一种基于频繁项集挖掘的推荐算法

帖军,吕琴艳,孙翀,王江晴,尹帆
(中南民族大学 计算机科学学院,武汉 430074)

摘 要 协同过滤是推荐系统中应用最成功的技术之一,现有基于项目的协同过滤算法在计算项目相似度时过度依赖用户对项目的评分数据,没有考虑项目间内在的关联性,导致推荐质量不高.为了全面客观地评估项目相似度,提出了一种基于频繁项集挖掘的推荐算法(BFIM).该算法提出将频繁项集作用于相似度计算中,可以提高相似度计算的准确性,进而提升推荐算法的推荐质量.实验结果表明:提出的改进算法较对比算法在公开数据集上能取得更好的推荐效果.

关键词 协同过滤;频繁项集;相似度;推荐算法
中图分类号 TP391.3 **文献标识码** A **文章编号** 1672-4321(2019)01-0144-06
DOI 10.12130/znmdzk.20190125
引用格式 帖军,吕琴艳,孙翀,等. 一种基于频繁项集挖掘的推荐算法[J].中南民族大学学报(自然科学版),2019,38(1):144-149.
TIE Jun,LV Qinyan,SUN Chong,et al. A recommendation algorithm based on frequent itemsets mining[J].Journal of South-Central University for Nationalities(Natural Science Edition),2019,38(1):144-149.

A recommendation algorithm based on frequent itemsets mining

TIE Jun,LV Qinyan,SUN Chong,WANG Jiangqing,YIN Fan
(School of Computer Science, South-Central University for Nationalities, Wuhan 430074, China)

Abstract Collaborative filtering is one of the most successful technology in the recommendation system. The existing item-based collaborative filtering algorithm excessively depends on the users' rating data when calculating the items similarity, which does not consider the inner relevance between the items, and these lead to lower recommendation quality. In order to fully and objectively evaluate the items similarity, a recommendation algorithm based on Frequent Itemsets Mining (BFIM) is proposed. The algorithm proposes to apply frequent itemsets to the similarity calculation, which can increase the accuracy of similarity calculations and further improve the recommendation quality of the algorithm. Experimental results show that the proposed improved algorithm can achieve better recommendation results than the compared algorithms on public dataset.
Keywords collaborative filtering; frequent itemsets; similarity; recommendation algorithm

随着互联网和信息技术的迅猛发展,网络上的信息数据量呈指数增长,人们逐渐陷入“信息过载”时代.在这个时代,消费者很难从众多商品中找到自己感兴趣的商品,同时生产者也很难让自己的商品在众多用户的关注中脱颖而出.推荐系统[1-3]则成为解决该问题的重要手段.它可以根据用户的喜好筛去不相关的项目,并推荐用户可能喜欢的项目.

在推荐系统中,关联规则挖掘[4]和协同过滤[5,6]是最常用和最重要的两种技术.挖掘关联规则的目的是在大型事务数据库中搜索项目集合之间的

收稿日期 2018-06-21
作者简介 帖军(1976-),男,副教授,博士,研究方向:数据挖掘、物联网应用,E-mail:tiejun@mail.scuec.edu.cn
基金项目 国家科技支撑计划项目子课题(2015BAD29B01);农业部软科学研究课题(D201721);中央高校基本科研业务费专项资助项目(CZY18048)

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

内在联系.协同过滤可以分为基于用户的推荐(User-based Recommendation)、基于物品的推荐(Item-based Recommendation)、基于模型的推荐(Model-based Recommendation)等子类.基于用户的协同过滤算法,其基本思想是先找到和目标用户有相似兴趣的用户,然后把这些用户感兴趣而目标用户没有接触过的物品推荐给目标用户;基于项目的协同过滤推荐算法,是给用户推荐那些与他们之前喜欢的物品相似的物品,主要通过分析用户的行为记录计算物品之间的相似度;基于模型的协同过滤推荐则是先使用历史数据训练得到一个模型,再用此模型进行预测.

虽然协同过滤算法取得了巨大的成功,但始终存在数据稀疏性问题.电子商务网站中用户和项目的数目非常庞大,而多数用户只会对少量的项目进行评分,导致用户之间评分的重叠部分很小,难以计算两个用户之间的相似程度.而相似性计算是基于协同过滤的推荐系统中一个非常重要的步骤.因此数据稀疏性大大降低了协同过滤的预测准确性.

