doi: 10.3969/j.issn.0490-6756.2018.03.010

基于综合相似度迁移的协同过滤算法

金　玉[1], 崔兰兰[2], 孙界平[1], 琚生根[1], 王　霞[1]
(1.四川大学计算机学院,成都 610065; 2.中国人民解放军 78123 部队, 成都 610017)

摘　要: 数据稀疏性问题是传统协同过滤算法的主要瓶颈之一.迁移学习通常是利用目标领域与辅助领域的潜在关系,对辅助领域进行知识迁移,以此来提高目标领域的推荐质量.现有的基于相似度迁移模型,普遍只利用了用户评分信息,并且在评分相似度计算上忽略了用户评分标准个性差异.针对这些问题,提出了一种综合相似度迁移模型,在相似度计算上,即利用了用户评分信息同时也利用了用户属性信息,并且考虑了用户间对满意度的打分标准的差异性,采用了用户评分分布一致性来衡量用户评分相似度的方法,提高了相似度计算的准确性,从而提高了数据迁移的质量.实验结果表明,该模型较其他算法能比较有效地缓解数据稀疏性问题.

关键词: 数据稀疏; 协同过滤; 迁移学习; 相似度迁移
中图分类号: TP391　　**文献标识码**: A　　**文章编号**: 0490-6756(2018)03-0477-06

Collaborative filtering algorithm based on integrated similarity transfer

JIN Yu[1], *CUI Lan-Lan*[2], *SUN Jie-Ping*[1], *JU Sheng-Gen*[1], *WANG Xia*[1]
(1. College of Computer Science, Sichuan University, Chengdu 610065, China; 2. 78123 Troop of the PLA, 610017, China)

Abstract: Data sparsity is one of the most challenges for traditional collaborative filtering algorithms. Transfer learning methods used the potential relationship between the target domain and the auxiliary domain to transfer the auxiliary domain knowledge, so as to improve the recommendation accuracy of the target domain. The existing transfer model based on similarity generally used only the rating information, and ignores the difference of user rating. To solve these problems, a transfer model based on comprehensive similarity is proposed, used user rating information and user attribute information, taking account of the difference of user rating, used the consistency of ratings, distribution to measure user rating similarity, improved the accuracy of similarity computation, thus improved the quality of data migration. Experimental results showed that the proposed model can effectively alleviate the sparsity of data compared with other algorithms.

Keywords: Sparse data; Collaborative filtering; Transfer learning; Similarity transfer

1　引　言

当前,网络信息量呈指数级增长,网络用户一方面可获取丰富信息,另一方面却面临信息过载问题[1],难以从海量信息中挖掘对自己有用的信息.推荐系统可根据用户兴趣,从海量数据筛选出用户感兴趣的部分.目前,推荐系统已得到广泛应用,如 Amazon, eBay, MovieLens, GroupLens 等电子商

收稿日期: 2017-07-19
基金项目: 国家自然科学基金(61332006)
作者简介: 金玉(1992－), 女, 硕士生, 研究方向为数据科学.
通讯作者: 琚生根. E-mail: jsg@scu.edu.cn

务平台.

协同过滤技术[2-4]是推荐系统中应用最广泛的技术之一,其基本思想是:利用用户的历史评分数据,来预测用户对未评分物品的兴趣度,选择兴趣度最高的几项物品作为推荐结果.对于传统的协同过滤算法,其最关键步骤是计算用户间或物品间的相似度,但随着数据的增长,用户评分数据会极度稀疏,而推荐质量也会随之下降.

目前,针对数据稀疏问题[5],有以下几种解决方案.

(1) 通过填充未评分物品来降低数据集的稀疏性[6-8],该算法适用于物品更新不频繁且物品数远小于用户数的场景,依赖于用户行为同时存在冷启动问题;

(2) 通过矩阵分解来降低数据集的稀疏性[9],该算法利用用户与项目之间的潜在关系,对评分矩阵进行奇异值分解,该方法训练代价大,且不适应用户兴趣的改变;

(3) 利用迁移学习思想[10,11],通过领域间的交叉部分来促进目标领域的学习[12],该算法通过发现目标领域与辅助领域的潜在关系,达到辅助目标领域训练的目的,由此,其依赖于潜在关系的可靠程度,如不可靠会导致负迁移.

