ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2018/12(02)-0208-10
doi: 10.3778/j.issn.1673-9418.1702012

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-89056056

融合社交网络特征的协同过滤推荐算法*

郭宁宁[1],王宝亮[1+],侯永宏[1],常 鹏[2]
1. 天津大学 电子信息工程学院,天津 300072
2. 天津大学 信息与网络中心,天津 300072

Collaborative Filtering Recommendation Algorithm Based on Characteristics of Social Network*

GUO Ningning[1], WANG Baoliang[1+], HOU Yonghong[1], CHANG Peng[2]
1. School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China
2. Information and Network Center, Tianjin University, Tianjin 300072, China
+ Corresponding author: E-mail: wangbl@tju.edu.cn

GUO Ningning, WANG Baoliang, HOU Yonghong, et al. Collaborative filtering recommendation algorithm based on characteristics of social network. Journal of Frontiers of Computer Science and Technology, 2018, 12 (2): 208-217.

Abstract: To solve the severe sparseness problem of traditional collaborative filtering recommendation algorithm, this paper proposes a novel collaborative filtering recommendation algorithm based on the characteristics of social network. On the basis of traditional matrix decomposition model, the algorithm obtains the trust and trusted characteristic matrix by integrating the characteristics of social network and user's preference degree, and then, predicts the rating of the commodity by the social identity matrix, the commodity characteristic matrix and the user rating preference similarity in common. In order to verify the reliability of the proposed algorithm, this paper uses the Epinions open dataset to compare the algorithm performance. The experimental results show that compared with the existing social recommendation algorithms, the proposed algorithm has smaller average absolute error and root mean square error. Meanwhile, there is a linear relationship between the time complexity of the proposed algorithm and the number of the dataset. Therefore, the proposed algorithm can effectively reduce the impact of data sparseness on recommendation results and improve the recommendation accuracy rate. In practice, the proposed algorithm can be consid-

* The National Natural Science Foundation of China under Grant No. 61571325 (国家自然科学基金).
Received 2017-02, Accepted 2017-04.
CNKI网络优先出版: 2017-04-19, http://kns.cnki.net/kcms/detail/11.5602.TP.20170419.1308.002.html

ered as an alternative and development of the large-scale data set recommendation.
Key words: recommender system; social network; collaborative filtering; user rating preference; rating prediction

摘　要：为了解决传统协同过滤算法中存在的严峻的数据稀疏性问题，提出了一种融合社交网络特征的协同过滤推荐算法。该算法在传统矩阵分解模型基础上，通过融合社交网络特征与用户评分偏好程度得到信任和被信任特征矩阵，然后利用社交特征矩阵、商品特征矩阵和用户评分偏好相似性共同预测用户对商品的评分值。为了验证该算法的可靠性，使用Epinions公开数据集对算法性能进行对比分析。实验结果显示，相比现有的社交推荐算法，所提算法有更小的平均绝对误差和均方根误差，同时算法的时间复杂度与数据集的数量之间为线性关系。因此，该算法可以有效缓解数据稀疏性对推荐结果的影响，并提高推荐准确率。在现实推荐中，该算法可以考虑作为大规模数据集进行商品推荐的一个选择方式。
关键词：推荐系统；社交网络；协同过滤；用户评分偏好；评分预测
文献标志码：A　　**中图分类号：**TP301

1　引言

大数据时代的快速发展造成日益严重的信息过载现象，信息检索已经无法满足用户日益增长的个性化信息获取的需求。个性化推荐系统因为其可靠性高，推荐结果准确，迅速成为解决信息过载的方式之一。其基本思想是依据用户的历史行为推荐用户感兴趣的用户或商品集，并且为获得更好的用户体验提供个性化服务[1]。目前，个性化推荐技术主要分为协同过滤推荐[2]、基于内容的推荐[3]、基于图的推荐[4]、混合推荐技术[5]。其中协同过滤推荐又包括基于模型的协同过滤[6]和基于记忆的协同过滤[7]，该技术在分析用户资源的基础上，充分挖掘用户潜在兴趣，并以此作为预测和推荐依据，现已成为推荐系统中发展最成熟、应用最广泛的推荐技术，也是本文主要的研究对象。

协同过滤推荐技术取得广泛应用，在信息海洋时代节省了用户获取信息的时间代价，但该技术也存在一些固有缺陷。首先，协同过滤技术依靠用户-商品评分矩阵进行推荐，但是在现实生活中评分矩阵存在严重的数据稀疏性问题；其次，部分用户只对很少部分商品进行评分，因此该技术存在冷启动问题；最后，传统的协同过滤技术仅仅依靠用户-商品评分矩阵为用户推荐，由于数据源单一，造成推荐结果失真[8]。

随着社交网络技术的发展，用户之间的联系更多依赖网络社交工具，比如Facebook、微信等。社交关系为推荐系统提供了一个独立的信息源，在社交推荐算法中起着越来越重要的作用[9]。基于社交关系的推荐方法[10-14]，在一定程度上缓解了用户稀疏性和冷启动问题，同时提高了推荐准确率，但是依然存在一些问题。首先现有的网络数据集中，只有少部分数据集有用户社交关系矩阵，且矩阵数值都是二值数据，因此很难区别用户间的信任程度；其次对用户社交数据建模时，大部分模型建立仅仅依靠用户显性信任关系，从而忽视了用户的隐性社交关系，如相似性等。

为解决上述存在的问题，本文提出了一种融合社交网络关系的协同过滤推荐方法，即融合用户间社交网络信息和用户评分信息，对传统基于分解模型的协同过滤算法进行优化。本文方法包含以下几个步骤：(1)将用户评分矩阵和用户信任矩阵分别映射到低维空间，即用户空间、商品空间、信任空间和被信任空间；(2)用社交特征、商品特征矢量和用户相似性近似估计稀疏用户评分矩阵；(3)依据密集用户评分矩阵选择评分最高的N个商品，形成推荐列表。最后在Epinions公开数据集上验证本文算法，证明了该算法有效缓解了数据稀疏性对推荐结果的影响，并有效降低了平均绝对误差，提高了推荐准确率。

