1.2 改进的信任关系模型

改进 1. 传统的信任模型虽然可以成功表示两个用户的信任关系, 但是忽略了恶意用户或者倾向于给高低分的用户, 如果一个用户比较严格, 评价分数普遍过高或者过低, 则认为其不是一般性用户, 此用户在信任机制中的权重则应该减少. 因此这里我们引入一个用户本身可信任度, 公式如下:

$$STrust(b) = \frac{\min(C(R_b=1,2), C(R_b=4,5))+1}{\max(C(R_b=1,2), C(R_b=4,5))+1} \tag{5}$$

其中, $C(R_b=1,2)$代表用户 b 评价分数为 1 和 2 的总次数, $C(R_b=4,5)$代表用户 b 评价分数为 4 和 5 的总次数. 此式可以衡量出低分和高分之间的比例, 若两者相差较大证明该用户趋向于高分或者低分, 从而可信任度下降. 将其融合到信任模型中得到式 (6):

$$DTrust(a,b) = Init(a,b)\frac{C_s}{C_a} \times STrust(b) \tag{6}$$

式 (6) 在原有的信任机制中引入用户本身可信任度, 经过计算之后, 如果打分比较正常的用户, 其信任度变化不大, 而恶意用户或者比较严格的用户其信任度会有所下降, 通过这种改进有效地区分了两种用户对相似度计算的影响.

改进 2. 增加两个用户的本身属性到可信度中. 一般而言, 人们更倾向于相信与自己特征类似的群体. 这里我们采取 3 个用户属性融合到信任模型中, 分别是用户年龄、性别和职业, 这里参考文献 [9] 及其实验结果, $S(a,b), A(a,b), O(a,b)$分别代表性别、年龄和职业的相似性, 公式分别如下:

$$S(a,b) = \begin{cases} 0, & S_a \neq S_b \\ 1, & S_a = S_b \end{cases} \tag{7}$$

$$A(a,b) = \begin{cases} 1, & |A_a - A_b| \leq 5 \\ \dfrac{5}{|A_a - A_b|}, & |A_a - A_b| > 5 \end{cases} \tag{8}$$

$$O(a,b) = \begin{cases} \dfrac{H_{a,b}}{Height}, & a \neq b \\ 1, & a = b \end{cases} \tag{9}$$

得到用户之间的属性相似性[9]如式 (10) 所示:

$$ATrust(a,b) = 0.5S(a,b) + 0.1A(a,b) + 0.4O(a,b) \tag{10}$$

组合到信任模型最终得到直接信任公式为:

$$DTrust(a,b) = Init(a,b)\frac{C_s}{C_a} \times STrust(b) + ATrust(a,b) \tag{11}$$

2.2 间接信任

由于评分矩阵一般都是稀疏矩阵, 所以直接信任数据过于稀少, 因此需要引入间接信任, 即用户间无直接信任关系, 但存在至少一条可达路径, 从而建立的信任关系. 间接信任用$ITrust(a,b)$来表示. 根据六度分割理论可知信任传递最大路径长度为 6, 依据文献 [10] 的研究设置信任传递最大路径长度 d 值为 3.

假设用户 a 和 b 之间存在至少一条路径, 则路径集合表示为$Paths(a,b) = \{p_1(a,b), p_2(a,b), \cdots, p_n(a,b)\}$每条路径可表示为$p_i(a,b) = (a_i, c_{i1}, c_{i2}, \cdots, b_i)$. 每条路径权重 α_i 随路径长度呈递减趋势, 计算方式如式 (12) 所示, 其中$dp_i(a,b)$表示第 i 条路径的长度:

$$\alpha_i = \frac{d - dp_i(a,b) + 1}{d} \tag{12}$$

则每条路径的信任度可表示为:

$$ITrustp_i(a_i,b_i) = \alpha_i \times (DTrust(a_i,c_{i1}) \times DTrust(c_{i1},c_{i2}) \cdots DTrust(c_{ip},b_i)) \tag{13}$$

由此得到间接信任模型如式 (14) 所示, 其中 n 表示用户 a 和用户 b 可以可达路径总数, 最终得到的用户 a 和 b 的间接信任度通过取平均值的方式计算得出.