目前,一些学者提出了利用多领域数据来缓解目标领域数据稀疏问题.如 Jamali[13,14]等人提出了一种基于上下文的矩阵分解模 HeteroMF,其主要思想是利用多领域间共同实体,并共享实体的特征因子来同时对多个矩阵进行联合分解,其算法需要训练较多参数且需消耗大量时间计算梯度;Li Bin[15]等人提出了一种评分矩阵生成模型 RMGM,其主要思想是通过找到共享的隐式集群级别的评级矩阵,然后利用这个矩阵填充目标领域中原始矩阵的空值,该方法使用与强相关领域且并没有理论支持;李超[16]等人提出的一种基于用户相似度迁移模型 TSUCF,其主要思想是交叉领域数据建立起辅助领域与目标领域的联系,达到辅助目标领域的目的,该方法仅利用用户评分信息,而且在衡量评分相似度时,仅采用共同物品数目来衡量,没有考虑用户的偏好.

虽然以上算法均采用辅助领域知识来提高推荐精度,但仍有以下不足:

(1) 基于矩阵变换模型的训练参数较多;

(2) 要求辅助领域与目标领域满足强相关,模型适用场景少;

(3) 计算用户评分相似度时,忽略了用户间对满意度的打分标准的差异性.

针对以上问题,本文提出了一个基于综合相似度迁移的推荐算法.本文算法同时利用用户属性信息和用户评分信息来计算相似度,并且在计算用户评分相似度时,考虑了用户对满意度打分标准的差异性,采用了用户评分分布一致性来衡量用户评分相似度,忽略了评分的具体数值.

2 相关研究

2.1 基于综合相似度迁移的推荐算法

本文提出了一种基于综合相似度迁移的推荐算法,利用辅助领域信息来缓解目标领域数据稀疏性问题.

下面将以两个电影平台为例对本文算法进行说明.假设有两个平台 e_1 和 e_2,U_1 表示只在平台 e_1 中存在历史行为信息的用户,U_2 表示只在平台 e_2 中存在历史行为信息的用户,U_c 表示在平台 e_1 和 e_2 中均有过历史行为信息的用户,定义为交叉用户.用户行为矩阵如图 1 所示.

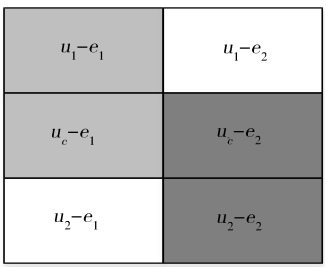

图 1 用户评分矩阵

Fig. 1 User rating matrix

在实际情况中,交叉用户的数量远远小于非交叉用户的数量.

传统推荐算法是利用所占比例较少的交叉用户对所占比例较大的非交叉用户进行推荐,由此会出现冷启动问题和数据稀疏问题,使得推荐质量较低.

本文算法是通过交叉用户,为非交叉用户 U_1 和 U_2 建立起相似度联系,以此帮助目标领域进行推荐.

2.2 相似度迁移

如图 1 所示,非交叉用户 U_1 和用户 U_2 无法直接计算相似性,但是,用户 U_1 和用户 U_2 分别与交叉用户 U_c 的相似度是可以计算的,所以,可将交叉用户 U_c 作为纽带来建立用户 U_1 和用户 U_2 的相似度.

相似度迁移步骤：首先找出与平台 1 和平台 2 的公共用户集 U_c；然后分别计算 U_1 与 U_c 的相似性，记为向量$\overrightarrow{S_{1c}}$，U_2 与 U_c 的相似性，记为$\overrightarrow{S_{c2}}$；最后计算$\overrightarrow{S_{1c}}$与$\overrightarrow{S_{c2}}$的内积，即为 U_1 和 U_2 的传递相似度. 相似度迁移如图 2 所示.