本文组织结构如下：第2章简要介绍传统分解模型的协同过滤算法和本文提出的基于社交网络关系

的协同过滤方法;第3章对本文推荐算法进行仿真验证与实验结果分析;第4章对全文进行总结。

2 基于社交关系的协同过滤推荐算法

2.1 问题分析

传统协同过滤推荐算法往往只分析用户的评分矩阵数据,容易忽视用户之间存在的社交信息。但在实际推荐应用中,社交网络信息在推荐系统中的重要性越来越明显,越来越多的研究者将社交网络中的信任关系引入推荐系统中。Massa等人在2004年首次提出将社交中的信任关系融入推荐算法中,用用户间的信任度替代传统相似度对用户空缺值进行预测评分[11],该方法对比传统协同过滤推荐算法准确性有很大提升。Ma等人在推荐系统中引入社交规则的概念,阐述了所提出的两种社交规则对推荐系统的贡献,实验证明基于社交规则的推荐可以有效提高推荐的准确性[12]。文献[7]融合用户社交信任度和评分相似性,提出了一个新矩阵填充的推荐方法,使预测评分准确度明显提升,改善了推荐过程中存在的稀疏性问题。文献[9,14]将高维用户评分矩阵映射到低维特征矩阵,融合用户的社交信息以及各自的隐性数据源进行推荐,实验结果证明该方法可以提高推荐准确度,但会造成部分信息丢失。

假设研究的推荐系统含有 m 个用户和 n 个商品,用户对商品的评分矩阵为 $\boldsymbol{R}=[\boldsymbol{R}_{u,i}]_{m\times n}$,如图1(a)。$\boldsymbol{R}_{u,i}\in[1,5]$ 表示用户 u 对商品 i 的评分值,5表示最喜欢,1表示最讨厌,评分值为空表示用户未对该商品评分。其中,$\boldsymbol{U}=\{u_1,u_2,\cdots,u_m\}$ 表示全部的用户集,$\boldsymbol{I}=\{i_1,i_2,\cdots,i_n\}$ 代表全部的商品集。如何有效得到未评分商品的预测值 $\widehat{\boldsymbol{R}_{u,i}}$,是个性化推荐至关重要的一步。最后计算真实评分和预测评分之间的差异最小值来评估推荐系统的推荐准确性,则上述问题变成求最优解问题,目标函数如式(1):

$$\min\sum_{u=1}^{m}\sum_{i=1}^{n}\left\|\widehat{\boldsymbol{R}_{u,i}}-\boldsymbol{R}_{u,i}\right\|^2 \tag{1}$$

通常情况若只分析用户评分矩阵来预测缺失评分,易导致评分预测不准确。通过增加额外的社交网络数据源辅助用户评分数据,来提高预测值的准确性[12]。这种方法的评分预测依据是:两个用户之间的偏好具有相似性或存在信任的社交关系,如果其中一个用户对某商品的评分较高,则可以认为另一用户对该商品的评分也较高。社交网络中用户间的信任关系可以用矩阵 $\boldsymbol{T}$ 表示,$\boldsymbol{T}=[\boldsymbol{T}_{u,v}]_{m\times m}$,其中 $\boldsymbol{T}_{u,v}\in[0,1]$ 表示用户间信任程度,如图1(b),用户 u_1 信任用户 u_3、u_4、u_5;用户之间的不信任关系用矩阵 $\boldsymbol{D}$ 表示,$\boldsymbol{D}=[\boldsymbol{D}_{u,v}]_{m\times m}$,$\boldsymbol{D}_{u,v}\in(0,1]$ 表示用户间的不信任程度,如图1(c),用户 u_1 不信任 u_2。本文研究的内容主要是引入社交网络数据源对用户评分数据中的空缺值进行填充,从而完成相关推荐。

2.2 传统矩阵分解模型

矩阵分解(matrix factorization,MF)模型被广泛应用在协同过滤推荐算法中,适用于对用户-商品评分矩阵数据进行分析,近似预测缺失数据[15]。其思想是将高维用户评分矩阵分解成为低维用户特征矩阵 $\boldsymbol{U}\in\mathbb{R}^{l\times m}$ 和商品特征矩阵 $\boldsymbol{I}\in\mathbb{R}^{l\times n}$,其中 $l\leqslant\min(m,n)$,分解后的用户特征只由几个少量的重要特征决定[16],

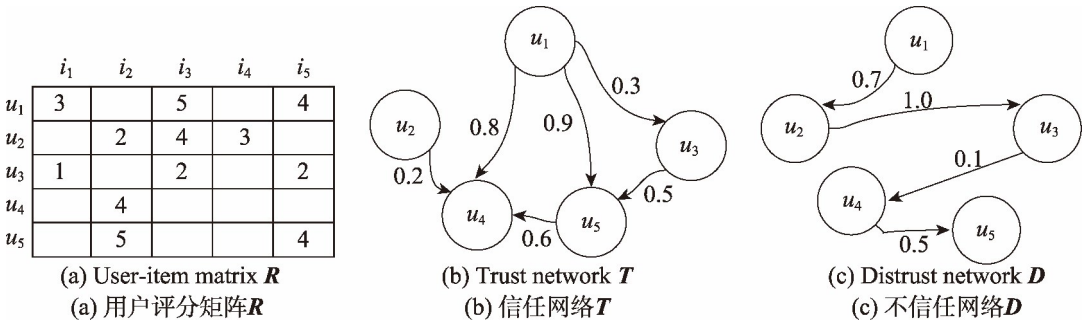

Fig.1 A sample of user-item matrix and users' relationship

图1 用户评分矩阵与用户关系举例

评分矩阵 $\boldsymbol{R}$ 可以用 $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{I}$ 近似替代，$\boldsymbol{U}^{\mathrm{T}}$ 为矩阵 $\boldsymbol{U}$ 的转置。为了方便研究，通常用函数 $f(x)=x/R_{\max}$ 将用户评分数据映射到 [0,1] 之间[1]，$R_{\max}$ 是用户评分的最大值。传统的基于矩阵分解模型的协同过滤方法利用简单的线性模型 $\boldsymbol{R}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{I}$ 近似拟合评分矩阵，容易造成预测评分过分偏离真实评分，使预测失真。本文引入非线性 logistic 函数 $g(x)=1/(1+\mathrm{e}^{-x})$，将预测评分值映射在 [0,1] 内。

为了避免过拟合现象，添加正则化约束项，求解最小代价函数 $\mathcal{L}$ 如式(1)时的用户特征矩阵 $\boldsymbol{U}$ 和商品特征矩阵 $\boldsymbol{I}$，则上述问题的目标函数如式(2)：

$$\mathcal{L}=\sum_{(u,i)\in\boldsymbol{R}}(\boldsymbol{R}_{u,i}-g(\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{I}_i))^2+\lambda_u\|\boldsymbol{U}\|_{\mathrm{F}}^2+\lambda_i\|\boldsymbol{I}\|_{\mathrm{F}}^2 \tag{2}$$

