



计算机科学

COMPUTER SCIENCE

一种改进的融合相似度和信任度的协同过滤算法

蔡晓娟, 谭文安

引用本文

蔡晓娟, 谭文安. 一种改进的融合相似度和信任度的协同过滤算法[J]. 计算机科学, 2022, 49(6A): 238-241.

CAI Xiao-juan, TAN Wen-an. Improved Collaborative Filtering Algorithm Combining Similarity and Trust[J].

Computer Science, 2022, 49(6A): 238-241.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[基于点割集图分割的矩阵变换与分解的推荐算法](#)

Matrix Transformation and Factorization Based on Graph Partitioning by Vertex Separator for Recommendation

计算机科学, 2022, 49(6A): 272-279. <https://doi.org/10.11896/jsjcx.210600159>

[基于遗憾探索的竞争网络强化学习智能推荐方法研究](#)

Study on Intelligent Recommendation Method of Dueling Network Reinforcement Learning Based on Regret Exploration

计算机科学, 2022, 49(6): 149-157. <https://doi.org/10.11896/jsjcx.210600226>

[基于注意力机制和门控网络相结合的混合推荐系统](#)

Hybrid Recommender System Based on Attention Mechanisms and Gating Network

计算机科学, 2022, 49(6): 158-164. <https://doi.org/10.11896/jsjcx.210500013>

[融合用户偏好的图神经网络推荐模型](#)

Graph Neural Network Recommendation Model Integrating User Preferences

计算机科学, 2022, 49(6): 165-171. <https://doi.org/10.11896/jsjcx.210400276>

[结合物品相似性的社交信任推荐算法](#)

Social Trust Recommendation Algorithm Combining Item Similarity

计算机科学, 2022, 49(5): 144-151. <https://doi.org/10.11896/jsjcx.210300217>

一种改进的融合相似度和信任度的协同过滤算法

蔡晓娟¹ 谭文安^{1,2}

1 南京航空航天大学计算机科学与技术学院 南京 211106

2 上海第二工业大学计算机与信息工程学院 上海 201209

(ava_tsai@163.com)

摘要 电子商务的迅猛发展在给用户提供更多商务选择的同时也导致了信息的泛滥。推荐系统作为信息过滤技术中必不可少的一种方法获得了社会的普遍关注。协同过滤算法是推荐系统中应用最广泛的技术,但其面临数据稀疏性、冷启动、数据扩展性等问题。文中提出了一种改进的融合相似度和信任度的协同过滤算法,该算法包括3个步骤:首先,计算用户之间的信任度;其次,计算用户之间的相似度;最后,融合信任度和相似度以计算用户之间的信任值,从而得到最终的评分预测方程。实验结果表明,针对不同的邻居集,所提算法的性能均优于传统协同过滤算法。

关键词: 电子商务;推荐系统;协同过滤;相似性;信任度

中图法分类号 TP182

Improved Collaborative Filtering Algorithm Combining Similarity and Trust

CAI Xiao-juan¹ and TAN Wen-an^{1,2}

1 College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

2 School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai 201209, China

Abstract The rapid development of e-commerce not only gives consumers more choice but has also causes information overload. As an indispensable method in information filtering technology, recommendation system has been widely concerned by the society. Collaborative filtering algorithm is the most widely used technology in recommendation systems, but it faces problems such as data sparsity, cold start and data scalability. This paper proposes an improved collaborative filtering algorithm model based on the fusion of trusted values and user similarity. This algorithm comprises three steps: first, we calculate the trust values between users; then we calculate the similarity between users; at last, we integrate the trust and the similarity to re-calculate the trust value between users and get the final rating prediction equation. Experimental results show that for different neighborhood sets, the performance of the proposed algorithm is better than that of traditional collaborative filtering algorithms.

