软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(12):3747−3763 [doi: 10.13328/j.cnki.jos.005322]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于信任机制下概率矩阵分解的用户评分预测*

杜东舫[1,2], 徐 童[1,2], 鲁亚男[1,2], 管 楚[1,2], 刘 淇[1,2], 陈恩红[1,2]

[1](大数据分析与应用安徽省重点实验室(中国科学技术大学),安徽 合肥 230027)

[2](中国科学技术大学 计算机科学与技术学院,安徽 合肥 230027)

通讯作者: 徐童, E-mail: tongxu@ustc.edu.cn

摘 要: 互联网的蓬勃发展,在为用户提供便利的同时,其海量信息也为用户选择造成了困难,基于用户理解的信息推荐服务正成为应时之需.相较于面向单个用户信息的传统推荐技术,基于社交信息的推荐技术通过引入影响力建模,可以更真实地还原用户属性及行为.然而,已有的社交推荐技术往往停留于对用户影响的笼统归纳,并没有对其内在机制进行清晰分类和量化.针对这一问题,通过对用户评分行为中的信任关系进行分析,着重研究了信任用户间接影响用户偏好和直接影响用户评分两种不同机制,进而提出了基于用户间信任关系融合建模的概率矩阵分解模型 TPMF,从而实现对上述两种机制的有效融合.在此基础之上,针对不同用户受两种机制影响权重不同的问题,通过借助评分相关性对用户进行聚类并映射到相应权重,实现了用户模型参数的个性化选择.公开数据集的多项实验结果表明:提出的 TPMF 及其衍生算法在各项指标上优于现有代表性算法,验证了所提出的影响机制及技术框架的有效性.

关键词: 聚类分析;概率矩阵分解;推荐系统;信任关系

中图法分类号: TP181

中文引用格式: 杜东舫,徐童,鲁亚男,管楚,刘淇,陈恩红.基于信任机制下概率矩阵分解的用户评分预测.软件学报,2018, 29(12):3747−3763. http://www.jos.org.cn/1000-9825/5322.htm

英文引用格式: Du DF, Xu T, Lu YN, Guan C, Liu Q, Chen EH. User rating prediction based on trust-driven probabilistic matrix factorization. Ruan Jian Xue Bao/Journal of Software, 2018,29(12):3747−3763 (in Chinese). http://www.jos.org.cn/1000-9825/5322. htm

User Rating Prediction Based on Trust-Driven Probabilistic Matrix Factorization

DU Dong-Fang[1,2], XU Tong[1,2], LU Ya-Nan[1,2], GUAN Chu[1,2], LIU Qi[1,2], CHEN En-Hong[1,2]

[1](Anhui Province Key Laboratory of Big Data Analysis and Application (University of Science and Technology of China), Hefei 230027, China)

[2](School of Computer Science and Technology, University of Science and Technology of China), Hefei 230027, China)

Abstract: The development of Internet has brought convenience to the public, but also troubles users in making choices among enormous data. Thus, recommender systems based on user understanding are urgently in need. Different from the traditional techniques that usually focus on individual users, the social-based recommender systems perform better with integrating social influence modeling to achieve more accurate user profiling. However, current works usually generalize influence in simple mode, while deep discussions on intrinsic mechanism have been largely ignored. To solve this problem, this paper studies the social influence within users who affects both rating and user attributes, and then proposes a novel trust-driven PMF (TPMF) algorithm to merge these two mechanisms. Furthermore, to

* 基金项目: 国家杰出青年科学基金(61325010); 国家自然科学基金(U1605251, 61703386, 61403358); 安徽省自然科学基金(1708085QF 140); 中央高校基本科研业务费专项资金(WK2150110006)

Foundation item: National Natural Science Funds for Distinguished Young Scholar (61325010); National Natural Science Foundation of China (U1605251, 61703386, 61403358); Anhui Natural Science Foundation (1708085QF140); Fundamental Research Funds for the Central Universities (WK2150110006)

收稿时间: 2016-09-08; 修改时间: 2016-11-29, 2016-12-20; 采用时间: 2017-06-15

deal with the task that different user should have personalized parameters, the study clusters users according to rating correlation and then maps them to corresponding weights, thereby achieving the personalized selection of users' model parameters. Comprehensive experiments on open data sets validate that TPMF and its derivation algorithm can effectively predict users' rating compared with several state of the art baselines, which demonstrates the capability of the presented influence mechanism and technical framework.