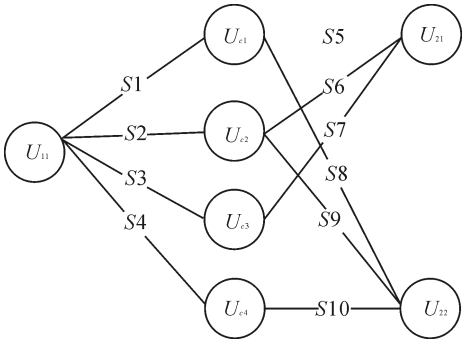

图 2　相似性迁移

Fig. 2　Similarity transfer

其中，U_{11} 表示平台 1 中的非交叉用户 1；U_{21}、U_{22} 表示平台 2 中的非交叉用户 1；U_{c1}、U_{c2} 等表示交叉用户；S_1、S_2 等表示相似度. 如果要计算 U_{11} 与 U_{21} 之间的相似度，则可通过 U_{c1}，U_{c2}，U_{c3} 过渡，间接计算 $S_{U_{11}U_{21}}=S_1.S_5+S_2.S_6+S_3.S_7$. 综上，则 U_1 和 U_2 之间的相似度计算可形式化为

$$S_{U_1U_2}=S_{U_1U_c}\cdot S_{U_cU_2} \tag{1}$$

2.3　相似度计算

基于以上分析，计算非交叉用户 U_1 与 U_2 的相似度之前，需先计算非交叉用户 U_1，U_2 分别与交叉用户的相似度，相似度计算如下所示.

1）用户评分相似度：本文通过评分分布一致性、可信度两方面衡量用户评分相似度.

一致性是由两用户评价过的相同物品的评分分布决定. 评分分布越一致，说明两用户的兴趣越相似. 设$\{ur_1,ur_2,\cdots,ur_n\}$，$\{ur_1,ur_2,\cdots,ur_n\}$分别为用户 u 与用户 v 对共同物品的评分集，将两组数据分别进行递增排序，即$\{ur_1,ur_2,\cdots,ur_n\}$，$\{ur_1,ur_2,\cdots,ur_n\}$，如果 $1,2,\cdots,n$ 与 $x_1,x_2,\cdots,x_n$ 的匹配度越大，则表明两者的一致性越高. 计算公式如下.

$$dist(u,v)=\frac{\sum_{i=1}^{n}(x_i-i)^2}{\sum_{i=1}^{n}[(n-i)-i]^2} \tag{2}$$

可信度是根据两用户评价过的相同物品的数量决定的，若数量很小，即使评分分布一致，也不代表两者一定相似. 计算公式如下.

$$conf(u,v)=\frac{|I_u\cap I_v|}{|I_u\cup I_v|} \tag{3}$$

其中，I_u 表示用户 u 评价的物品集.

用户评分相似度计算公式如下.

$$sim_1(u,v)=dist(u,v)conf(u,v) \tag{4}$$

2）用户属性相似度：本文通过用户属性衡量用户属性相似度. 一般认为，拥有相同属性的用户在一定程度上具有相似的兴趣. 计算公式如下.

$$sim_2(u,v)=\frac{1}{n}\sum_{i=1}^{n}d_isim(u,v,i) \tag{5}$$

其中，n 表示属性个数；$sim(u,v,i)$ 表示在第 i 个属性上两用户是否相同，如相同，则为 1，反之为 0，d_i 表示第 i 个属性的区分度，如果具有某属性的用户对所有物品都进行了评分则表明该属性没有区分度，其值由不同数据集决定.