Keywords E-commerce, Recommendation system, Collaborative filtering, Similarity, Trust

1 引言

根据 CNNIC 发布的第 44 次《中国互联网络发展状况统计报告》,截至 2020 年 3 月,我国网购用户规模达到 9.04 亿人,同比增长 18.1%。网络消费已经成为社会消费的主流趋势,解决数据信息过载并提高推荐的准确率是个性化推荐系统的最终目的。目前推荐系统应用最广泛的算法是协同过滤算法,协同过滤算法的主要思想是利用与当前用户具有相似兴趣或历史行为的用户的评分来预测当前用户的评分,归结于一句话就是“物以类聚,人以群分”。其缺陷是只考虑了用户间的相似性,导致用户-项目评分矩阵非常稀疏,易造成推荐准确率不高^[1]。

针对用户-项目评分矩阵的稀疏性、推荐结果准确率不高的问题,本文提出了一种改进的融合信任度和相似度的协同过滤算法,实验结果表明,所提算法运行结果的 MAE 和

RMSE 明显优于传统协同过滤算法。本文的贡献在于:1)考虑了用户之间的信任度,并融合了相似度与信任度寻找最近邻居集;2)基于信任值的权重来预测用户的评分,提高了系统推荐的准确率。

2 相关工作

许多改进的协同算法都致力于建立推荐系统,这些算法可以分为基于用户的方法和基于项目的方法。基于项目的方法^[2]首先分析用户-项目矩阵以确定不同项之间的关系,然后使用这些关系间接计算用户的建议。基于用户的协同过滤^[3-4]属于第一代协同过滤,其基本思想是针对用户之间的相似度提出建议。但是这些算法仍然存在数据稀疏、冷启动、推荐准确率不高等问题。为此,很多学者开始研究用户的行为,不少学者将用户之间的信任度引入到协同过滤算法中。Jiang 等^[5]提出了一种基于改进的融合用户信任数据和用户

基金项目:国家自然科学基金(61672022,61272036,U1904186);上海第二工业大学校重点学科资助项目(XXKZD1604)

This work was supported by the National Natural Science Foundation of China(61672022,61272036,U1904186) and Foundation of the Key Disciplines of Shanghai Polytechnic University(XXKZD1604).

通信作者:谭文安(watan@sspu.edu.cn)

相似性的 slope one 算法,该算法利用 Amazon 数据集中的 helpfulness 属性作为信任率来确定信任数据,解决了传统 slope one 算法低准确率和不信任数据的问题。Duricic 等^[6]为了解决冷启动用户问题,提出使用网络科学中的一种度量方法,即正则等价,将其应用到信任网络中生成一个相似度矩阵,用于选择 k 个最近邻进行推荐。Alejandro 等^[7]研究了用户相似度对获得高质量项目推荐的重要程度,并提出根据用户偏好与目标用户偏好的重叠程度来选择邻居进行推荐。另外,许多使用人工智能(AI)方法的个性化推荐系统已经被开发出来,一些著名的电子商务网站,如 Amazon,采用推荐技术向客户推荐产品,从而提高了其服务的质量和效率^[8-9]。Zarzour 等^[10]提出了一种基于降维和聚类技术的算法,提出并评价了一个有效的两阶段推荐系统,该系统能够产生准确、高效的推荐。Liu 等^[11]引入时间衰减函数对用户评分进行预处理,利用项目属性向量对项目进行特征化,将用户兴趣向量用于用户,并利用聚类算法对用户和用户进行聚类各项目,然后利用改进的相似性度量方法寻找用户最近邻,并在聚类中提出推荐候选集来提高推荐准确率。

针对上述研究中用户-项目评分矩阵的稀疏性,本文提出了一种改进的融合用户信任度和相似度的协同过滤算法。首先计算用户间的信任度;其次融合信任度和相似度再计算用户间的信任值;最后根据计算的信任值,获得最近邻居集,并计算当前用户未评分项目的评分。

3 基于用户的最近邻推荐

基于用户的最近邻推荐(UserCF)的算法思想是将用户的评分数据和用户 ID 作为输入,分析当前用户的行为数据,找到与当前用户兴趣相似的近邻用户,然后根据最近邻居集中用户的评分对当前用户没有评分过的项目进行评分。算法的实现过程一般分为 3 步。