$$ITrust(a,b) = \frac{\sum_{i \in n} ITrustp_i(a_i,b_i)}{n} \tag{14}$$

2.3 融合信任模型

融合信任模型指的是由直接信任和间接信任采用加权因子融合而成, 用来综合考虑用户间信任关系的一种评估方式, 其中的加权因子λ取值在 0−1 之间, 用

来衡量两个信任度的权重, 其计算公式如式 (15) 所示:

$$\lambda=\frac{DTrust(a,b)}{DTust(a,b)+ITrust(a,b)} \tag{15}$$

最后通过计算参数值, 得到最终的信任模型$Trust(a,b)$如式 (16) 所示:

$$Trust(a,b)=\lambda DTrust(a,b)+(1-\lambda)ITrust(a,b) \tag{16}$$

3 相似度算法改进

3.1 融合热点的皮尔逊相关系数

由第 1 节介绍可知目前大多是用皮尔逊相关系数进行相似度的计算. 但是该计算方式忽略了热点新闻的影响程度, 一条新闻越受欢迎, 多个用户共同点击的概率就会越大, 对相似度影响的权重就会越低, 基于此, 提出一种融合热点的皮尔逊相关系数 HS-P 如式所示:

$$MSim(a,b)=\frac{\sum_{i\in I_{a,b}}\left(\frac{\overline{C}}{C_i}\right)(R_{a,i}-\overline{R_a})(R_{b,i}-\overline{R_b})}{\sqrt{\sum_{i\in I_{a,b}}(R_{a,i}-\overline{R_a})^2}\sqrt{\sum_{i\in I_{a,b}}(R_{b,i}-\overline{R_b})^2}} \tag{17}$$

其中, $I_{a,b}$代表用户 a 和 b 共同评价过的项目集合, $R_{a,i}$代表用户 a 对项目 i 的评价值, $\overline{R_a}$为用户 a 的评价平均值, C_i 为用户 a 和 b 共同评价的某个项目 i 的被评价次数, $\overline{C}$为所有项目被评价次数的平均值. 从上述公式可以看出一个项目被评价的次数越多, 则其在相似度计算中所占权重越低.

3.2 最终的用户相似度公式

结合第 2 节的用户信任关系以及上述的 HS-P 公式, 得到最终改进的衡量用户间相似度公式为:

$$USim(a,b)=\mu Trust(a,b)+(1-\mu)MSim(a,b) \tag{18}$$

4 算法过程描述

4.1 用户聚类迭代算法

依据之前对相似度算法的改进, 可以将其用到聚类算法之中, 通过不断迭代的方式调整用户聚类簇, 让其达到稳定, 这种方式可以更加准确的寻找到目标用户的最近邻. 具体的算法步骤如算法 1 所示.

4.2 推荐算法整体流程

本文的算法整体流程如图 1 所示. 当用户聚类算法迭代完成, 也就是用户聚类集合趋于稳定之后, 再寻找最近邻的可靠性就会大大增加. 假设查找目标用户 a 的 k 个邻居, 首先计算用户 a 与所有聚类中心的相似度, 然后优先选取相似度最高的聚类, 计算 a 与该聚类中各个用户的相似度, 选取前 k 个即为最近邻, 记为 $UKN(a)$, 则用户 a 对项目 i 的预测评分为:

$$P_{a,i}=\overline{R_a}+\frac{\sum_{b\in UKN(a)}USim(a,b)\times(R_{b,i}-\overline{R_b})}{\sum_{b\in UKN(a)}USim(a,b)} \tag{19}$$

其中, $\overline{R_a},\overline{R_b}$分别表示用户 a 和 b 的平均评分.

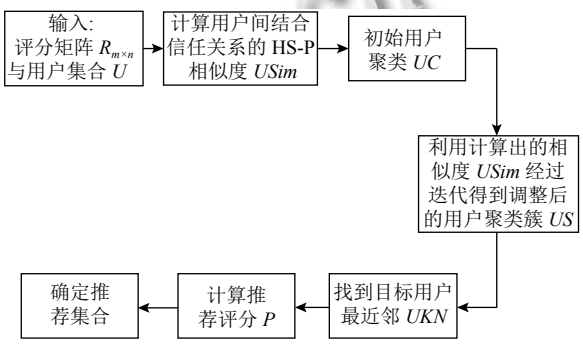

图 1 算法整体流程图

最终通过预评分公式预测出目标用户 a 对项目的评分值, 选取评分值最高的前 N 个项目作为推荐结果.