其中，$\|\cdot\|_{\mathrm{F}}^2$ 表示二阶范数；$\|\boldsymbol{A}\|_{\mathrm{F}}^2=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left|\boldsymbol{A}_{ij}\right|^2}$；$\lambda_u$、$\lambda_i$ 代表特征矩阵 $\boldsymbol{U}$、$\boldsymbol{I}$ 的正则化系数；$\lambda_u\|\boldsymbol{U}\|_{\mathrm{F}}^2$、$\lambda_i\|\boldsymbol{I}\|_{\mathrm{F}}^2$ 为正则化项。用户-商品预测评分的预测值则可以用式(3)计算：

$$\widehat{\boldsymbol{R}_{u,i}}=g(\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{V}_i)\cdot\boldsymbol{R}_{\max} \tag{3}$$

2.3 基于社交网络的协同过滤方法

传统基于矩阵分解模型的协同过滤推荐算法依据用户对商品的评分来预测空缺评分数据，忽视了用户、商品属性以及用户之间的社交关系，推荐预测并不一定准确[10]。目前社交关系被广泛应用于社交推荐系统中，社交推荐系统基于如下基本假设：如果用户 u 与 v 之间存在正相关社交信任关系，则认为用户 u、v 之间的兴趣偏好程度高于不相关的陌生人[1]，近几年的很多研究已经证明在社交网络中具有正相关社交关系的用户之间可以很好地传播这种正相关特性，通过利用这种正相关性能够有效地提高推荐准确率[7,17]。基于以上依据可以认为在实际社交推荐过程中，用户更容易接受与其具有正相关社交关系的用户的推荐。此处用户之间的信任关系可以由信任评分矩阵显性表示，如果数据集中用户之间的信任程度由显性的信任数据表示，信任数据范围是[0,1]，无评分则为空，如 Epinions 数据集；若没有显性信任数据时，信任度可以依据用户共同评分项等隐性信任数据构建，即用户之间的共同评分项的评分趋向越一致，则用户之间的信任程度越高，反之越低，如 UPS（user position similarity）方法[17]、用户间接可信度[11]等方法构建用户之间的信任度。本文主要基于已知社交信任数据下的协同过滤推荐算法进行研究。

社交网络关系一般存在以下几个特点：(1)用户信任关系具有传递机制，如果用户 u、v 之间，v、w 之间分别存在信任关系，则 v、w 之间被认为也存在信任关系，但是三者间的信任程度不同，$\boldsymbol{T}_{uw}\leqslant\min(\boldsymbol{T}_{uv},\boldsymbol{T}_{vu})$；(2)用户间社交网络信任关系不存在对称性，即使用户 u 信任用户 v，不一定用户 v 信任用户 u；(3)社交网络中的不信任关系不存在传递机制。

依据 2.2 节中的矩阵分解模型的分解方法，可以将用户间的信任关系矩阵 $\boldsymbol{T}\in\mathbb{R}^{m\times m}$ 映射成两个低维特征矩阵，即信任特征矩阵 $\boldsymbol{P}\in\mathbb{R}^{k\times m}$ 和被信任特征矩阵 $\boldsymbol{Q}\in\mathbb{R}^{k\times m}$，则信任矩阵 $\boldsymbol{T}$ 的信任值可通过 $\boldsymbol{T}=\boldsymbol{P}^{\mathrm{T}}\boldsymbol{Q}$ 线性组合近似估计，其中 m 表示用户数量，$k\leqslant m$ 表示特征数量。假如某个用户 u 是否信任用户 v 由 k 个特征因素决定，那么可以用一个 k 维向量 $\boldsymbol{P}_u=[p_1,p_2,\cdots,p_k]^{\mathrm{T}}$ 表示用户 u 的信任标准。用 $\boldsymbol{Q}_v=[q_1,q_2,\cdots,q_k]^{\mathrm{T}}$ 表示被信任用户 v 本身的特征，用户 u 对用户 v 的信任值 $\boldsymbol{T}_{uv}$ 可以用 $\boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{Q}_v$ 近似表示，但由于噪声的存在，这种分解并不准确，从而用代价函数 $\mathcal{L}$ 评估真实值与预测值之间的差异，代价函数 $\mathcal{L}$ 最小时对应求得的特征矩阵 $\boldsymbol{P}$、$\boldsymbol{Q}$ 即为最优解。代价函数为式(4)：

$$\mathcal{L}=\sum_{(u,v)\in\boldsymbol{T}}(\boldsymbol{T}_{u,v}-g(\boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{Q}_v))^2+\lambda_p\|\boldsymbol{P}\|_{\mathrm{F}}^2+\lambda_q\|\boldsymbol{Q}\|_{\mathrm{F}}^2 \tag{4}$$

其中，λ_p、λ_q 为正则化系数；$\lambda_p\|\boldsymbol{P}\|_{\mathrm{F}}^2$、$\lambda_q\|\boldsymbol{Q}\|_{\mathrm{F}}^2$ 是正则化项。预测到的社交信任数据通过非线性 logistic 函数 $g(x)=1/(1+\mathrm{e}^{-x})$ 映射到[0,1]之间。

通常用户之间不是完全相互独立的，而是具有一定相关性的。相关性度量方法通常有余弦相似性和 Pearson 相关性，本文采用 Pearson 相关性方法计算用户之间的相关程度，计算方法如式(5)：

$$sim_1(u,v)=\frac{\sum_I(r_{ui}-\bar{r}_u)\cdot(r_{vi}-\bar{r}_v)}{\sqrt{\sum_I(r_{ui}-\bar{r}_u)^2}\sqrt{\sum_I(r_{vi}-\bar{r}_v)^2}} \tag{5}$$

其中，$\boldsymbol{I}$ 表示用户 u、v 的共同评分商品集，$i\in\boldsymbol{I}$；r_{ui} 和 r_{vi} 分别表示用户 u 和 v 对项目 i 的评分值；$\bar{r}_u$、$\bar{r}_v$