Key words: clustering analysis; probabilistic matrix factorization; recommendation system; trust relationship

随着互联网技术的飞速发展,网络信息呈现指数级增长的态势.一方面,传统的传播模式与商业模式正在发生深刻变革,数字化、在线化程度不断加深;另一方面,社交媒体、C2C 等由个人提供的信息服务的繁荣,也推动了数据来源和种类的进一步丰富.统计显示,2014 年,Amazon 在线图书种类已超过 340 万,平均每小时新增 12 本以上.而在 Amazon 的应用商店,上架的应用种类已达 24 万,媒体文件更是超过了 2 700 万件.这在为用户带来更加丰富多样的选择同时,其海量数据的特性也加剧了用户获取所需信息或商品的困难[1].因此,建立在准确理解用户行为和意图基础之上,实现信息与服务有效筛选的智能推荐系统技术[2−4],成为当下的研究热点和应时之需.

一般而言,传统推荐技术主要包括基于内容的推荐技术[5]和协同过滤推荐技术[6,7]等.然而,这些推荐技术往往着眼于单个用户的行为记录分析,受用户动机和数据稀疏性干扰较强,其效果较为有限.社交网络和社交媒体服务的发展,催生了用户之间普遍关联和信息传递的新时代,使得获取用户之间的关系成为可能,也为推荐技术发展提供了新的契机.结合社交信息的推荐算法融入了人与人之间的关联及其引发的相互影响,从而使得建模更加真实和完善,推荐性能得以进一步提升[8−10].

基于社交的推荐技术往往具有以下基本假设:存在社交关系的用户在选择或倾向上,往往基于其相互信任而体现出一定的相似性,且信任关系越强,相似性越强[11].在真实的社交评论网站 Epinions 上,用户可以添加信任好友到自己的信任列表.图 1 是简单的社交信任网络图,单向箭头表示用户的信任关系,对应数字表示用户的信任权重.从图 1 中可以看到:用户 v_4 对用户 v_5 的信任权重为 0.9,体现出很强的信任关系.因此按照通常的基本假设,v_4 在面临选择时,很可能与 v_5 体现出很强的趋同性.然而,已有的社交推荐技术虽然都依赖于社交趋同性的基本假设,但对于趋同性的成因仍停留在简单的笼统概括上,对于其内在的因素缺乏明晰的区分和量化.例如,多数现有工作往往单纯认为社交趋同性的原因是好友之间具有相似的属性,进而由相似的属性导致了相似的选择.然而最新的研究表明,好友关系并非直接影响用户属性,而是直接对用户决策产生影响[12],甚至引发社交学习行为[13].显然,综合考虑不同成因下的社交影响融合建模,才能更有效地提升社交推荐技术的效果.

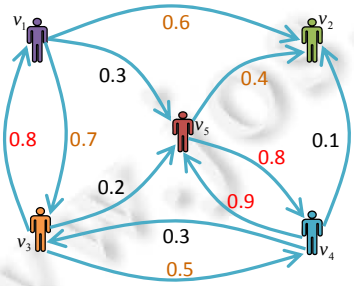


Fig.1 Simple graph of social trust network

图 1 简单的社交信任网络图

针对上述问题,本文将从社交关系影响属性的间接影响和社交关系影响决策的直接影响这两种机制入手,讨论社交关系对用户评分所产生的影响.具体说来,本文提出了一种新的基于用户信任关系的概率矩阵分解模型(trust-driven probabilistic matrix factorization,简称 TPMF),从而实现了对上述两种机制的融合建模.进而,针对不同用户受两种机制影响程度不同的问题,本文借助用户与信任好友的评分相关性对用户进行聚类,并映射各类别中的用户到相应的个性化权值.然后,将个性化权值算法应用到 TPMF 模型,通过 TPMF 模型得到用户和项

目的潜在特征向量,并最终借此实现对用户评分的个性化预测,由此构建了一套完整的推荐系统框架.

在Epinions公开数据集上的大量实验结果表明,本文提出的TPMF及其衍生算法在各项指标上均优于现有的代表性算法,从而验证了本文所提出的影响机制及技术框架的有效性.

