DOI:10.7544/issn1000-1239.2018.20170253
55(5):968-976,2018

一种基于社区专家信息的协同过滤推荐算法

张凯涵　梁吉业　赵兴旺　王智强
(山西大学计算机与信息技术学院　太原　030006)
(计算智能与中文信息处理教育部重点实验室(山西大学)　太原　030006)
(752750403@qq.com)

A Collaborative Filtering Recommendation Algorithm Based on Information of Community Experts

Zhang Kaihan, Liang Jiye, Zhao Xingwang, and Wang Zhiqiang
(*School of Computer and Information Technology, Shanxi University, Taiyuan* 030006)
(*Key Laboratory of Computational Intelligence and Chinese Information Processing (Shanxi University), Ministry of Education, Taiyuan* 030006)

Abstract　Collaborative filtering recommendation algorithm has been widely used because it is not limited by the knowledge in a specific domain and easy to implement. However, it is faced with the problem of several issues such as data sparsity, extensibility and cold start which affect the effectiveness of the recommendation algorithm in some practical application scenarios. To address the user cold start problem, by merging social trust information (i.e., trusted neighbors explicitly specified by users) and rating information, a collaborative filtering recommendation algorithm based on information of community experts is proposed in this paper. First of all, users are divided into different communities based on their social relations. Then, experts in each community are identified according to some criteria. In addition, in order to alleviate the impact of the data sparsity, ratings of an expert's trusted neighbors are merged to complement the ratings of the expert. Finally, the prediction for a given item is generated by aggregating the ratings of experts in the community of the target user. Experimental results based on two real-world data sets FilmTrust and Epinions show the proposed algorithm is able to alleviate the user cold start problem and superior to other algorithms in terms of *MAE* and *RMSE*.

Key words　recommendation system; collaborative filtering; cold start; community; expert information

摘　要　协同过滤推荐算法由于不受特定领域知识限制、简单易实现等优点,得到了广泛的应用.但是,在实际应用中,该类算法往往面临着数据稀疏性、可扩展性、冷启动等问题.为了解决其中的用户冷启动问题,将用户社交信息和评分信息进行融合,提出了一种基于社区专家信息的协同过滤推荐算法.首先,依据用户的社交关系将用户划分为不同的社区;其次,根据一定的准则确定各个社区的专家,并利用社交信息和评分信息对专家评分进行填充进而缓解稀疏性;最后,对冷启动用户根据其所属社区的专家信

收稿日期:2017-04-11;**修回日期**:2017-08-09
基金项目:国家自然科学基金项目(61432011,U1435212,61603230);山西省自然科学基金项目(201601D202039)
This work was supported by the National Natural Science Foundation of China (61432011, U1435212, 61603230) and the Natural Science Foundation of Shanxi Province of China (201601D202039).
通信作者:梁吉业(ljy@sxu.edu.cn)

息进行预测评分.在数据集 FilmTrust 和 Epinions 上与已有协同过滤推荐算法进行了比较分析.实验结果表明,提出的算法可以有效缓解协同过滤推荐算法中的用户冷启动问题,并在平均绝对误差和均方根误差 2 个评价指标上优于已有算法.

关键词 推荐系统;协同过滤;冷启动;社区;专家信息

中图法分类号 TP311

近年来,为了提高推荐系统的准确性与多样性,研究者针对不同问题提出了一系列推荐算法.其中,协同过滤推荐算法由于不受特定领域知识限制、简单易实现等优点,成为了应用最为广泛的算法之一[1].然而,在实际应用过程中,协同过滤推荐算法面临着冷启动问题,对于没有或仅有少量评分信息的新用户,在传统协同过滤推荐算法中无法利用评分信息查找与其兴趣相似的用户.同理,对于新物品也面临着相应问题.冷启动是协同过滤推荐算法中被广泛关注的一个经典问题,冷启动问题的存在严重影响了推荐系统的推荐质量[2-4].例如在电子商务系统中,存在着大量的新用户及新物品,如果推荐系统不能为新用户提供高质量的推荐,将会逐渐失去新用户的信任,从而丢失大量客源;而对于新物品,如果不能及时地将其推荐出去,将会降低这些商品的销售量,使得商家损失经济利益,不利于电子商务系统的长远健康发展.

传统的协同过滤推荐算法假设用户之间是相互独立的.然而现实生活中,用户对一个物品的喜好不仅由其自身决定,还受到身边朋友的影响[5-7].随着互联网技术的飞速发展,目前已有很多网站建立了用户之间的社交关系网络.已有研究表明,合理利用用户的社交关系,可以有效缓解冷启动问题,进而产生更多有意义的推荐[8-10].

社会化推荐算法在解决冷启动问题时不仅利用了用户-物品评分信息,还结合了用户间的社交关系信息.对于一个新用户,只要社会网络中存在与此用户有直接或间接社交关系的用户,就可以根据这种社交关系和已知用户的评分信息,对新用户产生推荐.Massa 等人[9]基于信任的传播性提出一种新的信任度指标 MoleTrust,利用目标用户的信任用户对其产生推荐缓解冷启动问题,但是该方法受信任传播距离影响较大,不够稳定;Guo 等人[10]在传统的协同过滤推荐算法中结合社交网络中用户信任关系,利用信任用户对各物品的评分补充并代表目标用户对各物品的喜好,缓解数据的稀疏性和冷启动问题;Liu 等人[11]利用社交信息改进传统协同过滤推荐算法寻找最近邻的过程,从而缓解无法找到邻居的问题;Jamali 等人[12]提出的 TrustWalker 方法把基于信任的方法与基于物品的推荐方法相结合,有效地缓解了冷启动问题.