3）最终相似度：一般情况下，当用户对某物品评分之后，应该尽量利用用户对物品评分信息，当用户对某物品没有评分，则应尽量利用用户属性信息. 当用户所评分的物品数量增多时，算法应平滑过渡到使用评分信息进行推荐，本文使用 sigmoid 函数进行平滑处理，最终用户相似度计算公式如下.

$$sim(u,v)=\alpha sim_1(u,v)+(1-\alpha)sim_2(u,v) \tag{6}$$

$$\alpha=2\times\left(1-\frac{1}{1+\exp(-|C_{uv}|)}\right) \tag{7}$$

其中，C_{uv} 表示用户 u 和用户 v 共同评价的物品集合. 由上述公式表示，用户相似度计算会随着用户所评价物品数量的增多，平滑过渡到使用评分信息，这种平滑过渡可以提高在冷启动状态下预测准确率.

2.4　算法描述

2.4.1　*计算用户相似度算法*　（1）根据用户属性信息，计算用户属性相似度；（2）根据用户评分信息，计算用户评分相似度；（3）根据用户属性相似度与用户评分相似度，计算最终用户相似度.

2.4.2　*基于迁移学习的推荐算法*　（1）计算 U_1 与 U_c 之间的相似度 $S_{U_1U_c}$；（2）计算 U_2 与 U_c 之间的相似度 $S_{U_cU_2}$；（3）计算迁移相似度 $S_{U_1U_2}$；（4）利用迁移相似度 $S_{U_1U_2}$，结合 UCF 算法进行推荐.

3　实　验

3.1　实验数据

实验采用 MovieLen 电影网站的数据集（http://grouplens.org/datasets/movielens/）. 数据集描述如下所示.

表 1 Movielens 数据描述

Tab. 1 Movielens data description

用户数	物品数	评分数	稀疏度
943	1682	100000	93.7%

实验数据集划分如下所示.

表 2 数据集划分

Tab. 2 Dataset partition

划分方案	A 组	B 组	C 组	D 组	E 组
目标用户	212	188	165	141	118
辅助用户	636	566	495	425	354
交叉用户	95	189	283	377	471

3.2 评价指标

为了衡量算法的预测准确度,本实验采用均方根误差 RMSE(Root Mean Squared Error,RMSE)来验证本文算法所得预测结果与用户真实评分的差距. RMSE 计算公式如下.

$$RMSE = \sqrt{\frac{\sum_{u,i\in T}(r_{ui} - pre_{ui})^2}{|T|}} \tag{8}$$

其中,r_{ui} 表示用户 u 对物品 i 的真实评分;pre_{ui} 表示用户 u 对物品 i 的预测评分;T 为测试集,$|T|$ 表示测试集大小. RMSE 越小,说明预测值与实际值越近,预测结果的准确率越高.

3.3 对比算法

(1) UCF 算法:只能利用交叉用户进行推荐.

(2) TSUCF 算法:利用所占比例较少的交叉用户的评分信息作为纽带,将两个不同电商的用户建立联系,达到推荐的效果.

(3) 本文算法:本文算法在 TSUCF 算法上做出改进:1) 充分利用了用户属性信息;2) 考虑了用户的评分标准的差异性,采用共同物品的评分分布一致性来衡量用户评分相似性.

3.4 实验结果

考虑到最近邻居数 N 的大小会对结果有影响,实验分别在最近邻居数为 5,10,20,30,40 前提下进行算法对比.

A 组数据集在不同最近邻居数下,算法的 RMSE 值,如图 3 所示.

由图 3 可得,本文算法的 RMSE 值均小于其他算法.

B 组数据集在不同最近邻居数下,算法的 RMSE 值,如图 4 所示.