本文第1节对所要解决的基于信任关系的用户评分预测问题进行数学化定义,进而提出TPMF模型并介绍其技术细节.第2节介绍用户的个性化权重方案,并在此基础上提出完整的推荐框架.第3节介绍实验方案及实验结果.第4节介绍推荐系统及其与社交领域相结合的已有研究工作.最后,第5节对全文进行总结并展望未来工作.

1 基于用户信任关系的概率矩阵分解模型

本节提出了基于用户信任关系的概率矩阵分解模型.首先探讨了信任用户间接影响用户偏好和信任用户直接影响用户评分这两种不同机制及其融合建模;然后给出其联合后验概率形式,并利用梯度下降方法实现优化求解;最后,对上述模型的时间复杂度进行了分析讨论.

1.1 问题定义及符号说明

本文着重研究基于社交信任的评分预测问题.形式化地,将该问题进行如下定义.

定义1. 给定用户评分矩阵 $\boldsymbol{R}$ 以及用户间的信任权重矩阵 $\boldsymbol{T}$,训练得到模型 F,使得对于任意用户 u_i 以及任意项目 v_j,都可以用该模型预测 u_i 对 v_j 评分 r_{ij},即 $r_{ij}=F(\boldsymbol{R},\boldsymbol{T},u_i,v_j)$.

本文所涉及的符号总结在表1中.

Table 1 Symbols definition

表1 符号定义

符号	解释
$V=\{v_1,v_2,\ldots,v_m\}$	项目集合,共 m 个项目
$U=\{u_1,u_2,\ldots,u_n\}$	用户集合,共 n 个用户
$\boldsymbol{U}\in\mathbb{R}^{d\times n},\boldsymbol{u}_i\in\mathbb{R}^{d\times 1}$	用户特征矩阵以及用户 u_i 的特征向量,d 为特征维度
$\boldsymbol{V}\in\mathbb{R}^{d\times m},\boldsymbol{v}_j\in\mathbb{R}^{d\times 1}$	项目特征矩阵以及项目 v_j 的特征向量,d 为特征维度
$N_i=\{u_{v_1},u_{v_2},\ldots,u_{v_S}\}$	用户 u_i 的信任用户集合,$s=\lvert N_i\rvert$
$\boldsymbol{T}\in\mathbb{R}^{n\times n},t_{i,v}$	用户间的信任权重矩阵及用户 u_i 对信任用户 u_v 的信任权重
r_{ij}^1	用户 u_i 的信任用户对项目 v_j 的评分影响其对项目 v_j 的评分
r_{ij}^2	用户 u_i 的信任用户对其特征产生影响,从而影响其对项目 v_j 的评分
$\boldsymbol{R}\in\mathbb{R}^{n\times m},r_{ij}$	用户评分矩阵以及用户 u_i 对的项目 v_j 的最终评分
ω_{ij}	用户 u_i 对项目 v_j 评分的重要性
$k_i\in[1,n]$	用户 u_i 的声誉排名
Q_i	用户 u_i 评分过的项目集合
$\boldsymbol{f}_i\in\mathbb{R}^{1\times 5}$	用户 u_i 的信任特征

1.2 TPMF模型介绍

在引言中已经提到,针对用户行为的社交趋同性,传统方法往往笼统地认为其成因是信任用户之间的属性相似性[3].然而近年来的相关研究发现:信任关系往往并非间接影响用户的属性,而是直接作用于用户的决策[12].

事实上,对于不同的用户,其所受上述两种机制的影响也不同.例如,某个用户可能更认同具有相似偏好的朋友的影响,而其他用户则更会听从喜好不同但具有亲友关系的人的意见.由此可见:在实现这两种机制的融合建模时,需要对不同机制赋予不同的权重.在此基础之上,本文提出了一种新的基于两种信任机制融合建模的概率矩阵分解模型TPMF(trust-driven probabilistic matrix factorization).

TPMF的图模型如图2所示,其中,r_{ij}^1 表示信任好友的评分对用户评分的直接影响,r_{ij}^2 表示信任好友的特征对用户评分的间接影响.引入α权衡这两种不同的影响机制,得到最终的评分 $r_{ij}=\alpha r_{ij}^2+(1-\alpha)r_{ij}^1$.