在上述融合用户社交信息缓解冷启动问题的研究中,仅仅考虑了用户的行为受其信任用户行为的影响.然而,在现实生活中,用户的行为决策往往受到多种因素的影响,只考虑信任用户而忽略其他因素会导致对用户行为的预测不够准确.尤其对于新用户而言,往往更倾向于参考领域内专家用户的意见,因为专家的意见更客观,在其特定领域内更具有代表性.

针对上述问题,本文提出了一种基于社区专家信息的协同过滤推荐算法,旨在更好地解决协同过滤推荐算法所面临的用户冷启动问题.通过社区划分算法挖掘用户间存在的社区结构,进而在不同社区内寻找代表性强的用户作为专家,并利用新用户与专家在社交网络中的相似性代替传统协同过滤推荐算法中基于评分信息计算的相似度.为了寻找专家,本文提出从用户的评分信息及社交信息 2 方面共同量化用户所具有的代表性,避免了仅利用评分信息带来的局限性.另外,考虑到评分信息的高度稀疏性,充分利用信任信息对专家评分进行填充,弥补数据稀疏对算法性能的影响.最后,在数据集 FilmTrust 和 Epinions 上进行了实验比较分析,结果表明本文所提出的算法可以有效缓解用户冷启动问题,并在平均绝对误差和均方根误差 2 个评价指标上优于已有算法.

1 相关研究

1.1 协同过滤推荐算法

协同过滤推荐算法由 Goldberg 等人[13]在 1992 年提出,由于计算过程仅依赖于用户的历史行为,而无需用户或物品的特征信息,简单高效的计算方法

使其得到广泛应用.在协同过滤推荐算法中,用户的历史行为通常表示为用户-物品评分矩阵 $\boldsymbol{R}_{m\times n}$. $U=\{u_1,u_2,\cdots,u_m\}$ 表示用户集合, $I=\{i_1,i_2,\cdots,i_n\}$ 表示物品集合, R_{ui} 表示用户 u 对物品 i 的评分.

Breese 等人[14]将协同过滤推荐算法分成基于模型(model-based)和基于内存(memory-based)2 类.基于模型的协同过滤推荐算法首先根据训练集数据采用概率统计模型或者机器学习方法建立模型(比如潜在语义模型、贝叶斯模型、决策树模型、图模型等)进而通过模型预测目标用户对目标物品的评分值[15].基于内存的协同过滤推荐算法根据推荐目标不同又分为基于用户(user-based)和基于物品(item-based)两种.本文算法是在基于用户的协同过滤推荐算法框架下提出的,因此,以基于用户的协同过滤推荐算法为例对推荐流程进行介绍.

在基于用户的协同过滤推荐算法中,需要首先利用评分信息计算目标用户与其他用户之间的相似性.用户间相似性的度量方法尤为重要,常见的相似性度量方法包括皮尔逊相关性和余弦相似性,本文采用皮尔逊相关性度量方式[1]:

$$sim(u,v)=\frac{\sum\limits_{i\in I_{uv}}(R_{ui}-\bar{R}_u)(R_{vi}-\bar{R}_v)}{\sqrt{\sum\limits_{i\in I_{uv}}(R_{ui}-\bar{R}_u)^2}\sqrt{\sum\limits_{i\in I_{uv}}(R_{vi}-\bar{R}_v)^2}},\tag{1}$$

其中, $sim(u,v)$ 表示用户 u 和用户 v 之间的评分相似性, I_{uv} 表示用户 u 和用户 v 共同评分过的物品集合, $\bar{R}_u$ 和 $\bar{R}_v$ 分别表示用户 u 和用户 v 的评分均值.其次,按照相似性大小对其他用户进行排序,取前 k 个相似用户作为目标用户的邻居.最后,利用邻居用户对目标物品的评分来预测目标用户对目标物品的评分:

$$\hat{R}_{ui}=\bar{R}_u+\frac{\sum\limits_{v\in NU_u}sim(u,v)(R_{vi}-\bar{R}_v)}{\sum\limits_{v\in NU_u}sim(u,v)},\tag{2}$$

其中, $\hat{R}_{ui}$ 表示用户 u 对物品 i 的预测评分值, NU_u 表示用户 u 的相似邻居集合.根据以上描述,基于用户的协同过滤推荐算法描述如下:

算法 1. 基于用户的协同过滤推荐算法.

输入:用户-物品评分矩阵 $\boldsymbol{R}_{m\times n}$、邻居个数 k、目标用户 u、目标物品 i;

输出:预测评分值 $\hat{R}_{ui}$.

步骤 1. 利用式(1)计算目标用户 u 与其他用户的相似性;

步骤 2. 对相似性计算结果按从大到小排序,选取前 k 位用户作为目标用户 u 的邻居,构成相似邻居集合 NU_u;

步骤 3. 利用式(2)预测目标用户 u 对目标物品 i 的评分 $\hat{R}_{ui}$.