算法 1. 用户聚类迭代算法

输入: 用户集合 U, 评分矩阵 $R_{m\times n}$
输出: 调整后的用户聚类

(1) 首先用 K-mean 聚类算法对初始的用户集合进行聚类, 得到初始用户聚类$UC=\{UC_1,UC_2,\cdots,UC_k\}$和初始用户聚类中心为$C=\{C_1,C_2,\cdots,C_k\}$
(2) //迭代得到最佳用户聚类集合
repeat
for each user $U_i\in U$
for each user cluster $C_j\in C$
计算 U_i 与聚类中心 C_j 的相似度 $USim(U_i, C_j)$
end for
$USim(U_i,C_t)=\max\{\text{sim}(U_i,C_1),\cdots,\text{sim}(U_i,C_k)\}$
U_i 所属聚类 $UC_s=UC_s-U_i$
$UC_t=UC_t+U_i$
end for
for each $UC_i\in UC$
更新聚类中心 C_i
until 聚类簇中元素不再变化或者达到设定迭代次数

5 实验

5.1 实验数据

本次实验选择了明尼苏达大学 GroupLens 项目组收集的 MovieLens 数据集, 该数据包含了 943 位用户

对 1682 部电影的十万条打分记录, 打分标准从 1 到 5, 经过计算得到稀疏度达到了 93.7%. 另外, 该数据集还包含了用户的特征属性信息, 满足本文对信任关系的改进要求. 实验时随机选取数据集中的 80% 作为训练集, 20% 作为测试集.

5.2 实验评分标准

实验标准采取均绝对误差 *MAE* (Mean Absolute Error) 进行评估, 该值是用来衡量预测值与真实值之间的误差, 值越小, 误差就越小, 推荐的准确率也就更高. 其计算公式为:

$$MAE = \frac{\sum_{i=1}^{n} P_i - R_i}{n} \tag{20}$$

其中, n 表示预测的项目个数, P_i 表示预测的评分值, R_i 表示真实的评分值.

5.3 实验结果与分析

实验 1. 在聚类算法中, 首先需要确定好聚类的数值大小. 这里采取传统的用户聚类协同过滤算法进行实验, 其中迭代时计算用户间相似度采用原始的皮尔逊相关系数. 聚类数值从 10 增加到 50, 查看对应的 *MAE* 值大小, 实验结果如图 2 所示.

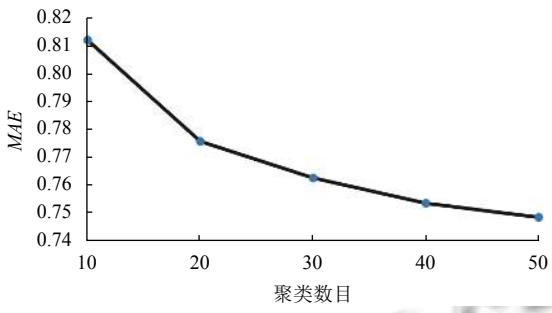

图 2 不同聚类次数对应的 *MAE* 值

从图 2 可以看出 *MAE* 值的变化随着聚类数值的增大而缩小, 当 K 达到 50 时 *MAE* 有最低值, 这里考虑到算法效率问题, 选取聚类数为 50 进行后续实验.

实验 2. 对于用户聚类迭代算法, 不容易判断何时聚类簇稳定, 因此需要考虑迭代次数对 *MAE* 的影响. 实验中设定相似度权重因子μ=0.5, 邻居个数设置 N=10, 调整聚类迭代次数从 1 到 10, 观察不同迭代次数对 *MAE* 的影响如图 3 所示.

从图中可看出当迭代次数从 7 开始 *MAE* 值保持稳定, 此时可以设定聚类簇达到稳定状态. 因此后续实验中, 设定迭代次数为 7 次.

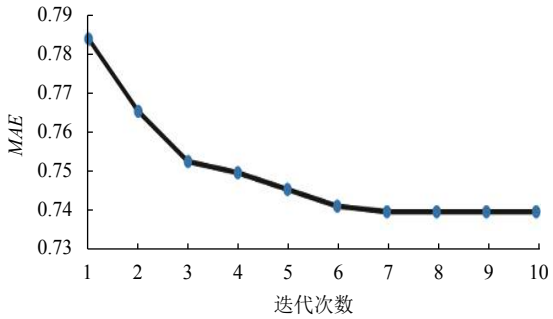

图 3 不同迭代次数对应的 *MAE* 值

实验 3. 该实验的目的是确定式 (15) 中权重因子μ的取值, 实验中确设定迭代次数为 7 次, 邻居个数取值为 10, 图 4 比较了当μ取值在 0–1 之间时 *MAE* 的相应值.

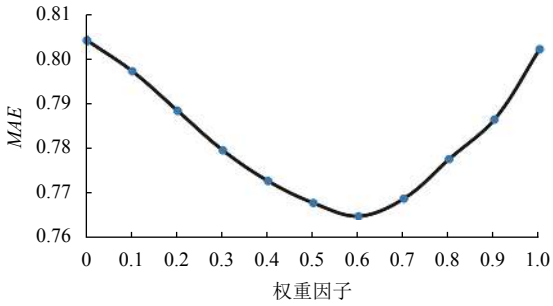

图 4 不同权重因子对应的 *MAE* 值

从图 4 中可以看出, μ的值不宜过大或者过小, 当μ=0.6 时, 即信任关系占据权重 0.6, HS-P 相似度算法占据权重 0.4 时, *MAE* 有最低值, 此时推荐结果最准确.