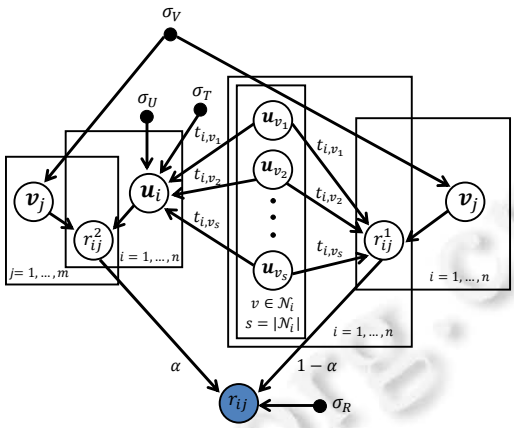


Fig.2 TPMF graph model

图 2 TPMF 图模型

针对信任用户直接影响用户评分这种机制,定义信任用户对用户 u_i 的评分影响为信任权重的加权平均,即:

$$r_{ij}^1 = \frac{\sum_{s\in N_i} t_{i,s}\boldsymbol{u}_s^T\boldsymbol{v}_j}{\sum_{s\in N_i} t_{i,s}} \tag{1}$$

其中,$t_{i,s}$ 表示用户 u_i 对信任用户 u_s 的信任权重.

同样地,针对信任用户间接影响用户偏好这种机制,定义信任用户对用户 u_i 的特征影响 $\hat{\boldsymbol{u}}_i$ 为信任权重的加权平均,即:

$$\hat{\boldsymbol{u}}_i = \frac{\sum_{s\in N_i} t_{i,s}\boldsymbol{u}_s}{\sum_{s\in N_i} t_{i,s}} \tag{2}$$

以上两种不同影响机制的区别在于:前者是通过信任好友的评分影响待推荐用户对该项目的评分;而后者是信任好友通过影响待推荐用户的兴趣特征,进而间接影响该用户的评分.

为了方便处理,对每一个用户的信任权重进行归一化,从而得到修正后的 r_{ij}^1 和 $\hat{\boldsymbol{u}}_i$ 分别为

$$r_{ij}^1 = \sum_{s\in N_i} t_{i,s}\boldsymbol{u}_s^T\boldsymbol{v}_j \tag{3}$$

$$\hat{\boldsymbol{u}}_i = \sum_{s\in N_i} t_{i,s}\boldsymbol{u}_s \tag{4}$$

进一步,定义用户 u_i 在受其信任好友间接的偏好影响后对项目 v_j 的评分为

$$r_{ij}^2 = \boldsymbol{u}_i^T\boldsymbol{v}_j \tag{5}$$

在分析了信任用户对待推荐用户的评分影响 r_{ij}^1 以及信任用户对待推荐用户的特征影响 $\hat{\boldsymbol{u}}_i$ (进而通过 r_{ij}^2 影响待推荐用户的最终评分)后,定义观测评分的条件概率:

$$p\left(R\,|\,\boldsymbol{U},\boldsymbol{V},\boldsymbol{T},\sigma_R^2\right) = \prod_{i=1}^{n}\prod_{j=1}^{m}\left[N\left(r_{ij}\,|\,(1-\alpha)r_{ij}^1+\alpha r_{ij}^2,\sigma_R^2\right)\right]^{\omega_{ij}} \tag{6}$$

其中,ω_{ij} 是指示函数,通常情况下,用于判断用户是否对项目有过评分.当用户 u_i 对项目 v_j 有评分时,ω_{ij}=1;否则,ω_{ij}=0.但事实上,并不是所有用户的评分都是同等重要的[14,15].文献[16]指出,用户间的社交关系反映了用户在整个网络中的声誉,而声誉作为用户地位的直接体现,有着重要的作用.根据地位理论[17],用户倾向于链接比自己地位高的用户,从而高地位的用户有着更强的影响力与号召力.因此,高地位用户的评分相对于其他用户的评分而言会显得更加重要.基于以上分析,我们首先利用最流行的 PageRank 算法[18]计算各个用户的声誉排名.令 $k_i\in[1,n]$表示用户 u_i 的声誉排名,k_i 值越小,表示其排名越高,也就是用户地位越高.接着,我们定义函数将用户 u_i 的声誉排名映射到其评分重要性,即,当用户 u_i 对项目 v_j 有评分时:

$$\omega_{ij}=\sqrt{\frac{1}{1+\ln(k_i)}} \tag{7}$$

该函数保证了$\omega_{ij}\in[0,1]$,且当用户声誉排名越靠前时,用户的评分重要性就越大.同时,经过和其他一些函数实验比较后发现,该函数能够更好体现高声誉用户有着更大评分重要性的特点.