基于上述分析,本文将频繁项集挖掘与协同过滤相结合,提出一种新的相似性度量方法,提高相似度计算的准确率,从而提升协同过滤算法的推荐质量.

1　相关工作

数据稀疏性[7,8]和冷启动一直是影响推荐系统的推荐性能的重要因素,针对这些问题,一些研究人员提出将数据挖掘算法与协同过滤相结合,对用户-项目评分矩阵的缺失值进行预测并填充.缺失值填补法是根据已有的用户评分数据,以某种计算方法对用户未评分的数据进行估计并填充,可以显式地解决数据稀疏问题.

Shambour 等人[9]提出将来自用户社交信任网络的附加信息和项目的语义领域知识结合起来,以提高推荐的准确性和覆盖范围.Fan[10]等人提出一种基于预测值填充的 UBCF 推荐算法,该算法在计算用户相似度之前,通过整合基于内容的推荐算法和用户活动水平来预测用户项目矩阵中的缺失值,一定程度缓解了数据稀疏性对推荐精度的影响.Zhou 等人[11]提出了一种基于 SVD(奇异值分解,Singular Value Decomposition)推荐系统的增量式方法,每次迭代计算原始矩阵的奇异值分解,以解决稀疏性问题和用户兴趣的动态性.张玉芳等人[12]提出一种结合条件概率和传统协同过滤算法的非固定 K 近邻算法.该算法在分步填充评分矩阵思想的基础上,第一步只接受相似度和共同评分项目数量达到阈值的邻居用户作为目标用户邻居,然后计算并填充未评分项目;第二步使用第一步填充后的矩阵计算剩余未评分项目的评分.Chujai 等人[13]同时使用用户信息和电影信息挖掘频繁项集,填补缺失数据.Insuwan 等人[14]提出 SVDUPMedianCF 算法,该算法利用改进的 K-means 算法进行聚类,得到聚类的中心来填补缺失值.刘枚莲等人[15]提出基于双向关联规则项目评分预测的推荐算法,利用双向关联规则挖掘事务数据库中相互关联的项目,找到目标项目的关联集,利用已评分项目初步预测用户对目标项目的偏好程度,从而进行推荐.

缺失值填补法可以直观、显著地改善数据稀疏问题,但其本身是对评分缺失值的一种预测,并不能真正代表用户偏好,而预测的评分对推荐结果有较大的影响,可能导致最终的推荐结果不准确.当数据十分稀疏时,使用传统的相似度计算方法往往不能得到很好的推荐效果.因此,本文提出一种结合频繁项集与协同过滤的相似度改进算法,算法考虑项目之间相互关联的特性,对项目进行频繁项集挖掘,设计一种新的相似性计算方法,对传统相似性计算方法进行改进,从而寻找项目的最近邻居并进行推荐.

2　算法设计

2.1　问题描述

协同过滤的相似度[16,17]计算只考虑用户对项目的评分数据(评分数据具有很大的稀疏性),忽略了项目在空间上存在的相互关联性.而且,如果有恶意的评分,通过某种手段对某一特定的项目评分,并且评分很高,这无疑会带来噪声样本.这些问题都会导致推荐误差大,推荐结果不准确.如果引入事务数据库和频繁项,则可以认为频繁项是比较正常和可靠的事务数据,也可以认为频繁被用户购买的商品理论上具有较高的相似性.本文利用 Apriori 算法挖掘事务数据库中的频繁项集,根据频繁项集设计新的相似性度量函数,再与用户对项目的评分相似性进行加权综合,从而寻找项目的最近邻居并进行推荐.

协同过滤推荐算法是以用户-项目评分数据作为基础进行推荐的,假设 $U=\{u_1,u_2,\cdots,u_m\}$ 和 $I=\{i_1,i_2,\cdots,i_n\}$ 分别是用户和项目的集合,用户对项

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

目的评分矩阵为 $X \in R^{m\times n}$,即矩阵 X 有 m 行 n 列,第 i 行第 j 列的元素 $X_{i,j}$ 表示第 i 个用户对第 j 个项目的评分.