分别表示用户 u 和用户 v 的评分均值。通常用户间的共同评分项的评分倾向越一致，则认为二者对项目的关注程度越一致，二者的相似性值也越高。定义用户评分偏好程度如式(6)：

$$P(u,v)=\left(\frac{|\boldsymbol{I}_u\cap\boldsymbol{I}_v|^{(\geqslant t)}}{|\boldsymbol{I}_u|^{(\geqslant t)}}+\frac{|\boldsymbol{I}_u\cap\boldsymbol{I}_v|^{(<t)}}{|\boldsymbol{I}_u|^{(<t)}}\right)\times\left(\frac{|\boldsymbol{I}_u\cap\boldsymbol{I}_v|^{(\geqslant t)}}{|\boldsymbol{I}_v|^{(\geqslant t)}}+\frac{|\boldsymbol{I}_u\cap\boldsymbol{I}_v|^{(<t)}}{|\boldsymbol{I}_v|^{(<t)}}\right) \tag{6}$$

其中，t 是区分评价好坏的阈值，超过阈值记为积极评分，否则为消极评分；$\boldsymbol{I}_u$、$\boldsymbol{I}_v$ 分别表示用户 u、v 的评分商品集合。将用户偏好程度融入相似度计算中，即基于偏好程度的相似性度量方法，计算如式(7)：

$$sim(u,v)=P(u,v)\times sim_1(u,v) \tag{7}$$

其中，$P(u,v)$ 表示用户间评分偏好程度；$sim_1(u,v)$ 为用户间Pearson相关系数。社交信任关系的代价函数为式(8)：

$$\mathcal{L}=\sum_{(u,v)\in\boldsymbol{T}}s(u,v)(\boldsymbol{T}_{u,v}-g(\boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{Q}_v))^2+\lambda_p\|\boldsymbol{P}\|_{\mathrm{F}}^2+\lambda_q\|\boldsymbol{Q}\|_{\mathrm{F}}^2 \tag{8}$$

通过对基于社交网络的研究，用户对商品预测评分可用社交特征和商品特征修正，如式(9)：

$$\widehat{\boldsymbol{R}_{u,i}}=g(\beta\boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{V}_i+\gamma\boldsymbol{Q}_u^{\mathrm{T}}\boldsymbol{V}_i+\theta sim(u,v)) \tag{9}$$

其中，$\beta,\gamma,\theta\in(0,1)$，且 $\beta+\gamma+\theta=1$，是调控信任特征向量、被信任特征向量和相似性的贡献率参数。

在社交网络关系中，如果两个用户 u、v 之间不存在信任关系，而是怀疑关系，那么认为两个用户的兴趣偏好存在较大偏差，对同一商品的评分差别可能较大。在本算法中，可以计算用户特征空间的欧氏距离描述用户之间兴趣偏好的差异。

$$\sum_{u=1}^{m}\sum_{w\in\boldsymbol{D}(u)}\|\boldsymbol{P}_u-\boldsymbol{P}_w\|^2+\sum_{u=1}^{m}\sum_{w\in\boldsymbol{D}(u)}\|\boldsymbol{Q}_u-\boldsymbol{Q}_w\|^2 \tag{10}$$

基于以上描述，最小代价函数修正成式(11)：

$$\begin{aligned}\mathcal{L}=&\sum_{(u,i)\in\boldsymbol{R}}\left(g\left(\beta\boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{V}_i+(1-\beta)\boldsymbol{Q}_u^{\mathrm{T}}\boldsymbol{V}_i\right)-\boldsymbol{R}_{u,i}\right)^2+\\&\sum_{(u,v)\in\boldsymbol{T}}s(u,v)\left(g(\boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{Q}_v)-\boldsymbol{T}_{u,v}\right)^2+\lambda_p\|\boldsymbol{P}\|_{\mathrm{F}}^2+\\&\lambda_q\|\boldsymbol{Q}\|_{\mathrm{F}}^2+\lambda_v\|\boldsymbol{V}\|_{\mathrm{F}}^2-\lambda_1\sum_{u=1}^{m}\sum_{w\in\boldsymbol{D}(u)}\|\boldsymbol{P}_u-\boldsymbol{P}_v\|^2-\\&\lambda_2\sum_{u=1}^{m}\sum_{u\in\boldsymbol{D}(u)}\|\boldsymbol{Q}_u-\boldsymbol{Q}_v\|^2\end{aligned} \tag{11}$$

其中，λ_1、λ_2 是社交网络差异特征矩阵正则化系数；λ_p、λ_q 和 λ_v 是矩阵 $\boldsymbol{P}$、$\boldsymbol{Q}$ 和 $\boldsymbol{V}$ 正则化系数，以防止过拟合现象。

根据上述模型描述，使损失函数最小时 $\boldsymbol{P}$、$\boldsymbol{Q}$ 和 $\boldsymbol{V}$ 的值即为最优解。为了求代价函数的最优解，算法采用随机梯度下降法求代价函数关于 $\boldsymbol{P}_u$、$\boldsymbol{Q}_u$ 和 $\boldsymbol{V}_i$ 的偏导数，如式(12)、(13)、(14)即为梯度方向。

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{P}_u}=2\Bigg[&\sum_{i\in\boldsymbol{R}(u)}\left(\beta\boldsymbol{V}_i g'\left(\beta\boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{V}_i+(1-\beta)\boldsymbol{Q}_u^{\mathrm{T}}\boldsymbol{V}_i\right)\right)\times\\&\left(g\left(\beta\boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{V}_i+(1-\beta)\boldsymbol{Q}_u^{\mathrm{T}}\boldsymbol{V}_i\right)-\boldsymbol{R}_{u,i}\right)+\\&\sum_{v\in\boldsymbol{T}(u)}\boldsymbol{Q}_v sim(u,v)g'(\boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{Q}_v)\left(g(\boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{Q}_v)-\boldsymbol{T}_{u,v}\right)+\\&\sum_{v\in\boldsymbol{D}(u)}(\lambda_p\boldsymbol{P}-\lambda_1|\boldsymbol{P}_u-\boldsymbol{P}_v|)\Bigg]\end{aligned} \tag{12}$$

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{Q}_u}=2\Bigg[&\sum_{i\in\boldsymbol{R}(u)}\left((1-\beta)\boldsymbol{V}_i g'\left(\beta\boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{V}_i+(1-\beta)\boldsymbol{Q}_u^{\mathrm{T}}\boldsymbol{V}_i\right)\right)\times\\&\left(g\left(\beta\boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{V}_i+(1-\beta)\boldsymbol{Q}_u^{\mathrm{T}}\boldsymbol{V}_i\right)-\boldsymbol{R}_{u,i}\right)+\\&\sum_{v\in\boldsymbol{T}(u)}\boldsymbol{P}_u^{\mathrm{T}}sim(u,v)g'(\boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{Q}_v)\left(g(\boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{Q}_v)-\boldsymbol{T}_{u,v}\right)+\\&\sum_{v\in\boldsymbol{D}(u)}(\lambda_q\boldsymbol{Q}-\lambda_2|\boldsymbol{Q}_u-\boldsymbol{Q}_v|)\Bigg]\end{aligned} \tag{13}$$