公式(6)中的 $\boldsymbol{U},\boldsymbol{V}$ 为随机变量,类似于文献[19]中的假设,其满足:

$$p\left(\boldsymbol{U}\mid\sigma_U^2\right)=\prod_{i=1}^{n}N\left(\boldsymbol{u}_i\mid 0,\sigma_U^2\boldsymbol{I}\right) \tag{8}$$

$$p\left(\boldsymbol{V}\mid\sigma_U^2\right)=\prod_{j=1}^{m}N\left(\boldsymbol{v}_j\mid 0,\sigma_V^2\boldsymbol{I}\right) \tag{9}$$

其中, $N(x\mid 0,\sigma_U^2\boldsymbol{I})$ 是指均值为 0、方差为 $\sigma_U^2\boldsymbol{I}$ 的 d 维高斯分布的概率密度函数.

由于用户的特征不仅受到自身影响,还受到其信任用户的影响,所以易得:

$$p\left(\boldsymbol{U}\mid\boldsymbol{T},\sigma_U^2,\sigma_T^2\right)\propto p\left(\boldsymbol{U}\mid\sigma_U^2\right)p\left(\boldsymbol{U}\mid\boldsymbol{T},\sigma_T^2\right)=\prod_{i=1}^{n}N\left(\boldsymbol{u}_i\mid 0,\sigma_U^2\boldsymbol{I}\right)\times\prod_{i=1}^{n}N\left(\boldsymbol{u}_i\mid\sum_{s\in N_i}t_{i,s}\boldsymbol{u}_s,\sigma_T^2\boldsymbol{I}\right) \tag{10}$$

从而,再次利用贝叶斯公式后可以得到关于特征矩阵 $\boldsymbol{U},\boldsymbol{V}$ 的后验概率:

$$p\left(\boldsymbol{U},\boldsymbol{V}\mid\boldsymbol{R},\boldsymbol{T},\sigma_R^2,\sigma_T^2,\sigma_U^2,\sigma_V^2\right)\propto p\left(\boldsymbol{R}\mid\boldsymbol{U},\boldsymbol{V},\boldsymbol{T},\sigma_R^2\right)p\left(\boldsymbol{U}\mid\sigma_U^2\right)p\left(\boldsymbol{U}\mid\boldsymbol{T},\sigma_T^2\right)p\left(\boldsymbol{V}\mid\sigma_V^2\right) \tag{11}$$

1.3 TPMF模型求解

对公式(11)取对数,可得:

$$\left.\begin{aligned}\ln p\left(\boldsymbol{U},\boldsymbol{V}\mid\boldsymbol{R},\boldsymbol{T},\sigma_R^2,\sigma_T^2,\sigma_U^2,\sigma_V^2\right)&=\ln p\left(\boldsymbol{R}\mid\boldsymbol{U},\boldsymbol{V},\boldsymbol{T},\sigma_R^2\right)+\ln p\left(\boldsymbol{U}\mid\boldsymbol{T},\sigma_T^2\right)+\ln p\left(\boldsymbol{U}\mid\sigma_U^2\right)+\ln p\left(\boldsymbol{V}\mid\sigma_V^2\right)\\&=-\frac{1}{2}\left(nd\ln\sigma_T^2+nd\ln\sigma_U^2+md\ln\sigma_V^2\right)-\\&\quad\frac{1}{2\sigma_R^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_{ij}\left(r_{ij}-\left((1-\alpha)r_{ij}^1+\alpha r_{ij}^2\right)\right)^2-\\&\quad\frac{1}{2\sigma_U^2}\sum_{i=1}^{n}\boldsymbol{u}_i^T\boldsymbol{u}_i-\frac{1}{2\sigma_T^2}\sum_{i=1}^{n}\left(\boldsymbol{u}_i-\sum_{s\in N_i}t_{i,s}\boldsymbol{u}_s\right)^T\left(\boldsymbol{u}_i-\sum_{s\in N_i}t_{i,s}\boldsymbol{u}_s\right)-\\&\quad\frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_{ij}\times\ln\sigma_R^2\right)-\frac{1}{2\sigma_V^2}\sum_{j=1}^{m}\boldsymbol{v}_j^T\boldsymbol{v}_j+c\end{aligned}\right\} \tag{12}$$