1.2 社会化推荐算法

自 1997 年社会化推荐系统被提出以来,社会化推荐系统吸引了大量学者的关注,尤其近年来微博、微信、Facebook 等社交媒体的迅速发展,更促进了学者们对社会化推荐算法的研究.文献[16]概括了社会化推荐算法的狭义定义和广义定义,其中狭义的社会化推荐是指任何将社交关系(例如信任关系、朋友关系等)作为附加输入的推荐算法,而广义的社会化推荐是指以社会媒体(例如物品、标签、社区等)为推荐目标的推荐算法,利用的数据源也不仅是社交信息,还包括各种可利用的社会化标签、用户间的交互信息以及用户的点击行为等.目前,已有相关研究利用用户间的社交信息提高推荐系统的性能,例如 MoleTrust[9], TrustSVD[17], SoRec[18]等算法.

在社会化推荐算法中,除了利用传统协同过滤推荐算法中的用户-物品评分矩阵 $\boldsymbol{R}_{m\times n}$,还需利用用户之间的社交信息.本文所利用的社交信息为用户间的信任关系,通常使用矩阵 $\boldsymbol{T}_{m\times m}$ 表示. $T_{uv}=1$ 表示用户 u 对用户 v 具有信任关系, $T_{uv}=0$ 表示没有关系.注意,信任关系为非对称关系,即用户 u 对 v 有信任连边,但 v 对 u 可能并没有信任连边.社交信息为推荐系统提供了一个新的信息源,为传统协同过滤推荐算法因评分信息匮乏所产生的冷启动问题提供了新的解决策略.

2 基于社区专家信息的协同过滤推荐算法

通过上述分析,本文在基于用户的协同过滤推荐算法框架下,将社交信息与专家信息融入推荐过程中,利用填充的专家用户评分对新用户的评分进行预测,从而缓解冷启动问题.下面将重点关注 3 个问题:1)如何利用社交信息与评分信息选择专家;2)如何对专家评分进行填充;3)如何利用专家信息对目标用户进行评分预测.表 1 列出了本文使用的主要符号.图 1 为本文算法示意图.

Table 1 The Main Symbols Used in the Paper

表 1 本文用到的主要符号

Symbol	Description
$\boldsymbol{R}_{m\times n}$	user-item rating matrix
$\boldsymbol{T}_{m\times m}$	user-user trust relations matrix
R_{ui}	the rating of user u on item i
$\hat{R}_{ui}$	the predicted rating of user u on item i
T_{uv}	the trust value of user u for user v
I_{uv}	co-rated items of user u and user v
$sim(u,v)$	the similarity between user u and user v
$\bar{R}_u$	the average rating of user u
NU_u	the set of similar users of user u
TU_u	the set of trusted neighbors of user u
C_i	the ith user community
$C(u)$	the community that user u belongs to
k	the number of similar users
$Reputation(u)$	the reputation of user u

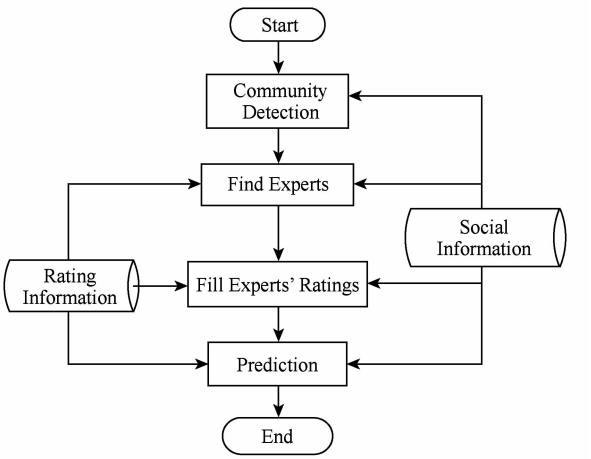


Fig. 1 The diagram of the proposed algorithm

图 1 本文算法示意图

2.1 用户社区划分

为了综合考虑目标用户所属领域内专家用户对其行为产生的影响,首先根据用户的社交信息将用户划分为不同的社区. 用 $C=\{C_1,C_2,\cdots,C_p\}$ 表示用户社区集合,其中 $C_i=\{u_1^{C_i},u_2^{C_i},\cdots,u_{|C_i|}^{C_i}\}$,$u_k^{C_i}$ 表示社区 C_i 中的第 k 位用户,$|C_i|$ 表示社区 C_i 中用户的个数. 本文采用 SLM(smart local moving)[19] 社区划分算法对用户进行划分. 该算法通过最大化模块度来完成,在保证社区质量的情况下所需要的时间较少,已经成功应用到有数百万节点和亿万条边的网络中.

通过用户社区的划分得到多个社区集合,各个用户社区所包含的用户数量不一定相等,社区的划分潜在地反映了部分用户群体对某类物品的偏好. 对于用户量较大的社区,可理解为多数用户对某类热门物品的偏好,而用户量较少的社区,反映了小众用户对某类物品的特殊偏好.