通过用户-项目评分数据 X 可以形成事务数据库矩阵 $D \in R^{m\times n}$,D 中的每一项 $D_{i,j}(i=1,2,\cdots,m;j=1,2,\cdots,n)$ 表示第 i 个用户是否对第 j 个项目进行评分.具体计算方式如公式(1):

$$D_{i,j}=\begin{cases}1, X_{i,j}\neq 0,\\ 0, X_{i,j}=0.\end{cases} \tag{1}$$

2.2 BFIM 算法

本文的 BFIM 算法针对基于项目的协同过滤算法,在计算项目相似度时完全依赖用户评分数据,忽略了项目在空间上的相互关联性问题,重新定义了项目相似度计算方法.算法提出对项目进行频繁项集挖掘,根据频繁项集设计一种相似性度量方法,融入到传统相似度计算方法中,然后利用该相似度更好地预测用户对项目的评分.

2.2.1 频繁项集矩阵的构建

本文采用 Apriori 算法进行频繁项集挖掘,根据最小支持度 $minSup$ 得出全部的频繁项集.现假设对事务数据库挖掘出的频繁项集总条数为 k,构建频繁项集矩阵 $F \in R^{k\times n}$,具体计算方法如下:

$$F_{k,i}=\begin{cases}1, \text{第 } k \text{ 条频繁项集包含项目 } i,\\ 0, \text{otherwise}.\end{cases} \tag{2}$$

矩阵 F 的可能取值如表1所示.

表1 频繁项集矩阵示例

Tab.1 An example of frequent itemsets matrix

频繁项数目	I_1	I_2	…	I_n
1	1	0	…	0
2	0	1	…	0
⋮	⋮	⋮	⋮	⋮
k	0	1	…	1

由表1可知,F 矩阵中$\{I_1,I_2,\cdots,I_n\}$ 表示 n 个项目,$\{1,2,\cdots,k\}$ 表示第 k 个频繁项集,$F_{2,2}=1$ 表示第2条频繁项集包含项目 I_2,$F_{2,1}=0$ 表示第2条频繁项集不包含项目 I_1.

2.2.2 相似度计算

频繁被购买的项目间存在一种相似性.BFIM 算法将基于频繁项集的项目相似性定义为 $S^{(1)}$,其中每一项 $S_{i,j}^{(1)}$ 表示基于频繁项集的项目 i 和项目 j 之间的相似度,计算公式如下:

$$S_{i,j}^{(1)}=\frac{\sum_{h=1}^{k}F_{h,i}\cdot F_{h,j}}{\sum_{h=1}^{k}F_{h,i}+\sum_{h=1}^{k}F_{h,j}-\sum_{h=1}^{k}F_{h,i}\cdot F_{h,j}}, \tag{3}$$

其中 k 表示挖掘的频繁项集总数目,$F_{h,i}$,$F_{h,j}$ 分别表示项目 i 与项目 j 在频繁项集矩阵中的取值,分子表示项目 i 与项目 j 在频繁项集中共同出现的次数,分母表示项目 i 和项目 j 分别在频繁项集中出现的次数.在协同过滤中,相似度计算比较常用的是 Pearson 相关系数,其取值范围为 $[-1,1]$,相关系数绝对值越大,相关性越强,而公式(1)取值范围为 $[0,1]$.为同时考虑基于频繁项集的项目相似性和基于协同过滤的相似度,以增强项目之间的相关性,因此将两种相似性进行加权运算.基于协同过滤的项目之间的相似度定义为 $S^{(2)}$,其中每一项 $S_{i,j}^{(2)}$ 表示修订后的 Pearson 相关系数,即项目 i 和项目 j 的相似度,计算公式如下:

$$S_{i,j}^{(2)}=\frac{\sum_{u\in U_{i,j}}|X_{u,i}-\bar{x}_i|\cdot|X_{u,j}-\bar{x}_j|}{\sqrt{\sum_{u\in U_{i,j}}(X_{u,i}-\bar{x}_i)^2}\sqrt{\sum_{u\in U_{i,j}}(X_{u,j}-\bar{x}_j)^2}}, \tag{4}$$

其中,$U_{i,j}$ 表示对项目 i 和项目 j 共同评分过的用户集合,$X_{u,i}$ 和 $X_{u,j}$ 分别表示用户 u 对项目 i 和 j 的评分,$\bar{x}_i$ 和 $\bar{x}_j$ 分别表示用户对项目 i 和项目 j 的平均评分,最后将两个项目之间相似度矩阵进行加权,得到综合的相似度矩阵 $S_{i,j}$,其每一项代表项目 i 和项目 j 的综合相似度,定义为:

$$S_{i,j}=\partial\cdot S_{i,j}^{(1)}+(1-\partial)\cdot S_{i,j}^{(2)}, \tag{5}$$

其中 ∂ 表示权重,为可调的参数.该综合相似度计算方法不仅考虑用户项目评分矩阵,同时还考虑了项目间的内在联系,二者以一定权重结合在一起,可以更准确地寻找项目的最近邻居.