(1) 建立用户评分矩阵

将用户对项目的所有评分根据一定的规则转换成数值评分,形成用户-项目评分矩阵。

(2) 计算用户之间的相似度,得到用户之间的相似度矩阵,并构造最近邻居集。

计算用户的相似度的方法主要有 3 种:余弦相似性, Pearson 相关系数以及修正的余弦相似性^[12]。

余弦相似性的计算方法如式(1)所示:

$$S_{u,v} = \frac{\sum_{i \in II_{uv}} r_{ui} * r_{vi}}{\sqrt{\sum_{i \in II_{uv}} r_{ui}^2} * \sqrt{\sum_{i \in II_{uv}} r_{vi}^2}} \quad (1)$$

Pearson 相关系数的计算方法如式(2)所示:

$$S_{u,v} = \frac{\sum_{i \in II_{uv}} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in II_{uv}} (r_{ui} - \bar{r}_u)^2} * \sqrt{\sum_{i \in II_{uv}} (r_{vi} - \bar{r}_v)^2}} \quad (2)$$

其中, II_{uv} 是用户 u 和用户 v 共同评分的项目集合; r_{ui} 是用户 u 对项目 i 的评分; r_{vi} 是用户 v 对项目 i 的评分; \bar{r}_u 和 \bar{r}_v 分别表示用户 u 和 v 对项目的平均分。

假设用户数量为 n , 用户相似矩阵可表示为:

$$sim = \begin{bmatrix} 1 & s_{1,2} & \cdots & s_{1,n} \\ s_{2,1} & 1 & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n,1} & s_{n,2} & \cdots & 1 \end{bmatrix} \quad (3)$$

通过选择与当前用户 u 相似度最高的前 k 个用户并将其作为邻居,来构造最近邻居集。

(3) 评分预测

根据当前用户最近邻居集中用户的项目和其对应的评分对当前用户未作评分的项目进行预测,其预测评分计算式如式(4)所示:

$$pred(u,i) = \bar{r}_u + \frac{\sum_{v \in N(i) \cap sim_{u,k}} sim(u,v)(r_{vi} - \bar{r}_v)}{\sum_{v \in N(i) \cap sim_{u,k}} |sim(u,v)|} \quad (4)$$

其中, \bar{r}_u 和 \bar{r}_v 分别表示用户 u 和用户 v 对项目的平均分; $N(i)$ 表示评分过的项目 i 的用户集; $sim_{u,k}$ 表示与用户 u 相似度最高的前 k 个用户的集合; r_{vi} 表示用户 v 对项目 i 的评分。

4 改进的基于用户的最近邻推荐

4.1 信任推荐模型的整体框架

信任具有主观性、可度量性、弱传递性、时间衰减性等性质,根据角度、场景以及社会经验的不同,信任的定义也会不同。Manuel 等^[13]将信任定义为商务环境中最重要的角色。传统概念中的信任关系是,如果 u 信任 v ,那么 v 同样信任 u ,并且 u 和 v 之间的信任值是一样的,即用户的信任度矩阵是对称的。但是在真实场景中不是这样的,如在医患关系中,患者会因为一个医生经验丰富而对该医生非常信任,但是医生可能信任患者会配合其治疗,两者的信任度是不一样的,即在现实生活中用户之间的信任度矩阵是不对称的。本文通过分析用户的评分项目数量和评分差异来计算用户之间的信任度。

本文提出的改进的融合用户信任度和相似度的协同过滤算法(UserTrustCF)推荐模型如图 1 所示。在该推荐模型中,系统将用户-项目评分矩阵作为输入,将当前用户未知项目的预测评分作为输出。该算法主要包括以下 5 个步骤:信任度计算、相似度计算、信任值计算、寻找最近邻居集、预测评分。

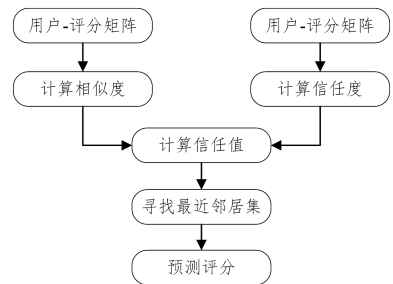


图 1 信任推荐模型的整体框架

Fig.1 Overall framework of trust recommendation model

4.2 信任度指标

对于用户 u 和用户 v ,本文表达式用用户 u 和用户 v 公共评分的项目的数量占用户 u 所有的评分项目的数量的比例与用户 u 和用户 v 共同评分的项目的数量占用户 u 和用户 v 的评分项目的总数量的比例的乘积来说明用户 u 和用户 v 之间的信任度,表达式如式(5)所示:

$$I_{u,v} = \frac{|I_u \cap I_v|}{|I_u|} * \frac{|I_u \cap I_v|}{|I_u \cup I_v|} \quad (5)$$