2.2 社区中专家的确定

在划分得到用户社区的基础上,需要根据一定的准则在各社区内寻找具有代表性的用户,构成该社区的专家集合. 假设经划分共得到 p 个社区集合,$E=\{E^1,E^2,\cdots,E^p\}$ 表示所有社区的专家集合,其中 $E^g=\{e_1^g,e_2^g,\cdots,e_{|E^g|}^g\}$ $(g=1,2,\cdots,p)$ 表示社区 C_g 中的专家集合,共 $|E^g|$ 位专家.

为了度量每个社区中各个用户所具有的代表性,以此判断该用户是否能够成为专家. 本文分别从用户的社交关系和用户对物品的评分信息出发,定义了可信度、权威性以及评分多样性 3 个指标,对用户成为专家的可能性进行评价.

定义 1. 用户可信度. 可信度反映用户被其他用户所信任的程度,通过在信任网络中入度的大小来衡量. 用户 u 的可信度表示为

$$B_u=\frac{d_u}{d_{\max}^{C(u)}}, \tag{3}$$

其中,d_u 表示用户 u 的入度,$d_{\max}^{C(u)}$ 表示 u 所属社区 $C(u)$ 中所有用户入度的最大值.

定义 2. 用户权威性. 权威性反映用户的活跃程度,通过用户评分数量的多少来刻画. 评分数量越多,说明用户在系统中越活跃,相对于评分数量少的用户其在评分预测时更具有参考价值,因此权威性更高. 用户 u 的权威性表示为

$$A_u=1-\frac{1}{N(u)}, \tag{4}$$

其中,$N(u)\geqslant 1$ 表示用户 u 对物品的评分数量.

定义 3. 用户评分多样性. 评分多样性反映用户对不同物品所具有评分值的差异性. 用户对不同物品应具有不同的评分值,如果用户对所有物品的评分值都一样,则不能体现对各物品的喜好程度. 因此通过评分方差来度量用户 u 的评分多样性,表示为

$$D_u=\mathrm{e}^{-\frac{1}{v_u+1}}, \tag{5}$$

其中,v_u 表示用户 u 对物品评分值的方差.

因此,社区内每个用户成为专家的可能性为

$$Reputation(u)=\frac{1}{3}(B_u+A_u+D_u). \tag{6}$$

根据此值从大到小的顺序对社区内每个用户所具有的代表性进行排序选择各社区中的专家,每个

社区中专家所占比例定义为 γ,则对于第 g 个社区来说,专家比例表示为

$$\gamma=\frac{|E^g|}{|C_g|}. \tag{7}$$

2.3 专家评分的填充

2.2 节通过对用户代表性的量化找到可以代表各社区的专家用户,考虑到用户-物品评分信息的高度稀疏性,因此本节提出利用专家用户的信任用户的评分信息对专家评分进行填充,缓解专家评分的稀疏问题. 专家 e 的信任用户集表示为

$$TN_e=\{v\,|\,T_{ev}>\theta, v\in U\}. \tag{8}$$

由于本文所利用的社交关系为用户信任关系,只用数值 0 和 1 表示,因此本文将设置 $\theta=0$,即专家 e 显性声明具有信任关系的用户视为 e 的信任用户.

对专家 e 的评分信息进行填充时,欲填充的候选物品集应是 e 的信任用户产生过评分,而 e 没有评分的物品,因此填充时的候选物品集表示为

$$\tilde{I}_e=\{i\,|\,\exists i\in I_e^-, |U_i\cap TN_e|\geqslant\beta\}, \tag{9}$$

其中,I_e^- 表示专家 e 未产生过评分的物品集,U_i 表示对物品 i 有过评分的用户集合.

为了控制算法的复杂度及精度,本文在对专家 e 的评分进行填充时,候选物品集只考虑至少被 e 的 5 个信任用户所评分过的物品,即 $\beta=5$(参数 β 的选取在 3.3.2 节说明).

最后,对候选物品集中的物品使用下式填充专家 e 的评分值:

$$\tilde{R}_{ei}=\frac{\sum\limits_{v\in TN_e}T_{ev}R_{vi}}{\sum\limits_{v\in TN_e}T_{ev}}, \tag{10}$$

其中,$\tilde{R}_{ei}$ 表示专家 e 对物品 i 的填充评分值.

因此,专家 e 对物品 i 的评分值为

$$R_{ei}=\begin{cases}\tilde{R}_{ei}, & i\in\tilde{I}_e.\\ R_{ei}, & \text{others}.\end{cases} \tag{11}$$

2.4 预测评分

在新用户 u 所属的社区 $C(u)$ 中利用专家信息预测 u 对目标物品 i 的评分.

首先,利用 Salton 指标[20]在社交网络中计算目标用户与专家之间的相似性:

$$Salton(u,e)=\frac{|\Gamma(u)\cap\Gamma(e)|}{\sqrt{k_uk_e}}, \tag{12}$$

其中,$\Gamma(u)$ 和 $\Gamma(e)$ 分别表示信任网络中用户 u 和专家 e 所信任的用户集合,k_u 和 k_e 分别表示用户 u 和专家 e 的出度.