忽略上式中的无关常量 c 并整理后,得到如下目标函数:

$$\left.\begin{aligned}L=&\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_{ij}\left(r_{ij}-\left((1-\alpha)\sum_{s\in N_i}t_{i,s}\boldsymbol{u}_s^T\boldsymbol{v}_j+\alpha\boldsymbol{u}_i^T\boldsymbol{v}_j\right)\right)^2+\\&\frac{\lambda_T}{2}\sum_{i=1}^{n}\left(\boldsymbol{u}_i-\sum_{s\in N_i}t_{i,s}\boldsymbol{u}_s\right)^T\left(\boldsymbol{u}_i-\sum_{s\in N_i}t_{i,s}\boldsymbol{u}_s\right)+\frac{\lambda_U}{2}\sum_{i=1}^{n}\boldsymbol{u}_i^T\boldsymbol{u}_i+\frac{\lambda_V}{2}\sum_{j=1}^{m}\boldsymbol{v}_j^T\boldsymbol{v}_j\end{aligned}\right\} \tag{13}$$

其中, $\lambda_U=\sigma_R^2/\sigma_U^2,\lambda_V=\sigma_R^2/\sigma_V^2,\lambda_T=\sigma_R^2/\sigma_T^2$ 均作为常数处理,其值在实验时选定.需要注意上式中用户 u_i 的信任好友通过 $\lambda_T/2\times\sum_{i=1}^{n}\left(\boldsymbol{u}_i-\sum_{s\in N_i}t_{i,s}\boldsymbol{u}_s\right)^T\left(\boldsymbol{u}_i-\sum_{s\in N_i}t_{i,s}\boldsymbol{u}_s\right)$ 影响用户特征向量 $\boldsymbol{u_i}$,从而影响用户 u_i 对项目 v_j 的评分 $\alpha r_{ij}^2=\alpha\boldsymbol{u}_i^T\boldsymbol{v}_j$.求得使目标函数达到最小值时的特征矩阵 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 后,便可以此进行个性化项目推荐.本文采用梯度下降算法进行求解,目标函数关于特征向量的偏导数分别为

$$\left.\begin{aligned}\frac{\partial L}{\partial\boldsymbol{u}_i}=&\lambda_T\left(\boldsymbol{u}_i-\sum_{s\in N_i}t_{i,s}\boldsymbol{u}_s\right)-\lambda_T\sum_{\{k\mid i\in N_k\}}t_{k,i}\left(\boldsymbol{u}_k-\sum_{w\in N_k}t_{k,w}\boldsymbol{u}_w\right)+\\&\alpha\sum_{j=1}^{m}\omega_{ij}\boldsymbol{v}_j(g_1-r_{ij})+(1-\alpha)\sum_{\{k\mid i\in N_k\}}\sum_{j=1}^{m}\omega_{kj}t_{k,i}\boldsymbol{v}_j(g_2-r_{kj})+\lambda_U\boldsymbol{u}_i\end{aligned}\right\} \tag{14}$$

$$\frac{\partial L}{\partial\boldsymbol{v}_j}=\sum_{i=1}^{n}\omega_{ij}\left((1-\alpha)\sum_{s\in N_i}t_{i,s}\boldsymbol{u}_s+\alpha\boldsymbol{u}_i\right)(g_1-r_{ij})+\lambda_V\boldsymbol{v}_j \tag{15}$$

其中, $g_1=\left((1-\alpha)\sum_{s\in N_i}t_{i,s}\boldsymbol{u}_s^T\boldsymbol{v}_j+\alpha\boldsymbol{u}_i^T\boldsymbol{v}_j\right),g_2=\left((1-\alpha)\sum_{w\in N_k}t_{k,w}\boldsymbol{u}_w^T\boldsymbol{v}_j+\alpha\boldsymbol{u}_k^T\boldsymbol{v}_j\right)$.