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{V}_i}=2\Bigg[&\sum_{u\in\boldsymbol{R}(i)}\left(\beta\boldsymbol{P}_u^{\mathrm{T}}+(1-\beta)\boldsymbol{Q}_u^{\mathrm{T}}\right)\times\\&\left(g\left(\beta\boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{V}_i+(1-\beta)\boldsymbol{Q}_u^{\mathrm{T}}\boldsymbol{V}_i\right)-\boldsymbol{R}_{u,i}\right)\times\\&g'\left(\beta\boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{V}_i+(1-\beta)\boldsymbol{Q}_u^{\mathrm{T}}\boldsymbol{V}_i\right)+\lambda_v\boldsymbol{V}\Bigg]\end{aligned} \tag{14}$$

其中，$\boldsymbol{R}(u)$ 表示用户 u 评分过的商品集；$\boldsymbol{T}(u)$ 表示用户 u 信任的用户集；$\boldsymbol{D}(u)$ 表示用户 u 怀疑的用户集；$\boldsymbol{R}(i)$ 表示给商品 i 评分过的用户集。沿梯度的负方向不断迭代更新 $\boldsymbol{P}$、$\boldsymbol{Q}$ 和 $\boldsymbol{V}$ 直至收敛，即可认为得到最优解。

$$\begin{cases}\boldsymbol{P}_{t+1}=\boldsymbol{P}_t+\alpha\dfrac{\partial\mathcal{L}}{\partial\boldsymbol{P}_u}\\\boldsymbol{Q}_{t+1}=\boldsymbol{Q}_t+\alpha\dfrac{\partial\mathcal{L}}{\partial\boldsymbol{Q}_u}\\\boldsymbol{V}_{t+1}=\boldsymbol{V}_t+\alpha\dfrac{\partial\mathcal{L}}{\partial\boldsymbol{V}_i}\end{cases} \tag{15}$$

其中，α 表示学习速率；$\boldsymbol{P}_t$、$\boldsymbol{Q}_t$ 和 $\boldsymbol{V}_t$ 用随机数进行初始化，t 表示迭代次数，$t\in[1,\infty)$。通过不断迭代更新，$\boldsymbol{P}$、$\boldsymbol{Q}$ 和 $\boldsymbol{V}$ 的值会趋于稳定，稳定后的值即为最优解。高维评分矩阵、社交信任矩阵通过分解变成

低维用户信任特征矩阵、被信任特征矩阵、商品特征矩阵，利用式(9)对用户评分矩阵进行预测评估，稀疏矩阵变为密集矩阵，选择目标用户中评分最高的 N 个商品推荐给用户，即完成 Top-N 推荐。

2.4 复杂度分析

评价一个算法性能，主要计算算法的时间复杂度，即语句总的执行次数[18]。在本研究中，用户之间的相似性以及评分矩阵、社交矩阵分解等计算过程都是离线条件下完成的，因此本文基于社交网络的协同过滤推荐算法的计算复杂度主要受代价函数和梯度下降特征矩阵维数变量的影响。其中代价函数的计算复杂度是 $O(k(|\boldsymbol{R}|+|\boldsymbol{T}|+|\boldsymbol{D}|))$，$k$ 表示隐性特征维数，$|\boldsymbol{R}|$、$|\boldsymbol{T}|$ 和 $|\boldsymbol{D}|$ 表示存在的非空用户评分数量、用户信任连接数量和用户不信任连接数量。梯度 $\partial\mathcal{L}/\partial\boldsymbol{P}$、$\partial\mathcal{L}/\partial\boldsymbol{Q}$ 和 $\partial\mathcal{L}/\partial\boldsymbol{V}$ 的复杂度分别为 $O(k(|\boldsymbol{R}|+|\boldsymbol{T}|+|\boldsymbol{D}|))$、$O(k(|\boldsymbol{R}|+|\boldsymbol{T}|+|\boldsymbol{D}|))$ 和 $O(k|\boldsymbol{R}|)$，因此本算法的复杂度为 $O(k(|\boldsymbol{R}|+|\boldsymbol{T}|+|\boldsymbol{D}|))$，即时间复杂度与观察到的用户评分数、信任连接数与不信任连接数的和成线性关系，因此适合用于大规模数据集的推荐。

3 实验及结果分析

为了验证本文算法的可靠性和有效性，利用 Epinions 真实数据集，采用五折交叉验证方法，对本文算法与其他现有的社交推荐算法进行实验对比，统计分析不同方法在推荐准确性方面的平均绝对误差(mean absolute error，MAE)和均方根误差(root mean square error，RMSE)指标。同时对推荐系统在 Top-N 推荐时每个算法在不同的推荐数量 N 的情况下的查准率、查全率和 F1-Measure 值进行统计并分析，以此验证本文算法的可靠性。

3.1 数据集描述

本文选择 Epinions 公开数据集作为研究的真实数据集，它由 Massa 在 http://www.epinions.com 网站收集整理所得[19]。该数据集是一极度稀疏的数据集，稀疏度用空评分数据数量/(用户数量×商品数量)表示，为 99.991 35%，包含 49 290 个用户对 139 738 个不同商品的评分数据，用户评分数据为 1~5 内的整数，1 表示最差，5 表示最好；同时也包含 664 824 条用户间的社交数据，其中 487 181 条记录表示用户间的关系是积极的，认为是信任数据，信任值为 1，其余不存在信任关系即为 0。经对数据统计分析可得每类评分的贡献情况，评分数据中评分为 5 的记录数占总数的 45%，评分为 4 的占 29%，评分为 3 的占 11%，评分为 2 的占 8%，评分为 1 的占 7%，所有评分的平均值约为 3.9，接近一半的用户给商品的评分为最高值 5。如果用户的评论数量低于 5，则被认为是“冷启动”用户，该类用户数量为 26 037，那么数据集中有超过一半的用户属于冷启动用户。