其中, I_u 和 I_v 表示用户 u 和用户 v 评分过的项目集。信任度量不仅考虑了两个用户共同评分项目在用户 u 中的比例,还考虑了共同评分项目在两个用户评分项目中的比例。

4.3 相似度指标

在用户-项目的评分矩阵中,我们将用户的所有项目评分定义为向量,这样每个用户都可以表示为一个 m 维的评分向量。我们定义用户集、项目集,然后根据用户-项目评分矩阵计算用户相似度。

因为余弦相似性没有考虑用户的平均评分的差异,所以相似度计算的可靠性并不稳固,也就是说余弦相似性无法计算每个维度上的差异。举例说明余弦相似性的不足,用户 u 和用户 v 的评分-项目的情况如表 1 所列。

表 1 用户-项目评分表

Table 1 Rating of user-item

	I_1	I_2	I_3	I_4
u	5	5	-	3
v	1	1	-	1

根据式(1)计算 u 对 v 的相似度:

$$s_{u,v} = \frac{5 * 1 + 5 * 1 + 3 * 1}{\sqrt{(25 + 25 + 9) * (1 + 1 + 1)}} = 0.98$$

用户 u 和用户 v 根据余弦相似性计算的结果为 0.98,这意味着用户 u 和用户 v 的兴趣非常相似。然而,从评分的角度来看,用户 u 对项目 1、项目 2、项目 4 非常喜欢,但是用户 v 并不喜欢这 3 个项目。因此,余弦相似性对评分维度并不敏感导致相似度计算不可靠。

为了避免这种错误,用 Pearson 相关系数用户评分与其平均评分的偏差来代替原来的评分,因此本文采用 Pearson 相关系数计算用户之间的相似度。

4.4 计算信任值

基于相似度和信任度作为权重,我们可以得到信任值 $Trust_{u,v}$,信任值的计算式如式(6)所示:

$$Trust_{u,v} = sim_{u,v} * \tau_{u,v} \quad (6)$$

其中, $Trust_{u,v}$ 表示用户 u 和用户 v 之间的信任值; $sim_{u,v}$ 表示用户 u 和用户 v 之间的相似度; $\tau_{u,v}$ 表示用户 u 和用户 v 之间的信任度。

选取当前用户最信任的 k 个用户作为当前用户的邻居,构造最近邻居集 $Trust_{u,k}$ 。

4.5 预测评分

基于当前用户与最近邻居集中用户的信任值 $Trust(u, v)$ 的权重计算当前用户 u 的未知项目 i 的预测评分 $pred(u, i)$, $pred(u, i)$ 的计算式如式(7)所示:

$$pred(u, i) = \bar{r}_u + \frac{\sum_{v \in N(i) \cap Trust_{u,k}} Trust(u, v) (r_{vi} - \bar{r}_v)}{\sum_{v \in N(i) \cap Trust_{u,k}} |Trust(u, v)|} \quad (7)$$

其中, \bar{r}_u 和 \bar{r}_v 分别表示用户 u 和用户 v 对项目的平均分; $N(i)$ 表示评分过项目 i 的用户集; $Trust_{u,k}$ 表示用户 u 最信任的 k 个用户的用户集; r_{vi} 表示用户 v 对项目 i 的评分; $Trust(u, v)$ 表示用户 u 和用户 v 的信任值。