2.2.3 寻找最近邻居

假设目标项目为 I_a,根据相似性矩阵 S 筛选出所有项目与 I_a 的相似性值,并对其进行降序排列,选择相似性最高的 K 个项目作为目标项目 I_a 的最近邻居集合,设为 $N_a=\{N_1,N_2,\cdots,N_K\}$,用户 u 对项目 I_a 的预测评分是 P_{u,I_a},计算方法如下:

$$P_{u,I_a}=\overline{X_{I_a}}+\frac{\sum_{q\in N_a}S_{I_a,q}\cdot(X_{u,q}-\overline{X}_q)}{\sum_{q\in N_a}S_{I_a,q}}, \tag{6}$$

其中 $S_{I_a,q}$ 表示项目 I_a 与项目 q 之间的相似性,$X_{u,q}$ 表示用户 u 对项目 q 的评分,$\overline{X_{I_a}}$,$\overline{X_q}$ 分别表示对项目 I_a 和项目 q 的平均评分.

BFIM 算法以伪代码形式描述如下:

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

BFIM 算法	
输入	评分矩阵 $X \in R^{m\times n}$,目标用户 u,可调参数 ∂,最近邻居数目 K,推荐结果个数 $N(N <= K)$
输出	N 个推荐项目
1	对矩阵 X 筛选目标用户 u 未评分的项目集合 $I_s = \{I_{s1}, I_{s2}, \cdots, I_{sh}\}$;
2	利用 Apriori 算法挖掘频繁项集,构建频繁项集矩阵 $F \in R^{k\times n}$
3	根据矩阵 F 和公式(3) 计算基于频繁项集的项目相似度矩阵 $S^{(1)}$;
4	采用公式(4) 计算基于用户评分数据的项目相似度矩阵 $S^{(2)}$;
5	利用 ∂ 和公式(5) 计算项目间的综合相似性,得到相似性矩阵 S;
6	for $i = 1,2,3,\cdots,h$ do
7	对矩阵 S 筛选 I_{si} 与其他项目的相似度矩阵;
8	根据其相似度值取 Top-K 个项目作为 I_{si} 的最近邻居 N_a;
9	根据公式(6) 计算用户 u 对 I_{si} 的预测评分;
10	end for
11	对步骤 6 ~ 10 的结果进行降序排列,选择 Top-N 个项目作为用户 u 的推荐结果

本文算法实现中需要构建项目 -项目相似度矩阵,其所需要的时间复杂度为 $O(n^2)$,在对 h 个未评分项目中每一个项目进行预测时,需寻找 K 个最近邻居,因此本算法的时间复杂度为 $O(n^2 + h \cdot K)$.

由算法伪代码可知,本文提出的 BFIM 算法在项目相似度计算中考虑了项目内在的关联性,对相似度计算方法进行了改进,下面将通过实验来验证算法的相关性能.

3　实验与分析

3.1　数据集

本文的实验采用目前衡量推荐系统推荐质量的最常用的 MovieLens 数据集(https://grouplens.org/datasets/movielens/),该数据集是由美国 Minnesota 大学 GroupLens 研究小组公布的,它包含了 943 名用户对 1682 部电影的评分,共 10^5 条评分记录. 为了进行实验比较,将数据集进行十折交叉验证,进行 10 次实验,每次实验将数据集的 90%作为训练集,10%作为测试集.

3.2　评价指标

推荐算法的推荐效果评价标准有准确度、覆盖度、多样性等,其中准确度是推荐系统的推荐质量评估中最常用的指标,本文以平均绝对差 MAE 为判断算法推荐质量的标准,其计算方法如公式(7)、(8)所示.

$$MAE_u = \frac{\sum_{i=1}^{N_u} | P_{u,i} - q_{u,i} |}{N_u}, \quad (7)$$

其中,MAE_u 是用户 u 对 N_u 个项目预测评分的平均绝对偏差,$P_{u,i}$ 表示用户 u 对项目 i 的预测评分,$q_{u,i}$ 表示用户 u 对项目 i 的实际评分,N_u 是测试集所提供的用户 u 的评分项目数量.

$$MAE = \frac{\sum_{u=1}^{M} MAE_u}{M}, \quad (8)$$

其中 M 是全体用户总数,MAE 则是测试集全体用户的平均绝对偏差. 可看出,MAE 的值越小,则预测越精确,算法推荐的准确性越高.