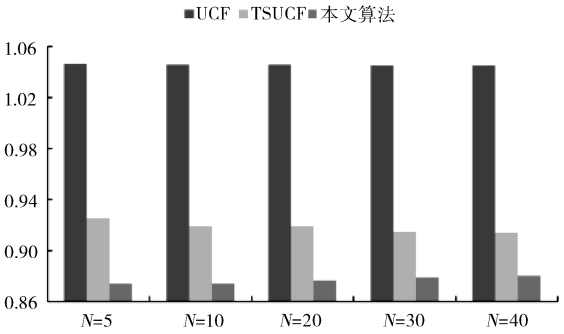

图 3 A 组下不同算法 RMSE 值对比图

Fig. 3 The RMSE of different algorithms in group A

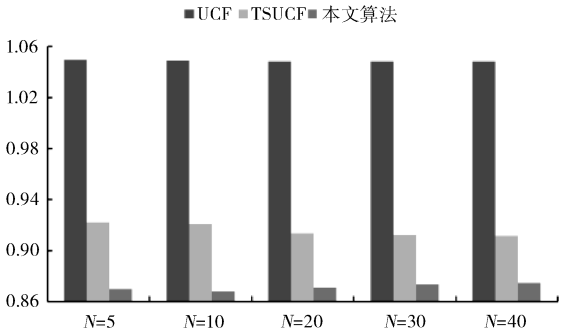

图 4 B 组不同算法 RMSE 值对比图

Fig. 4 The RMSE of different algorithms in group B

由图 4 可得,本文算法的 RMSE 值均小于其他算法. C 组数据集在不同最近邻居数下,算法的 RMSE 值,如图 5 所示.

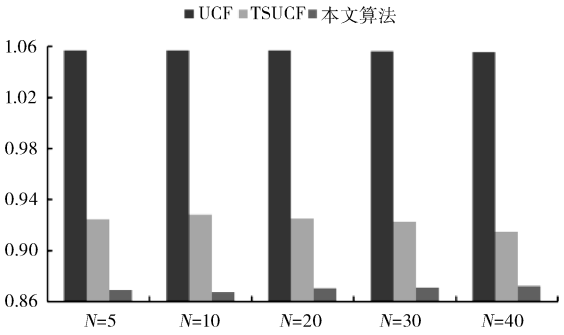

图 5 C 组不同算法 RMSE 值对比图

Fig. 5 The RMSE of different algorithms in group C

由图 5 可得,本文算法的 RMSE 值均小于其他算法.

D 组数据集在不同最近邻居数下,算法的 RMSE 值,如图 6 所示. 由图 6 可得,本文算法的 RMSE 值均小于其他算法.

E 组数据集在不同最近邻居数下,算法的 RMSE 值,如图 7 所示.

由图 7 可得,本文算法的 RMSE 值均小于其他算法.

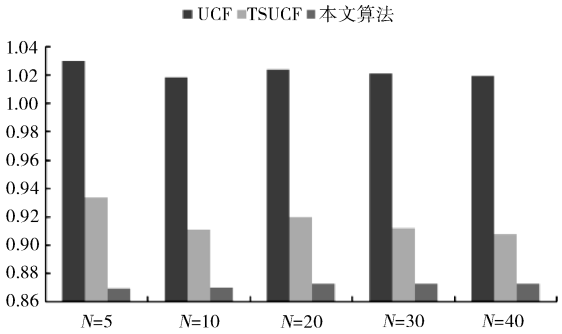

图 6　D 组不同算法 RMSE 值对比图

Fig. 6　The RMSE of different algorithms in group D

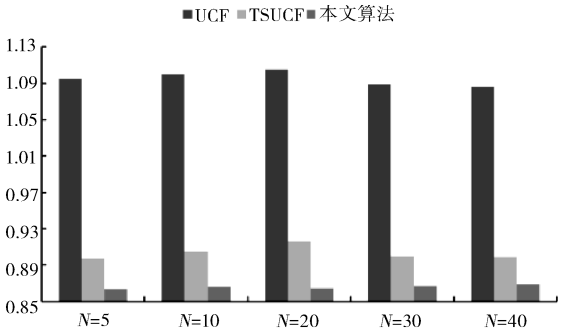

图 7　E 组不同算法 RMSE 值对比图

Fig. 7　The RMSE of different algorithms in group E

考虑到交叉用户数目对实验结果的影响，实验分别在交叉用户数目为 95、189、283、377、471 下进行算法对比.

在最近邻居数为 5 下，算法在不同数据组下的 RMSE 值，如图 8 所示.