实验 4. 为了验证本文提出的结合信任关系的用户聚类协同过滤推荐算法 (K-TUBCF), 实验选取了几个传统算法[10] 比较它们的 *MAE* 值, 算法包括传统的基于用户的协同过滤算法 (UBCF), 基于 K-means 聚类的协同过滤推荐算法 (K-UBCF), 融入传统信任关系的协同过滤算法 (TUBCF), 融入用户喜好度到信任关系中的协同过滤算法 (WTUBCF), 最终结合文献 [8,10] 的数据得出实验对比结果如图 5 所示.

从图 5 中可以看出, 这几种算法随着邻居个数的增加 *MAE* 值逐渐下降, 并且在邻居个数达到 30 时趋于稳定, 说明邻居个数 N=30 时会取得最好的推荐结果.

另外, 本文提出的算法在 *MAE* 值上明显低于之前的各类算法, 说明本文这种基于改进信任关系的用户聚类算法可以有效的提高算法准确度, 产生更准确的推荐结果.

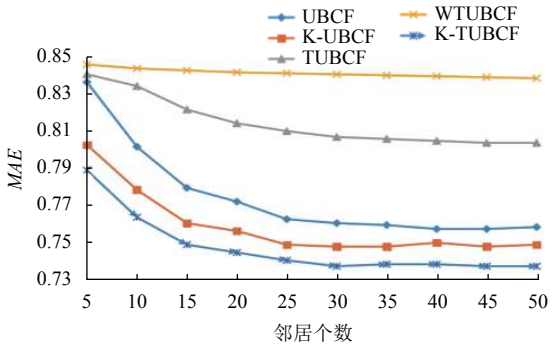

图 5 不同算法之间的比较

6 结论

本文提出的结合信任关系的用户聚类协同过滤推荐算法, 在直接信任与间接信任两个层面上都做了一定改进, 并且有效的结合用户的属性特征. 另外, 对于传统的皮尔逊相似度忽略热点新闻的缺点进行了补充, 最后采用迭代的聚类算法求出最近邻从而得到更为准确的推荐结果. 实验也证明所提出的优化算法比之前的算法误差值更小. 但是文章没有考虑到时间对推荐的影响, 下一步将结合时间因子对推荐算法进行更详细的研究, 使其更加满足实际情况.

参考文献

1 王绍卿, 李鑫鑫, 孙福振, 等. 个性化新闻推荐技术研究综述. 计算机科学与探索, 2020, 14(1): 18–29.

2 Lokhande A, Jain P. Hybrid collaborative filtering model using hierarchical clustering and PCA. Proceedings of Recent Advances in Interdisciplinary Trends in Engineering & Applications (RAITEA). 2019. [doi: 10.2139/ssrn.3365525]

3 Bobadilla J, Ortega F, Hernando A, *et al*. A collaborative filtering approach to mitigate the new user cold start problem. Knowledge-Based Systems, 2012, 26: 225 –238. [doi: 10.1016/j.knosys.2011.07.021]

4 Zhou K, Yang SH, Zha HY. Functional matrix factorizations for cold-start recommendation. Proceedings of the 34th international ACM SIGIR conference on Research and Development in Information Retrieval. Beijing, China. 2011. 315–324.

5 Yang B, Lei Y, Liu JM, *et al*. Social collaborative filtering by trust. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(8): 1633–1647. [doi: 10.1109/TPAMI.2016.2605085]

6 Xue FL, Liu JL. Improving collaborative filtering recommendation based on trust relationship among users. Data Analysis and Knowledge Discovery, 2017, (7): 90–99.

7 Du YP, Du XY, Huang L. Improve the collaborative filtering recommender system performance by trust network construction. Chinese Journal of Electronics, 2016, 25(3): 418–423. [doi: 10.1049/cje.2016.05.005]

8 刘智捷. 基于信任关系和兴趣变化的协同过滤算法研究 [硕士学位论文]. 杭州: 杭州电子科技大学, 2017.

9 蒋宗礼, 于莉. 基于用户特征的协同过滤推荐算法. 计算机系统应用, 2019, 28(8): 190–196. [doi: 10.15888/j.cnki.csa.007002]

10 刘智捷, 徐小良, 王宇翔. 基于融合信任关系的协同过滤推荐算法. 自然科学版, 2018, 38(3): 44–48.