3.3　实验结果与分析

实验分为两部分,第一部分是对影响频繁项集挖掘质量的最小支持度阈值 $minSup$ 和综合相似度的加权因子 ∂ 进行选取和比较,找到使算法推荐结果最优的 $minSup$ 和 ∂. 在第一部分实验取得最优的基础上,第二部分实验是对其他推荐算法与本文的 BFIM 算法在不同的最近邻取值下的性能进行比较,从而说明 BFIM 算法的有效性.

3.3.1　$minSup$ 与 ∂ 选取

首先确定最小支持度 $minSup$ 的大小. 将 ∂ 设为 0.2,$minSup$ 在 0.1 ~ 0.4 之间,以 0.05 递增,随着 $minSup$ 的增加,观察算法的 MAE 值的变化. 实验结果如图 1 所示.

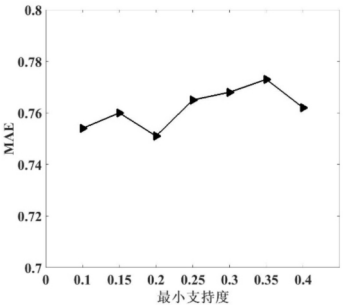

图 1　最小支持度对推荐精度的影响

Fig.1　Effect of minimum support on the accuracy of the recommendation

由图 1 可知,将加权因子 ∂ 设为 0.2 时,$minSup$ 取 0.2 时,MAE 最小,当 $0.2 < minSup < 0.35$ 时,随着支持度的增加,MAE 也在增大,因此,$minSup$ 取 0.2 时,推荐算法效果最好.

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

下一步确定加权因子 ∂ 的大小. 使 $minSup = 0.2$,改变加权因子 ∂ 的大小,$\partial \in [0,1]$,每次递增0.2. 实验结果如图2所示,$\partial = 0.4$ 时 *MAE* 值最小,因此 $minSup = 0.2$,$\partial = 0.4$,BFIM 算法取得最优推荐质量.

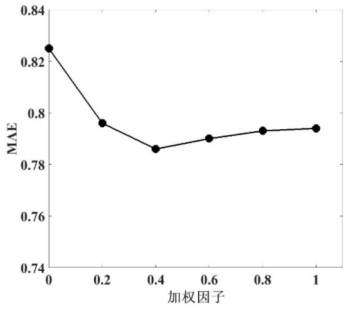

图2 加权因子对推荐精度的影响

Fig.2 Effect of weighting factors on the accuracy of the recommendation

3.3.2 算法比较

为验证本文提出的 BFIM 算法的性能,将基于项目的协同过滤推荐(IBCF 算法),基于项目评分预测的推荐(IRBCF 算法)与本文的 BFIM 算法进行比较,均采用十折交叉验证法进行实验,每次随机将数据集划分为10份,选9份为训练集,1份为测试集,然后得到10次实验的验证结果. 如图3所示,BFIM 算法的 *MAE* 值与其他算法相比大部分都较小,即本文的 BFIM 算法推荐效果更好.

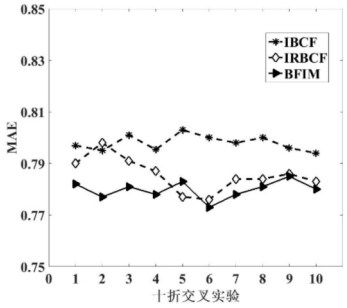

图3 实验次数对推荐精度的影响

Fig.3 Effect of the number of experiments on the accuracy of the recommendation

图4展示了最近邻数目与 *MAE* 值的关系,最近邻数目在5~40之间,随着最近邻数目的增加,3种推荐算法的 *MAE* 值均不同程度地降低,推荐质量不断提高,当最近邻数目一定时,BFIM 算法的 *MAE* 值明显小于传统推荐算法,表明文中提出的 BFIM 算法优于传统的推荐算法.