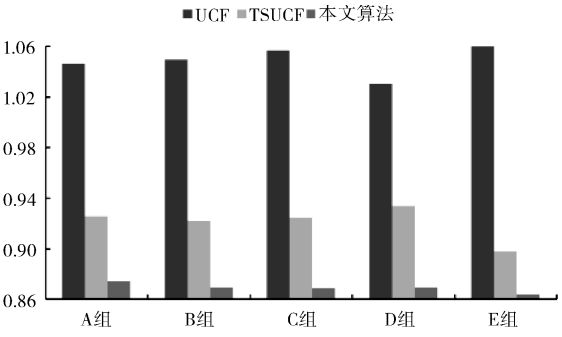

图 8　$N=5$ 下不同算法 RMSE 值对比图

Fig. 8　The RMSE of different algorithms

由图 8 可得，本文算法的 RMSE 值均小于其他算法.

在最近邻居数为 10 下，算法在不同数据组下的 RMSE 值，如图 9 所示.

由图 9 可得，本文算法的 RMSE 值均小于其他算法.

在最近邻居数为 20 下，算法在不同数据组下的 RMSE 值，如图 10 所示.

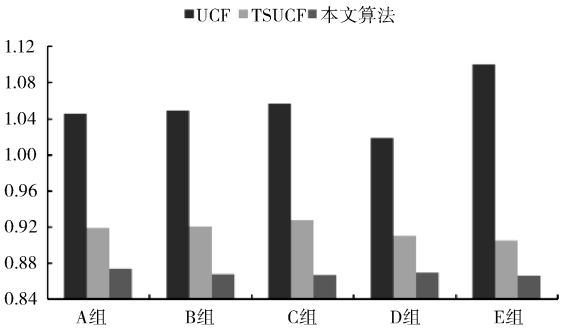

图 9　$N=10$ 不同算法 RMSE 值对比图

Fig. 9　The RMSE of different algorithms

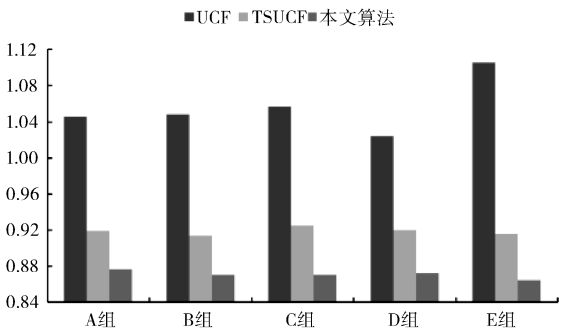

图 10　$N=20$ 不同算法 RMSE 值对比图

Fig. 10　The RMSE of different algorithms

由图 10 可得，本文算法的 RMSE 值均小于其他算法. 在最近邻居数为 30 下，算法在不同数据组下的 RMSE 值，如图 11 所示.

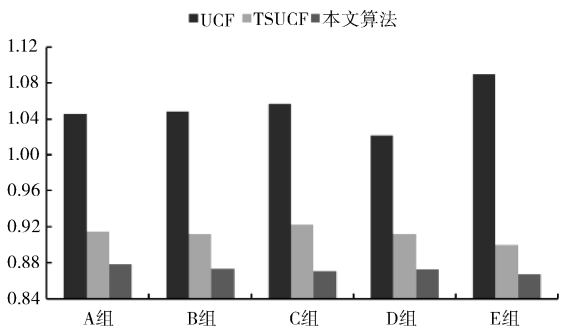

图 11　$N=30$ 不同算法 RMSE 值对比图

Fig. 11　The RMSE of different algorithms

由图 11 可得，本文算法的 RMSE 值均小于其他算法.

在最近邻居数为 40 下，算法在不同数据组下的 RMSE 值，如图 12 所示.

由图 12 可得，本文算法的 RMSE 值均小于其他算法.