5 实验结果与分析

5.1 实验数据

实验使用 Movielens 项目的部分用户评分数据集¹⁾, 采取离线实验进行 10 次实验来比较算法的预测准确率。按照一定

的格式生成一个标准的用户-项目评分数据集,并将其分为训练集和测试集,训练集和测试集的比例为 4:1。每个项目都有对应的评分,用稀疏度 S 反映数据集的稀疏程度, S 越高,数据系数程度越高,数据也就越稀少。 S 的定义如式(8)所示:

$$S = 1 - \frac{|R|}{|U| * |I|} \quad (8)$$

其中, $|R|$ 表示数据集中的评论数量; $|U|$ 表示数据集中的用户数量; $|I|$ 表示数据集中的项目数量。表 2 列出了该数据集的基本信息。

从表 2 可以看出,该数据集的数据稀疏度达到了 98.3%,如果仅靠单一的协同过滤方法,推荐系统的推荐质量将很难保证。本文算法提出用户信任度来提高推荐质量,引入用户的信任关系权重对当前用户的未知项目进行预测,具体方法在第 4 节中已经给出。

表 2 movielens 数据集的信息

Table 2 Information of movielens

数据集	用户数	项目数	评论数	稀疏度
movielens	10036	5986	1000952	0.983

5.2 评价指标

评价推荐算法的最基本指标是服务推荐的准确率。准确率指标有很多种,如衡量项目的预测评分与实际评分的接近程度、衡量预测评分与实际评分的相关性、考虑具体评分、只考虑推荐排名,本文主要考虑预测的准确率。

评价指标分为两大类:预测准确度指标和分类准确度指标。我们采用预测准确度指标,即通过比较预测评分与实际用户评分的偏差来预估算法的准确性。最经典的预测准确度指标是平均绝对误差(MAE)^[14], MAE 主要计算测试数据集中预测评分与实际评分之间的平均绝对误差。MAE 越小,预测就越准确,从而能可以制定更好的建议。MAE 的定义如式(9)所示:

$$MAE = \frac{\sum_{i=1}^N |p_i - r_i|}{N} \quad (9)$$

其中,实际评分集为 $\{r_1, r_2, r_3, \dots, r_N\}$; 预测评分集为 $\{p_1, p_2, p_3, \dots, p_N\}$ 。

比 MAE 更严格的是均方根误差(RMSE)^[14],它增加了对不准确预测评分的惩罚(平方惩罚),因此对系统的评价要求更高。RMSE 越小,预测就越准确,从而能构成更好的推荐列表。RMSE 的定义如式(10)所示:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (p_i - r_i)^2}{N}} \quad (10)$$

5.3 实验结果与结果分析

本文通过对比 MAE 和 RMSE 值来分析所提出的融合用户信任度和相似度的协同过滤算法(UserTrustCF)及传统的基于用户的最近邻推荐算法(UserCF)。

为了验证算法的有效性,本文为用户选取不同大小的邻域集,用于验证在不同邻域集上 UserTrustCF 和 UserCF 的预测准确度。图 2 给出了 UserTrustCF 和 UserCF 在不同邻域集情况下的 MAE 的实验结果。

¹⁾ <https://grouplens.org/datasets/movielens/>

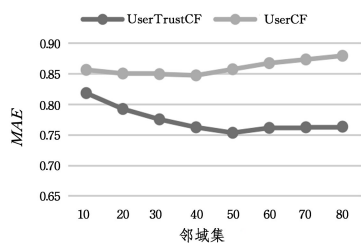


图2 UserTrustCF 和 UserCF 在不同邻域下的 MAE 值

Fig. 2 MAE values of UserTrustCF and UserCF in different neighborhood sets

图3给出了 UserTrustCF 和 UserCF 在不同邻域集情况下的 RMSE 的实验结果。

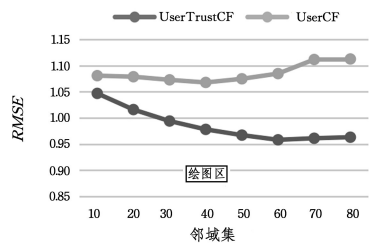


图3 UserTrustCF 和 UserCF 在不同邻域的 RMSE 值

Fig. 3 RMSE values of UserTrustCF and UserCF in different neighborhood sets

由图2和图3可知,随着邻域集的扩大,UserTrustCF在MAE和RMSE上的结果均优于UserCF,说明本文提出的UserTrustCF比UserCF有更好的预测准确度。另外,文献[50-60]达到最优值。UserCF在文献[40-50]达到最优值后准确度开始降低且误差的改进明显减少,两种算法的差距越来越大。这种现象是因为数据集的稀疏性,随着邻域集的扩大,邻域集中的用户和当前用户的相似度越来越低,导致预测评分不准确,而本文算法考虑了用户间的交互行为,利用用户间的信任关系较好地解决了上述问题。