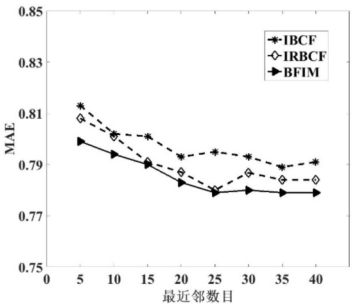

图4 3种算法 *MAE* 对比图

Fig.4 Comparison of three algorithms on *MAE*

4 结语

协同过滤是如今应用成功的推荐算法之一,然而该算法过度依赖评分数据,忽略了项目间关联的特性. 在数据极端稀疏的情况下,传统基于项目的协同过滤算法不能准确找到目标项目的最近邻居,导致推荐不准确. 对此,文中提出的结合频繁项集与项目相似度的协同过滤推荐算法,通过 Apriori 算法找到事务数据库频繁被购买的项目,计算基于频繁项集的项目相似度,再与 Pearson 相关系数加权计算项目的综合相似度,该算法不仅考虑了事务数据的频繁项集,还考虑了用户评分数据,减小了相似度计算值与实际值的偏差,从而提高了推荐算法的质量.

参 考 文 献

[1] WANG Q,WU S.A study of model to analyze the behavior of users of a personalized digital TV recommended system [C] // IEEE. International Conference on Artificial Intelligence, Management Science and Electronic Commerce.Zhengzhou: IEEE,2011: 927-930.

[2] ZHENG X,CHEN C C,HUNG J L,et al.A hybrid trust-based recommender system for online communities of practice [J] .IEEE Transactions on Learning Technologies, 2017,8(4): 345-356.

[3] SHARMA C,BEDI P.CCFRS-Community based collaborative filtering recommender system [J] . Journal of Intelligent & Fuzzy Systems,2017,32(4): 2987-2995.

[4] NGUYEN L T T,VO B,NGUYEN L T T,et al.ETARM: an efficient top-k association rule mining algorithm [J] . Applied Intelligence,2017(5): 1-13.

[5] BREESE J S,HECKERMAN D,KADIE C.Empirical analysis of predictive algorithms for collaborative filtering [J] . Uncertainty in Artificial Intelligence,2013,98(7): 43-52.

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[6] 王成,朱志刚,张玉侠,等.基于用户的协同过滤算法的推荐效率和个性化改进[J].小型微型计算机系统,2016,37(3):428-432.

[7] 黄震华,张佳雯,田春岐,等.基于排序学习的推荐算法研究综述[J].软件学报,2016,27(3):691-713.

[8] 于洪,李俊华.一种解决新项目冷启动问题的推荐算法[J].软件学报,2015,26(6):1395-1408.

[9] SHAMBOUR Q,LU J.A trust-semantic fusion-based recommendation approach for e-business applications[J]. Decision Support Systems,2012,54(1):768-780.

[10] FAN J,PAN W,JIANG L.An improved collaborative filtering algorithm combining content based algorithm and user activity[C].IEEE.2014 International Conference on Big Data and Smart Computing (BIGCOMP). Bangkok: IEEE,2014:88-91.

[11] ZHOU X,HE J,HUANG G,et al.SVD-based incremental approaches for recommender systems[J]. Journal of Computer & System Sciences,2015,81(4):717-733.

[12] 张玉芳,代金龙,熊忠阳.分步填充缓解数据稀疏性的协同过滤算法[J].计算机应用研究,2013,30(9):2602-2605.

[13] CHUJAI P,RASMEQUAN S,SUKSAWATCHON U,et al. Imputing missing values in collaborative filtering using pattern frequent itemsets[C]// IEEE. Electrical Engineering Congress.Chonburi: IEEE,2014:1-4.

[14] INSUWAN W,SUKSAWATCHON U,SUKSAWATCHON J. Improving missing values imputation in collaborative filtering with user-preference genre and singular value decomposition[C]// IEEE. International Conference on Knowledge and Smart Technology.Chonburi: IEEE,2014:87-92.

[15] 刘枚莲,刘同存,张峰.基于双向关联规则项目评分预测的推荐算法研究[J].武汉理工大学学报,2011(9):150-155.

[16] LUO H,NIU C,SHEN R,et al.A collaborative filtering framework based on both local user similarity and global user similarity[J]. Machine Learning,2008,72(3):231-245.

[17] YANG H.Collaborative filtering algorithm based on improved similarity calculation[C]// ICICA. International Conference on Information Computing and Applications. Berlin: Springer,2011:271-276.

(责任编辑　曹　东)

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net