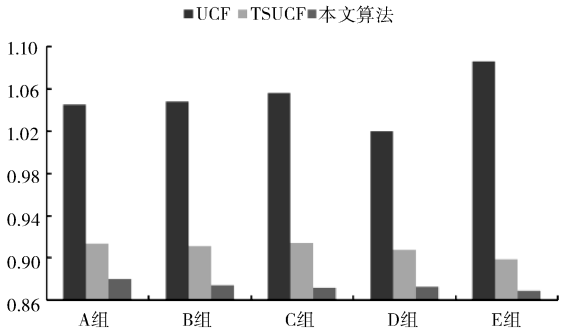

图 12 $N=40$ 不同算法 RMSE 值对比图
Fig. 12 The RMSE of different algorithms

4 结 论

本文算法利用用户属性相似度及用户评分相似度对辅助领域的数据进行迁移以解决目标领域的数据稀疏性问题,未来可以考虑结合项目相似度或其他知识,如文本信息等,对辅助领域的数据进行迁移,通过这种方式可以提高迁移数据的质量,从而提高推荐精确度.

参考文献:

[1] 王桂华, 陈黎, 于中华, 等. 一种建立在对客户端浏览历史进行 LDA 建模基础上的个性化查询推荐算法[J]. 四川大学学报:自然科学版, 2015, 52: 755.

[2] 刘汉清, 朱敏, 苏亚博, 等. 一种考虑用户兴趣转移特征的协同预测模型[J]. 四川大学学报:自然科学版, 2016, 53: 548.

[3] 张为民, 李坷露, 李永丽. 基于社交关系和条件补全的协同过滤推荐算法[J]. 吉林大学学报:理学版, 2017, 55: 1244.

[4] 王永, 万潇逸, 陶娅芝, 等. 基于 K-medoids 项目聚类的协同过滤推荐算法[J]. 重庆邮电大学学报:自然科学版, 2017, 29: 521.

[5] Pan W. A survey of transfer learning for collaborative recommendation with auxiliary data[J]. Neurocomput, 2016, 177: 447.

[6] Lemire D, Maclachlan A. Slope one predictors for online rating-based collaborative filtering[C]//Proceedings of the 2005 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics. Newport Beach, California: SIAM, 2005.

[7] Wang P, Ye H W. A personalized recommendation algorithm combining slope one scheme and user based collaborative filtering[C]// Proceedings of the 2009 International Conference on Industrial and Information Systems. New York, USA: IEEE, 2009.

[8] Sun Z, Luo N, Kuang W. One real-time personalized recommendation systems based on slope one algorithm [C]// Proceedings of the 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). Shanghai, China: IEEE, 2011.

[9] Sarwar B, Karypis G, Konstan J, *et al*. Analysis of recommendation algorithms for e-commerce[C]// Proceedings of the 2nd ACM Conference on Electronic Commerce. New York, USA: ACM, 2000.

[10] Pan W, Yang Q, Duan Y, *et al*. Transfer learning for semisupervised collaborative recommendation [J]. ACM Trans Interact Intel Syst, 2016, 6: 10.

[11] Pan W, Yang Q, Duan Y, *et al*. Transfer Learning for Semisupervised Collaborative Recommendation [J]. ACM Trans Interact Intel Syst, 2016, 6: 10.

[12] Pan W, Xiang E W, Liu N N, *et al*. Transfer learning in collaborative filtering for sparsity reduction[C]// Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence. USA: AAAI, 2010.

[13] Jamali M, Lakshmanan L. Heteromf: recommendation in heterogeneous information networks using context dependent factor models[C]//Proceedings of the 22nd international conference on World Wide Web. New York, USA: ACM, 2013.

[14] Wu S, Liu Q, Wang L, *et al*. Contextual operation for recommender systems[J]. IEEE Trans Knowl Data En, 2016, 28: 2000.

[15] Li B, Yang Q, Xue X. Transfer learning for collaborative filtering via a rating-matrix generative model[C]// Proceedings of the International Conference on Machine Learning. New York, USA: ACM, 2009.

[16] 李超, 周涛, 黄俊铭, 等. 基于用户相似性传递的跨平台交叉推荐算法[J]. 中文信息学报, 2016, 30: 90.