最后,结合社区 $C(u)$ 内专家与用户 u 的相似度以及对物品 i 的评分值进行加权求和,得到最终的预测结果:

$$\hat{R}_{ui}=\frac{\sum\limits_{e\in E^{C(u)}}Salton(u,e)R_{ei}}{\sum\limits_{e\in E^{C(u)}}Salton(u,e)}, \tag{13}$$

其中,$E^{C(u)}$ 表示用户 u 所属社区 $C(u)$ 的专家集合.

2.5 基于社区专家信息的协同过滤推荐算法

基于以上对算法各个主要阶段的介绍,本文提出的算法描述如下:

算法 2. 基于社区专家信息的协同过滤推荐算法.

输入:用户-物品评分矩阵 $\boldsymbol{R}_{m\times n}$、用户社交关系矩阵 $\boldsymbol{T}_{m\times m}$、专家数量占比 γ、目标用户 u、目标物品 i、参数 β;

输出:预测评分值 $\hat{R}_{ui}$.

步骤 1. 对社交关系 $\boldsymbol{T}$ 利用 SLM 算法将用户划分为不同社区.

步骤 2. 利用式(3)~(6)计算各社区内每个用户的代表性,由大到小对用户代表性排序,前 $\gamma|C(u)|$ 位用户选为社区专家.

步骤 3. 结合专家的信任用户,根据式(9)选择待填充评分的候选物品集,利用式(10)填充专家对候选物品集中各物品的评分.

步骤 4. 利用式(12)计算 u 与各专家之间的相似度,最后根据式(13)预测评分 $\hat{R}_{ui}$.

3 实验及结果分析

为验证本文所提算法的有效性,在真实数据集 FilmTrust 和 Epinions 上进行了实验,并与其他推荐算法进行比较,最后通过实验分析本文所提算法中参数的选取对实验性能的影响. 实验环境为:4 GB 内存、Intel® Core™ 2 Quad 处理器、2.66 GHz,Windows7 操作系统.

3.1 数据集

由于本文所提算法需要运用到用户的社交信息,因此选择常用数据集 FilmTrust 和 Epinions. 这 2 个数据集不仅具有用户-物品的评分信息,还具有社交网络中用户之间的信任关系信息.

数据集 FilmTrust 包含了 1 508 位用户对 2 071 部电影的 35 497 条评分信息,以及 1 642 位用户间 1 853 条信任关系. 信任关系表示了用户对其他用户是否产生信任,如果一个用户信任另一用户,在数据

集中用 1 表示,否则用 0 表示. 其中评分值在 0.5～4 之间.

数据集 Epinions 的评分信息表示了用户对电影、图书以及汽车等物品的评分,用数值 1～5 表示,该数据集中包含了 40 163 位用户对 139 738 个物品的 664 824 条评分数据. 此外,还包含了 487 183 条用户之间的信任关系. 表 2 统计了这 2 个数据集的相关信息.

Table 2 The Specifications of Two Data Sets

表 2 2 个数据集统计信息

Data Set	Rating Information				Social Information	
	# Users	# Items	# Ratings	Density/%	# Users	# Edges
FilmTrust	1 508	2 071	35 497	1.14	1 642	1 853
Epinions	40 163	139 738	664 824	0.011 8	49 289	487 183

3.2 评价指标

本文在衡量推荐性能时,为体现预测评分的准确度,采用了推荐系统中广泛使用的平均绝对误差(mean absolute error, *MAE*)和均方根误差(root mean squared error, *RMSE*)两个评价指标. 这 2 个评价指标的值越小表示预测效果越好.

MAE 可表示为

$$MAE = \frac{1}{|R_{\text{test}}|}\sum_{R_{ui}\in R_{\text{test}}}|\hat{R}_{ui}-R_{ui}|, \quad (14)$$

其中,$|R_{\text{test}}|$ 表示测试集中的评分数量.

RMSE 可表示为

$$RMSE = \sqrt{\frac{1}{|R_{\text{test}}|}\sum_{R_{ui}\in R_{\text{test}}}(\hat{R}_{ui}-R_{ui})^2}. \quad (15)$$

3.3 实验设置

为了验证本文所提算法对评分预测性能的提升以及对冷启动问题的处理效果,在 *MAE*,*RMSE* 指标上对以下算法进行比较:

1) 基于用户的协同过滤推荐算法(user-based collaborative filtering, UCF). 基于预先定义的相似性度量方法以及用户的邻居数量,通过用户邻居的评分信息对目标用户进行预测.

2) 基于物品的协同过滤推荐算法(item-based collaborative filtering, ICF). 基于预先定义的相似性度量方法以及物品的邻居数量,结合物品邻居的评分信息预测目标评分.

3) MoleTrust x[9]. 用户之间的信任关系在信任网络中以距离 x 进行传播,只有被目标用户所信任的用户才会被考虑参与到评分预测.

4) 融合相似用户与朋友的协同过滤推荐算法(combine neighbors and friends collaborative filtering, CNCF)[11]. 利用评分信息与社交信息,根据预先定义的相似性度量方法,由目标用户的信任用户与评分相似度最大的用户共同构成近邻用户,预测评分时与传统的基于用户的协同过滤推荐算法相同.