3.2 实验设置

3.2.1 实验方法

K 折交叉验证：为验证本文算法的有效性和真实性，本实验采用 5 折交叉验证的方法[7]将所研究的数据集平分成 5 份，每次实验随机选取数据集中的 1 组作为测试数据，剩下的 4 组数据集作为训练数据，每个实验进行 5 次，实验结果为 5 次实验的平均值，并进行比较分析。

3.2.2 实验评估指标

(1)为了验证推荐算法的准确性，常用的推荐性能评估指标主要包括平均绝对误差 MAE 和均方根误差 RMSE[20]，用来评估推荐结果的误差分布。MAE 计算的是所有测试用户对测试项目的预测评分和实际评分的平均误差大小，RMSE 计算真实评分与预测评分值的均方根误差，如式(16)、(17)：

$$MAE=\frac{1}{N}\sum_{u,i\in\boldsymbol{T}_u}|\widehat{\boldsymbol{R}_{u,i}}-\boldsymbol{R}_{u,i}| \tag{16}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{u,i\in\boldsymbol{T}_u}(\widehat{\boldsymbol{R}_{u,i}}-\boldsymbol{R}_{u,i})^2} \tag{17}$$

其中，$\boldsymbol{T}_u$ 表示测试集中的用户数据集合；N 表示测试集中商品数量；$\widehat{\boldsymbol{R}_{u,i}}$ 表示用户 u 对商品 i 的预测评分；$\boldsymbol{R}_{u,i}$ 表示真实评分。

(2)另一种预测推荐系统质量的方法是计算推荐的正确率(Precision)、召回率(Recall)和 F1-Measure 值[16]。准确率和召回率是评估推荐系统常用的两个度量指标。其中准确度可以衡量推荐系统的查准率，即推荐结果满足用户喜好的概率；召回率衡量的是推荐系统的查全率，即所有推荐结果与用户喜好

相关的概率。两者取值在0和1之间,数值越接近1,查准率或查全率就越高。*F*1-Measure是结合Precision和Recall两者给出的综合评价指标。定义如下:

$$Precision=\frac{\sum_{u\in T_u}|L_u\cap B_u|}{\sum_{u\in T_u}|L_u|} \tag{18}$$

$$Recall=\frac{\sum_{u\in T_u}|L_u\cap B_u|}{\sum_{u\in T_u}|B_u|} \tag{19}$$

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{20}$$

其中,L_u表示用户u由训练数据集得到的推荐商品集合;B_u表示在测试数据集给出正反馈评分的商品集合;T_u表示测试集中的用户数据集合。

3.2.3 方法比较

为了评估本文算法的性能,采用以下几种方法进行对比验证。

(1) SocialMF(matrix factorization in social networks):该算法是一种基于信任传递机制的社交推荐算法[21]。

(2)TDMF(matrix factorization with trust and distrust information):该算法利用社交信任和不信任信息进行推荐[22]。

(3)TDRec(matrix factorization with explicit trust and distrust side information for improved social recommendation):该算法通过显式信任和非信任数据进行社交推荐[23]。

(4)TDSRec(similarity social recommendation with trust and distrust information):该算法是本文算法模型,在考虑社交网络的同时,融合基于用户评分偏好的相似性,共同对用户评分矩阵中的数据值进行评分预测。

为了便于比较,将各个算法的参数设置为表1,参数的设置来自于参考文献或者本研究模型。在矩阵分解过程中,特征矩阵的维数设置为5和10,分别统计特征维数的误差值。

3.3 实验结果及分析

在表2、表3中,分别统计了所有用户和冷启动用户(用户对商品评分个数少于5)在矩阵特征数k为5和10时,本文算法与现有算法的平均绝对误差(MAE)和均方根误差(RMSE),并计算得到本文算法比现有算法的提升率,观察到本文算法的MAE和RMAE小于已有算法,实验证明了本文算法提升了推荐准确度。

由图2、图3可以得出,在特征矩阵维数$k=10$时,随着迭代次数的增加,MAE和RMSE逐渐趋于稳定。结果显示,本文算法的MAE和RMSE低于被比较对象,性能优于被比较算法。

Table 1 Parameter settings

表1 参数设置

方法	参数选择
SocialMF	$\lambda_u=\lambda_v=0.001$, $\lambda_T=1$
TDMF	$\lambda_u=\lambda_v=0.001$, $\lambda_S=10$
TDRec	$\lambda_p=\lambda_w=\lambda_v=\lambda_c=0.001$, $\lambda=\lambda_t=0.5$, $\beta=0.4$, $\alpha=0.001$
TDSRec	$\lambda_p=\lambda_q=\lambda_v=0.001$, $\lambda_1=\lambda_2=0.000\,1$, $\beta=0.3$, $\gamma=0.5$, $\theta=0.2$, $\alpha=0.001$

Table 2 MAE and RMSE results of different algorithms on all users

表2 所有用户在不同算法得到的MAE、RMSE值

维数k	评估方法	MAE	Improve/%	RMSE	Improve/%
5	SocialMF	0.422 2	6.04	0.623 2	3.05
	TDMF	0.413 5	4.07	0.619 8	2.52
	TDRec	0.396 9	0.06	0.615 3	1.81
	TDSRec	0.396 7	—	0.604 2	—
10	SocialMF	0.424 5	8.72	0.641 2	7.30
	TDMF	0.405 3	4.38	0.610 0	2.56
	TDRec	0.396 9	2.36	0.579 9	0.54
	TDSRec	0.387 5	—	0.594 4	—

Table 3 MAE and RMSE results of different algorithms on cold-start users

表3 冷启动用户在不同算法得到的MAE、RMSE值

维数k	评估方法	MAE	Improve/%	RMSE	Improve/%
5	SocialMF	0.954 6	15.23	1.212 3	23.73
	TDMF	0.908 4	10.92	1.191 0	22.37
	TDRec	0.848 8	4.67	1.009 7	8.43
	TDSRec	0.809 2	—	0.924 6	—
10	SocialMF	1.035 4	24.05	1.205 4	28.00
	TDMF	0.907 3	13.32	1.186 8	26.87
	TDRec	0.806 6	2.51	0.921 7	5.84
	TDSRec	0.786 4	—	0.867 9	—