结束语 针对传统的基于用户的最近邻推荐算法存在推荐准确率和推荐系统中评分可信性问题,提出了一种改进的融合用户信任度和相似度的协同过滤推荐算法。首先分析用户之间的交互行为和评分行为,构建用户之间的信任度,接着融合用户的相似度和信任度,计算用户之间的信任值权重,最后基于信任值的权重预测用户未知项目的评分。本文算法具有灵活性强、应用性广的特点,例如社交网络推荐系统、贷款服务平台等。实验结果表明,将信任度引入到传统的基于用户的最近邻推荐算法中是可行的,其推荐准确率得到了显著的提高。

参 考 文 献

- [1] WANG P, QIAN Q, SHANG Z, et al. An recommendation algorithm based on weighted slope one algorithm and user-based collaborative filtering [C] // 2016 Chinese Control and Decision Conference (CCDC). IEEE, 2016: 2431-2434.
- [2] LI D, CHEN C, LV Q, et al. An algorithm for efficient privacy-preserving item-based collaborative filtering [J]. Future Generation Computer Systems, 2016, 55: 311-320.
- [3] KOOHI H, KIANI K. User based Collaborative Filtering using fuzzy C-means [J]. Measurement, 2016, 91: 134-139.

- [4] SHI Y, LARSON M, HANJALIC A. Exploiting user similarity based on rated-item pools for improved user-based collaborative filtering [C] // Proceedings of the Third ACM Conference on Recommender Systems. 2009: 125-132.
- [5] JIANG L, CHENG Y, YANG L, et al. A trust-based collaborative filtering algorithm for E-commerce recommendation system [J]. Journal of Ambient Intelligence and Humanized Computing, 2019, 10(8): 3023-3034.
- [6] DURICIC T, LACIC E, KOWALD D, et al. Trust-based collaborative filtering: Tackling the cold start problem using regular equivalence [C] // Proceedings of the 12th ACM Conference on Recommender Systems. 2018: 446-450.
- [7] BELLOGÍN A, CASTELLS P, CANTADOR I. Improving memory-based collaborative filtering by neighbour selection based on user preference overlap [C] // Proceedings of the 10th Conference on Open Research Areas in Information Retrieval. 2013: 145-148.
- [8] THORAT P B, GOUDAR R M, BARVE S. Survey on collaborative filtering, content-based filtering and hybrid recommendation system [J]. International Journal of Computer Applications, 2015, 110(4): 31-36.
- [9] SMITH B, LINDEN G. Two decades of recommender systems at Amazon. com [J]. IEEE Internet Computing, 2017, 21(3): 12-18.
- [10] ZARZOUR H, AL-SHARIF Z, AL-AYYOUB M, et al. A new collaborative filtering recommendation algorithm based on dimensionality reduction and clustering techniques [C] // 2018 9th International Conference on Information and Communication Systems (ICICS). IEEE, 2018: 102-106.
- [11] LIU X J. An improved clustering-based collaborative filtering recommendation algorithm [J]. Cluster Computing, 2017, 20(2): 1281-1288.
- [12] XU J, LI X B, XU H. Research and analysis of collaborative recommendation algorithm based on user trust [J]. Data Communication, 2019, 2: 29-34.
- [13] MANUEL P. A trust model of cloud computing based on Quality of Service [J]. Annals of Operations Research, 2015, 233(OCT.): 281-292.
- [14] WANG W, LU Y. Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model [C] // IOP Conference Series: Materials Science and Engineering. 2018.



CAI Xiao-juan, born in 1997, postgraduate. Her main research interests include recommendation system and so on.



TAN Wen-an, born in 1965, Ph.D, professor. His main research interests include software service engineering, trustworthy service computing and composition, collaborative computing, business process intelligence technology.