5) 未填充专家评分的基于社区专家信息的协同过滤推荐算法(a collaborative filtering recommendation algorithm based on information of community experts without filling ratings, CECF). 该算法与本文 2.5 节所提算法区别在于不考虑 2.3 节对专家评分的填充,仅利用原有专家评分对目标用户进行预测.

6) 基于社区专家信息的协同过滤推荐算法(a collaborative filtering recommendation algorithm based on information of community experts, CEFCF). 本文 2.5 节所提算法,其中专家评分依据 2.3 节所述进行填充.

本文分别针对冷启动用户和全部用户(包含非冷启动用户和冷启动用户)进行实验. 在全部用户的实验中采用五折交叉验证方法,将数据集随机分为 5 份,每次取其中 1 份作为测试集,剩余 4 份作为训练集,最终结果为 5 次实验结果的平均值. 文献中通常将数据集中评分数量小于 5 的用户视为冷启动用户[10]. 为了模拟对冷启动用户的评分预测实验,从数据集 FilmTrust 和 Epinions 中分别选取部分用户,并将他们的部分评分信息隐藏,使每个用户的评分数量低于 5,将其作为冷启动用户进行分析.

本实验相似度计算方法均采用皮尔逊相似度. UCF 算法、ICF 算法与 CNCF 算法中邻居数量均设置为 30. MoleTrust 算法中信任的传播距离分别采用 1,2,3,表示为 MT-1, MT-2, MT-3. CECF 和 CEFCF 算法中专家占比采用 0.2,实验结果如表 3～6 所示.

Table 3 The Predictive Performance for All Users on the FilmTrust Data Set

表 3 数据集 FilmTrust 上对全部用户的预测性能

Criteria	UCF	ICF	MT-1	MT-2	MT-3	CNCF	CECF	CEFCF
MAE	0.732 1	0.832 0	0.879 6	0.823 5	0.783 6	0.732 1	0.779 2	**0.715 2**
RMSE	0.937 1	1.088 9	1.163 1	1.077 1	1.023 6	0.940 1	1.040 3	**0.931 0**

Table 4 The Predictive Performance for Cold Users on the FilmTrust Data Set

表 4 数据集 FilmTrust 上对冷启动用户的预测性能

Criteria	UCF	ICF	MT-1	MT-2	MT-3	CNCF	CECF	CEFCF
MAE	0.702 8	0.845 1	0.814 3	0.786 4	0.769 1	0.719 8	0.670 3	**0.640 8**
RMSE	0.912 4	1.080 7	1.066 6	1.039 7	1.024 6	0.934 6	0.962 8	**0.845 0**

Table 5 The Predictive Performance for All Users on the Epinions Data Set

表 5 数据集 Epinions 上对全部用户的预测性能

Criteria	UCF	ICF	MT-1	MT-2	MT-3	CNCF	CECF	CEFCF
MAE	1.079 2	1.010 6	**0.873 3**	0.923 8	0.906 7	0.972 8	0.910 4	0.876 7
RMSE	1.436 0	1.335 9	1.262 9	1.272 7	1.236 3	1.303 9	1.260 2	**1.172 2**

Table 6 The Predictive Performance for Cold Users on the Epinions Data Set

表 6 数据集 Epinions 上对冷启动用户的预测性能

Criteria	UCF	ICF	MT-1	MT-2	MT-3	CNCF	CECF	CEFCF
MAE	1.126 0	1.034 7	0.914 0	0.929 5	0.915 5	1.116 8	0.915 9	**0.881 7**
RMSE	1.471 7	1.373 0	1.326 3	1.299 7	1.270 2	1.462 8	1.268 5	**1.174 0**

从实验结果看出:在评分信息相对稠密的小规模数据集 FilmTrust 上,传统的 UCF 算法表现仍较为可观,然而该数据集物品量相对较少,因此传统 ICF 算法表现一般. 在稀疏的大规模数据集 Epinions 上,基于社会网络的推荐算法具有更好的推荐效果,说明信任信息的引入确实可以缓解协同过滤推荐算法所面临的稀疏性问题. 本文所提出的基于社区专家信息的 CEFCF 算法虽然在数据集 Epinions 的全部用户预测中 *MAE* 指标表现欠优,但是在 *RMSE* 指标上均胜过了其他算法,而本文也更关注对冷启动用户的预测问题. 如实验结果所示,引入专家信息的 CECF 算法和填充专家评分的 CEFCF 算法在 2 个数据集的冷启动用户上均具有良好的推荐性能,说明专家的引入确实能够提升系统对新用户的预测性能. 而对专家评分进行过填充的 CEFCF 算法较未填充的 CECF 算法更优,说明本文对专家评分的填充确实进一步弥补了稀疏性问题对系统的影响,对冷启动用户的预测也更加准确.