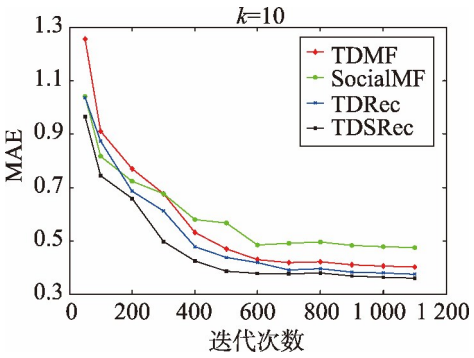

Fig.2 MAE at different iterations on all users

图2 所有用户不同迭代次数时的MAE

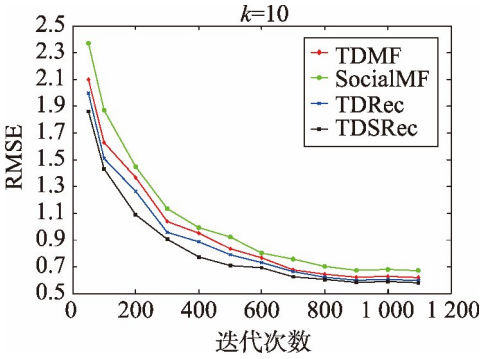

Fig.3 RMSE at different iterations on all users

图3 所有用户不同迭代次数时的RMAE

通过实验对比分析本文算法与已有算法，证明本文提出的协同过滤推荐算法在不同的推荐数量 N 下的查准率、查全率和综合指标 $F1$-Measure 都优于被比较算法。同时由图4、图5可知算法的查全率和查准率呈相反趋势增长，由图6可知综合两者特性的 $F1$-Measure 参数也有增长趋势。

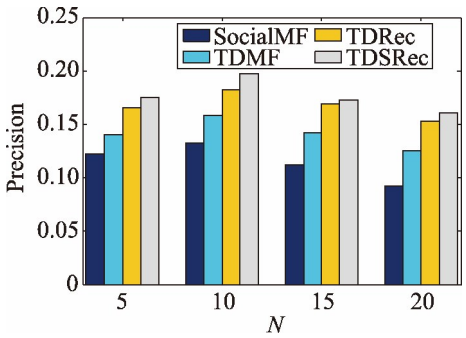

Fig.4 Precision

图4 查准率

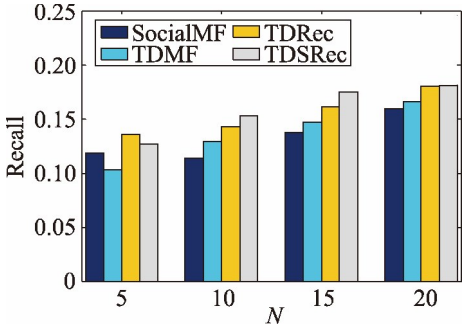

Fig.5 Recall

图5 查全率

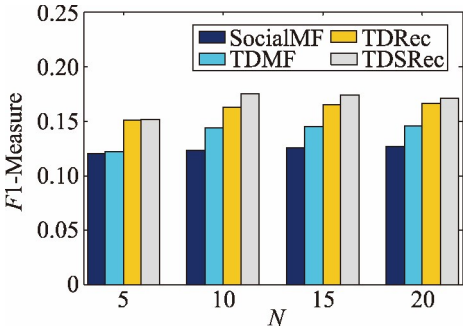

Fig.6 $F1$-Measure

图6 $F1$-Measure

4 结束语

协同过滤推荐技术是推荐系统中应用最广和最多的推荐技术，在过去的十几年内，已有很多基于社交网络特征的协同过滤推荐方法，数据稀疏性、推荐准确性方面在一定程度上都得到提升[9]，但是推荐系统的固有缺陷仍然存在，导致推荐质量较差，很难满足用户个性化需求。本文针对推荐系统中存在的固有的数据稀疏和冷启动问题，创新性地提出一种融合社交网络特征的协同过滤推荐算法。本文算法主要基于传统矩阵分解模型，分解用户评分矩阵为两个低维的用户特征矩阵和商品特征矩阵的同时，分解社交网络矩阵数据为信任特征矩阵和被信任特征矩阵；在求解分解矩阵过程中通过融合用户评分偏好程度计算用户相似性使代价函数最小，如式(11)，使用梯度下降法不断迭代更新得到对应分解后的特征矩阵，如式(12)、(13)、(14)；利用信任特征矩阵、被信任特征矩阵、商品特征矩阵预测用户对商品的评分值，如式(9)。以此缓解矩阵稀疏性，提高推荐结果的准确性和有效性。

为了验证本文算法的可靠性,基于Epinions公开数据集进行实验分析,采用五折交叉验证方法,分析对比了本文算法较现有的社交网络推荐算法如TDMF、SocialMF、TDRec等算法的先进性和优越性。实验结果表明,本文算法在使用相同数据集的情况下与其他算法相比,平均绝对误差、均方根误差都有所下降,有效地提高了推荐算法的性能。在现实推荐中,本文算法可以有效用于大规模用户商品集的推荐,有效缓解数据稀疏性,提高预测准确度和推荐准确度,改善推荐质量。

本文提出的融合社交网络的协同过滤推荐算法,虽然在一定程度上缓解了评分数据稀疏性对推荐结果的影响,并对于大量数据的推荐运算具有一定效果,但是不足以支撑超大规模数据推荐,此时可以在本文基础上考虑超大规模推荐的并行算法来克服此局限性。同时本文算法未考虑时间差异对推荐产生的影响,由于用户对商品的兴趣爱好时间长短不一,同一用户对同一商品的评分在不同时间点存在差异,由此可以发现可疑用户,以此提高推荐准确率。以上两点可以作为下一步的研究方向。

References:

[1] Bai Tiansheng, Yang Bo, Li Fei. TDRec: enhancing social recommendation using both trust and distrust information [C]//Proceedings of the 2nd European Network Intelligence Conference, Karlskrona, Sep 21-22, 2015. Washington: IEEE Computer Society, 2015: 60-66.

[2] Breese J S, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[C]//Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, Madison, Jul 24-26, 1998. San Francisco: Morgan Kaufmann Publishers Inc, 1998: 43-52.

[3] Wang Licai, Meng Xiangwu, Zhang Yujie. Context-aware recommender systems[J]. Journal of Software, 2012, 23(1): 1-20.

[4] Zhang Yanmei, Wang Lu, Cao Huaihu, et al. Recommendation algorithm based on user-interest-item tripartite graph [J]. Pattern Recognition and Artificial Intelligence, 2015, 28 (10): 913-921.