利用梯度下降算法求得所有用户和项目的特征向量,由此便可预测用户 u_i 对项目 v_j 的评分.

1.4 TPMF模型时间复杂度分析

下面计算基于用户信任关系的概率矩阵分解模型的时间复杂度,该模型的代价函数计算公式为公式(13),偏导函数计算公式为公式(14)、公式(15),用户数是 n,项目数是 m,且特征向量维度是 d.现在假设任意用户 u 平均打分的个数是 $\bar{r}$,平均信任用户的个数是 $\bar{t}$,则更新目标函数 L 的时间复杂度为

$$O(n\bar{r}\bar{t}d+n\bar{t}d+md+nd)=O(n\bar{r}\bar{t}d).$$

其中需要指出的是:现实推荐中,常见的项目个数 m 一般都小于总的评分数 $n\bar{r}$,从而可知上式是正确的.

同时,计算对全体用户和项目的偏导数所需的时间代价为

$$O(n\bar{r}\bar{t}d+n\bar{r}\bar{t}^2d+nd+n\bar{t}d+n\bar{t}^2d+md)=O(n\bar{r}\bar{t}^2d).$$

由此可见,TPMF 的计算代价与用户个数 n、特征维度 d、用户平均信任关系数 $\bar{t}$ 和用户平均评分数 $\bar{r}$ 这 4 个因素呈正相关.考虑到应用中的长尾特性[20],一般用户的信任用户和评分数都有限,因此,计算代价的主要决定性因素在于用户个数 n.

2 个性化权值算法

在基于用户信任关系的概率矩阵分解模型中,本文通过对每一个用户赋予相同的权值α来权衡不同影响机制的作用.然而,基本的 TPMF 算法并没有考虑到不同用户所受两种影响机制的不同,因此,本文借助评分相关性对用户进行聚类并映射到相应权重,从而得到个性化权值α.由于训练数据的稀疏片面性,直接为每个用户赋予个性化权值很可能降低模型的泛化性能,因此在本文中,我们讨论的是将用户按照特征进行聚类,进而为每个类别中的用户赋予相同权值.

2.1 用户信任特征

深入探究两种不同影响机制的规律,不难发现如果一名用户倾向于接受来自于好友属性特征的间接影响,那么在面对具有不同属性的好友时,其评分所受影响往往会体现出一定的强弱不同;反之,如果该用户更倾向于接受来自于用户评分的直接影响,那么即使面对不同好友,其评分的相似性也会相对一致.换言之,倾向于接受好友评分直接影响的用户,其评分与好友的相似性将体现得更为明显.因此,从用户与信任好友评分的一致性入手,将有助于我们衡量用户受两种不同机制的不同影响.

为此,本文从已有的评分数据入手,对目标用户及其信任好友对同一项目的评分信息提出如下直观假设:

假设 1. 对于一个给定的用户 u_i,如果其更易借鉴信任的好友对项目的评分,那么用户 u_i 实际对该项目的评分与其信任好友对该项目的评分之间的差距应更加小.

因此,将给定用户与信任好友对同一项目的评分差作为一种关键特征,考虑到用户对项目的评分只能从{1,2,3,4,5}中选取,所以评分差的取值范围为集合{0,1,2,3,4}.同时,考虑到不同用户有着不同的信任用户个数以及评分数目,相应的结果需进行归一化处理.具体地,定义用户 u_i 的信任用户集合为 N_i,用户 u_i 评分过的项目集合为 Q_i,且用户 u_i 对项目 v_j 的评分为 r_{ij}.$\forall u_s\in N_i, v_k\in Q_i$,如果用户 u_s 对项目 v_k 有评分 r_{sk},则记录用户和信任好友对该项目的评分差$|r_{ik}-r_{sk}|$.统计所有值为 $r\in\{0,1,2,3,4\}$ 的评分差的个数 k_r,并将其归一化后作为用户 u_i 的信任特征为 $\boldsymbol{f}_i$.