3.3.1 专家比例 γ 的选取对算法性能的影响

图 2 为数据集 FilmTrust 和 Epinions 上社区内

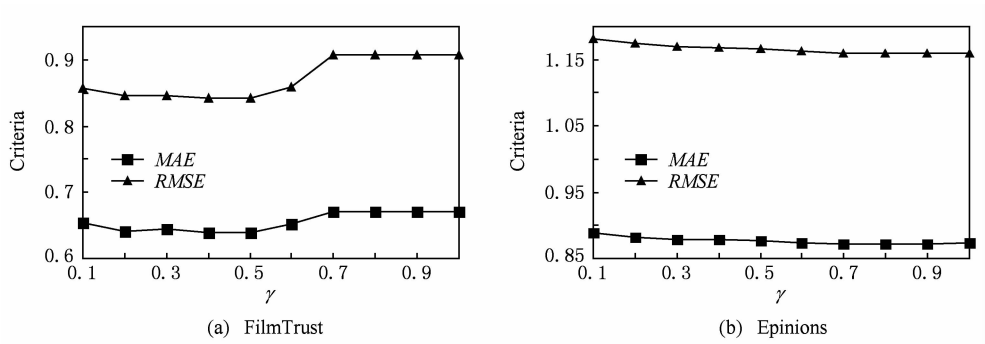


Fig. 2 The effect of γ on the system performance

图 2 专家比例 γ 对算法性能的影响

专家比例 γ 对 CEFCF 算法冷启动用户评分预测性能的影响.如图 2 所示,专家比例对规模较大的数据集 Epinions 影响很小,随着比例逐渐增大,*MAE* 与 *RMSE* 仅有小幅下降,并且在 $\gamma=0.6$ 处逐渐趋于稳定.而对于小规模数据集 FilmTrust,在 $\gamma=0.2$ 处推荐性能已经取得不错效果,随着 γ 的增长,当 $\gamma=0.6$ 时,推荐性能甚至有所下降.随着专家数量的增加,起初可以利用更多的专家信息预测目标评分,所以推荐性能有一定提升.但是当专家数量达到一定比例时,由于过多地将低质量用户选为专家,因此不会再产生高质量的预测效果,甚至在小规模数据集中使得推荐质量下降.因此本文算法中专家比例选择为 0.2.

3.3.2 参数 β 的选取对算法性能的影响

图 3 为参数 β 对算法预测性能及时间损耗的影响.其中图 3(a)(b)分别表示在数据集 FilmTrust 和 Epinions 上算法预测性能随 β 的变化,图 3(c)为 2 个数据集上 β 对算法时间损耗的影响.实验中计算了专家评分填充以及利用填充的专家评分产生推荐所消耗的时间.在数据集 FilmTrust 上,如图 3(a)所示,随着 β 的增长,算法预测性能小幅提升后在 $\beta=5$ 处逐渐趋于平稳,由于数据集 FilmTrust 很小,因此在图 3(c)中时间消耗趋近于 0,并不明显.在数据集 Epinions 上,图 3(b)显示 β 对算法预测性能只产生微弱影响.但是据图 3(c),在 β 下降过程中,算法的时间损耗成倍增长,并且 $\beta<3$ 后,由于填充评分量的增多使得算法因机器内存原因运行受限.因此为了在算法预测性能与时间损耗之间寻求折中,本文设置 $\beta=5$.

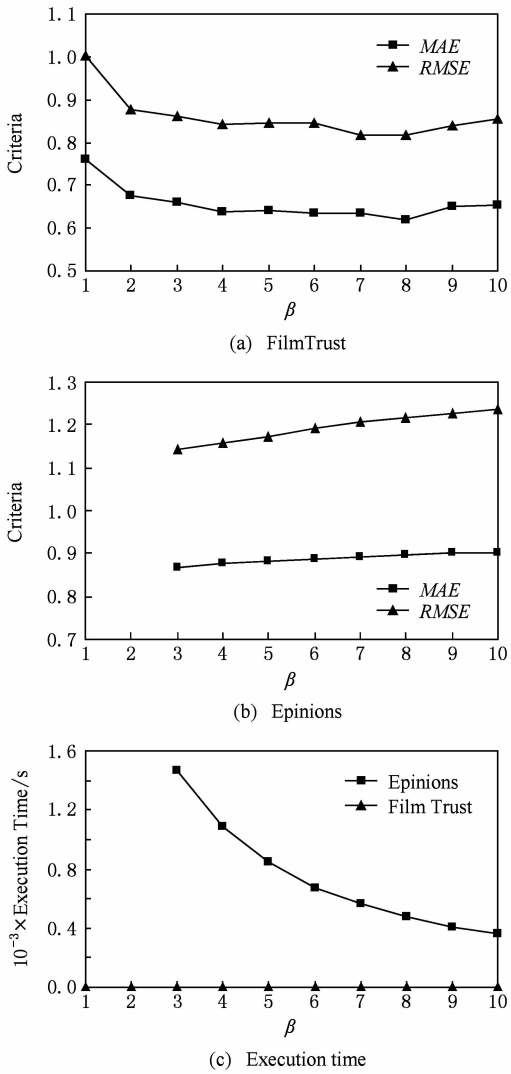


Fig. 3 The effect of β on the system performance

图 3 参数 β 对算法性能的影响

4 总 结

本文提出基于社区专家信息的协同过滤推荐算法,首先依据社交信息将用户划分为不同社区,在各社区内综合考虑用户的评分信息和社交信息,进而选取代表性强的用户作为专家.通过对专家评分的填充更有效地缓解了评分稀疏性的影响.利用各社区的专家对新用户产生推荐,有效缓解了传统协同过滤推荐算法所面临的用户冷启动问题.