[5] Lu Zhongqi, Dou Zhicheng, Lian Jianxun, et al. Content-based collaborative filtering for news topic recommendation [C]//Proceedings of the 29th Conference on Artificial Intelligence, Austin, Jan 25-30, 2015. Menlo Park: AAAI, 2015: 217-223.

[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.

[7] Wang Meiling, Ma Jun. A novel recommendation approach based on users' weighted trust relations and the rating similarities[J]. Soft Computing, 2015, 20(10): 3981-3990.

[8] Wang Shengsheng, Zhao Haiyan, Chen Qingkui, et al. Latent factor model for personalized recommendation[J]. Journal of Chinese Computer System, 2016, 37(5): 881-889.

[9] Felfernig A, Ninaus G, Grabner H, et al. An overview of recommender systems in requirements engineering[M]//Ma-alej W, Thurimella A K. Managing Requirements Knowledge. Berlin, Heidelberg: Springer, 2013: 315-332.

[10] Pan Junchi, Zhang Xingming, Wang Xin. Improved singular value decomposition recommender algorithm based on user reliability[J]. Journal of Chinese Computer System, 2016, 37(10): 2171-2176.

[11] Massa P, Bhattacharjee B. Using trust in recommender systems: an experimental analysis[C]//LNCS 2995: Proceedings of the 2nd International Conference on Trust Management, Oxford, Mar 29-Apr 1, 2004. Berlin, Heidelberg: Springer, 2004: 221-235.

[12] Ma Hao, Zhou Dengyong, Liu Chao, et al. Recommender systems with social regularization[C]//Proceedings of the 4th International Conference on Web Search and Web Data Mining, Hong Kong, China, Feb 9-12, 2011. New York: ACM, 2011: 287-296.

[13] Chen Wei. Multi-collaborative filtering trust network for online recommendation[J]. Information Systems Frontiers, 2015, 15(4): 533-551.

[14] Ma Hao, Lyu M R, King I. Learning to recommend with trust and distrust relationships[C]//Proceedings of the 2009 Conference on Recommender Systems, New York, Oct 23-25, 2009. New York: ACM, 2009: 189-196.

[15] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30-37.

[16] Wang Peiying. Community discovery and collaborative filtering recommendation in social networks[D]. Beijing: Beijing Jiaotong University, 2016.

[17] Du Yongping, Du Xiaoyan, Huang Liang. Improve the collaborative filtering recommender system performance by trust network construction[J]. Chinese Journal of Electronics, 2016, 25(3): 418-423.

[18] Fouss F, Saerens M. Evaluating performance of recommender systems: an experimental comparison[C]//Proceedings of the

2008 International Conference on Web Intelligence and Intelligent Agent Technology, Los Alamitos, Dec 9-12, 2008. Washington: IEEE Computer Society, 2008: 735-738.
[19] Rennie J D M, Srebro N. Fast maximum margin matrix factorization for collaborative prediction[C]//Proceedings of the 22nd International Conference on Machine Learning, Bonn, Aug 7-11, 2005. New York: ACM, 2005: 713-719.
[20] Mao Yiyu, Liu Jianxun, Hu Rong, et al. Sigmoid function-based Web service collaborative filtering recommendation algorithm[J]. Journal of Frontiers of Computer Science and Technology, 2017, 11(2): 314-322.
[21] Jamali M, Ester M. A matrix factorization technique with trust propagation for recommendation in social networks [C]//Proceedings of the 4th ACM Conference on Recommender Systems, Barcelona, Sep 26-30, 2010. New York: ACM, 2010: 135-142.
[22] Forsati R, Mahdavi M, Shamsfard M, et al. Matrix factorization with explicit trust and distrust side information for improved social recommendation[J]. ACM Transactions on Information Systems, 2014, 32(4): 17.
[23] Park C, Kim D, Oh J, et al. Improving top-K recommendation with truster and trustee relationship in user trust network[J]. Information Sciences, 2016, 374(C): 100-114.

附中文参考文献：

[3] 王立才, 孟祥武, 张玉洁. 上下文感知推荐系统[J]. 软件学报, 2012, 23(1): 1-20.
[4] 张艳梅, 王璐, 曹怀虎, 等. 基于用户-兴趣-项目三部图的推荐算法[J]. 模式识别与人工智能, 2015, 28(10): 913-921.
[6] 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.
[8] 王升升, 赵海燕, 陈庆奎, 等. 个性化推荐中的隐语义模型[J]. 小型微型计算机系统, 2016, 37(5): 881-889.
[10] 潘骏驰, 张兴明, 汪欣. 融合用户可信度的改进奇异值分解推荐算法[J]. 小型微型计算机系统, 2016, 37(10): 2171-2176.
[16] 王培英. 社会网络中的社区发现及协同过滤推荐技术研究[D]. 北京: 北京交通大学, 2016.
[20] 毛宜钰, 刘建勋, 胡蓉, 等. 采用Sigmoid函数的Web服务协同过滤推荐算法[J]. 计算机科学与探索, 2017, 11(2): 314-322.

GUO Ningning was born in 1992. She is an M.S. candidate at Tianjin University. Her research interests include recommendation system and data mining, etc.
郭宁宁(1992—),女,山东聊城人,天津大学电子信息工程学院宽带无线通信与3D成像研究所硕士研究生,主要研究领域为推荐系统,数据挖掘等。

WANG Baoliang was born in 1971. He received the Ph.D. degree from Tianjin University in 2010. Now he is a senior engineer and M.S. supervisor at Tianjin University. His research interests include data mining, mobile internet and image processing, etc.
王宝亮(1971—),男,山东潍坊人,2010年于天津大学获得博士学位,现为天津大学高级工程师、硕士生导师,主要研究领域为数据挖掘,移动互联,图像处理等。

HOU Yonghong was born in 1968. He received the Ph.D. degree in communication and information system from Tianjin University in 2009. Now he is an associate professor and Ph.D. supervisor at Tianjin University. His research interests include computer vision, artificial intelligence and multimedia signal processing, etc.
侯永宏(1968—),男,山西太原人,2009年于天津大学获得博士学位,现为天津大学电子信息工程学院副教授、博士生导师,主要研究领域为计算机视觉,人工智能,多媒体信号处理等。

CHANG Peng was born in 1980. He received the Ph.D. degree from Tianjin University in 2010. Now he is a research assistant at Tianjin University. His research interests include data mining, text mining and information retrieval, etc.
常鹏(1980—),男,山西太原人,2010年于天津大学获得博士学位,现为天津大学助理研究员,主要研究领域为数据挖掘,文本挖掘,信息检索等。