本文所提算法只考虑了用户的社交信息,在未来的研究中,将从多个角度综合考虑用户、物品的属性等信息,寻找解决推荐系统中冷启动问题更好的方法.

参 考 文 献

[1] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions [J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(6): 734-749

[2] Pereira A L V, Hruschka E R. Simultaneous co-clustering and learning to address the cold start problem in recommender systems [J]. Knowledge-Based Systems, 2015, 82: 11-19

[3] Wang Zhiqiang, Liang Jiye, Li Ru, et al. An approach to cold-start link prediction: Establishing connections between non-topological and topological information [J]. IEEE Trans on Knowledge and Data Engineering, 2016, 28(11): 2857-2870

[4] Li Xin, Liu Guiquan, Li Lin, et al. Circle-based and social connection embedded recommendation in LBSN [J]. Journal of Computer Research and Development, 2017, 54(2): 394-404 (in Chinese)

(李鑫, 刘贵全, 李琳, 等. LBSN 上基于兴趣圈中社会关系挖掘的推荐算法[J]. 计算机研究与发展, 2017, 54(2): 394-404)

[5] Meng Xiangwu, Liu Shudong, Zhang Yujie, et al. Research on social recommender systems [J]. Journal of Software, 2015, 26(6): 1356-1372 (in Chinese)
(孟祥武, 刘树栋, 张玉洁, 等. 社会化推荐系统研究[J]. 软件学报, 2015, 26(6): 1356-1372)

[6] Guo Lanjie, Liang Jiye, Zhao Xingwang. Collaborative filtering recommendation algorithm incorporating social network information [J]. Pattern Recognition and Artificial Intelligence, 2016, 29(3): 281-288 (in Chinese)
(郭兰杰, 梁吉业, 赵兴旺. 融合社交网络信息的协同过滤推荐算法[J]. 模式识别与人工智能, 2016, 29(3): 281-288)

[7] Guo Hongyi, Liu Gongshen, Su Bo, et al. Collaborative filtering recommendation algorithm combining community structure and interest clusters [J]. Journal of Computer Research and Development, 2016, 53(8): 1664-1672 (in Chinese)
(郭弘毅, 刘功申, 苏波, 等. 融合社区结构和兴趣聚类的协同过滤推荐算法[J]. 计算机研究与发展, 2016, 53(8): 1664-1672)

[8] Yang Bo, Lei Yu, Liu Jiming, et al. Social collaborative filtering by trust [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2017, 39(8): 1633-1647

[9] Massa P, Avesani P. Trust-aware recommender systems [C] //Proc of the 2007 ACM Conf on Recommender Systems. New York: ACM, 2007: 17-24

[10] Guo Guibing, Zhang Jie, Thalmann D. Merging trust in collaborative filtering to alleviate data sparsity and cold start [J]. Knowledge-Based Systems, 2014, 57: 57-68

[11] Liu Fengkun, Lee H J. Use of social network information to enhance collaborative filtering performance [J]. Expert Systems with Applications, 2010, 37(7): 4772-4778

[12] Jamali M, Ester M. TrustWalker: A random walk model for combining trust-based and item-based recommendation [C] // Proc of the 15th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2009: 397-406

[13] Goldberg D, Nichols D, Oki B M, et al. Using collaborative filtering to weave an information tapestry [J]. Communications of the ACM, 1992, 35(12): 61-70

[14] Breese J S, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering [C] //Proc of the 14th Conf on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann, 1998: 43-52

[15] Park D H, Kim H K, Choi I Y, et al. A literature review and classification of recommender systems research [J]. Expert Systems with Applications, 2012, 39(11): 10059-10072

[16] Tang Jiliang, Hu Xia, Liu Huan. Social recommendation: A review [J]. Social Network Analysis and Mining, 2013, 3(4): 1113-1133

[17] Guo Guibing, Zhang Jie, Smith N Y. TrustSVD: Collaborative filtering with both the explicit and implicit influence of user trust and of item ratings [C] //Proc of the 29th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2015: 123-129

[18] Ma Hao, Yang Haixuan, Lyu M R, et al. SoRec: Social recommendation using probabilistic matrix factorization [C] //Proc of the 17th ACM Conf on Information and Knowledge Management. New York: ACM, 2008: 931-940

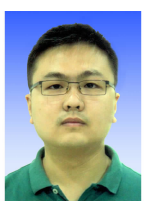
[19] Waltman L, Eck N J V. A smart local moving algorithm for large-scale modularity-based community detection [J]. The European Physical Journal B, 2013, 86(11): 1-14

[20] Gerard S, Michael J M. Introduction to Modern Information Retrieval [M]. Auckland: MuGraw-Hill, 1983

Zhang Kaihan, born in 1994. PhD candidate. Student member of CCF. Her main research interests include data mining and machine learning.

Liang Jiye, born in 1962. Professor and PhD supervisor. Distinguished member of CCF. His main research interests include granular computing, data mining and machine learning.

Zhao Xingwang, born in 1984. PhD candidate. Member of CCF. His main research interests include data mining and machine learning (zhaoxw84@163.com).



Wang Zhiqiang, born in 1987. PhD candidate. Member of CCF. His main research interests include social network analysis and machine learning (zhiq.wang@163.com)