下面举例说明如何求得给定用户 u_i 的信任特征 $\boldsymbol{f}_i$.如图 3 所示:给定用户 u_i、用户 u_i 对 5 个项目的评分(红色数字表示)以及用户 u_i 的信任好友对这 5 个项目的评分(蓝色数字表示),则在一条折线上的红、黄、蓝这 3 个点对应一个用户及其信任好友对同一个项目的评分差(黑色数字表示).从图中可以看到:给定的用户 u_i 一共评分过 5 个项目,而且对于每个用户评分过的项目,都存在部分信任用户对该项目进行过评分.图中一共存在 12 个用户 u_i 及其信任好友对同一项目的评分差,且值为 0~4 的分别有 4,5,3,0,0 个.从而,经过归一化后得到用户 u_i 的信任特征为[4/12 5/12 3/12 0/12 0/12].

在求得每个用户 u_i 的信任特征后,根据用户特征进行聚类,从而使得每一个类别中的用户具有相同的权值α,并且每个类别中的用户对应的权重大小与该类用户及其信任好友对同一项目的评分差的大小有关.下面将

具体对这两个问题进行讨论.

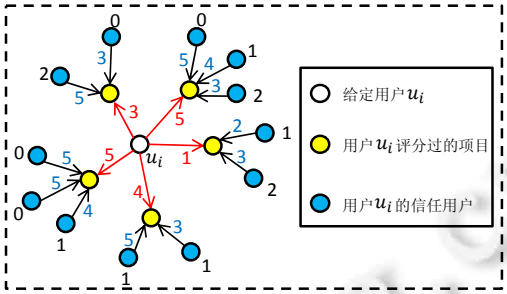


Fig.3 An example illustration

图 3 例子示意图

2.2 用户聚类

首先,随机选择 k 个用户的特征向量作为 k 个类别 $c_1,c_2,\ldots,c_k$ 的初始化特征向量.然后,对于每一个用户 u_i,计算其与各个聚类中心之间的余弦相似度大小:

$$sim(u_i,c_k)=\frac{\sum_{j=0}^{4} f_{i,j}\times c_{k,j}}{\sqrt{\sum_{j=0}^{4} f_{i,j}^2}\times\sqrt{\sum_{j=0}^{4} c_{k,j}^2}} \tag{16}$$

这里,$f_{i,j}$ 表示用户 u_i 的信任特征向量的第 j 个分量,$c_{k,j}$ 表示类别 c_k 的特征向量的第 j 个分量.

然后,将用户 u_i 的类别更新为与其相似度最大的聚类中心所在类别,并将每一个类别 c_k 的特征向量更新为该类中的所有用户信任特征的均值.

在对用户进行聚类之后,可以认为每一个类别中的用户具有相同的个性化权值α.下面引入权重函数将各个用户通过其所属类别的特征向量映射到个性化权值.

2.3 用户个性化权重

在得到每个用户所属类别后,其类别特征向量表示用户在各个评分差{0,1,2,3,4}上的倾向比重.因此,引入权重$\boldsymbol{w}$=[0 1 2 3 4],其含义为用户和信任好友对同一项目的评分差.则权重$\boldsymbol{w}$和每一个聚类中心的特征向量作内积,即为该类用户与信任好友的平均评分差 $\sum_{j=0}^{4} w_j\times c_{k,j}$.考虑到 TPMF 模型已经得到最佳固定权值α,进一步合理假设,可以将α视作用户个性化权值的均值,而借助一个较小的波动范围ε,对不同机制作用下的用户群体加以区分.为方便处理,不妨在α周围按照大小均匀地选取 k 个点作为 k 个类别对应的个性化权重.具体方案为:

以 TPMF 模型最佳固定权值α为中心,参数ε为半径的区间$[\alpha-\varepsilon,\alpha+\varepsilon]$中,依照平均评分差 $\sum_{j=0}^{4} w_j\times c_{k,j}$ 的大小按顺序平均取 k 个点构成每个类别的个性化权值(即这 k 个点在区间$[\alpha-\varepsilon,\alpha+\varepsilon]$上均匀分布).该方法的高效之处在于:可以先通过调节α来估计合理的权值范围,然后固定α调节,从而快捷地得到个性化权值.如:在以α=0.6为中心、ε=0.1 为半径的区间[0.5,0.7]中选取 3 个个性化权重,则个性化权值从大到小依次为 0.7,0.6 和 0.5.我们将在实验部分讨论参数ε的敏感性.

至此,通过提取用户特征,聚类分析,利用权值函数进行特征映射后,便得到了个性化权值.下面将个性化权重融入到 TPMF 模型,得到完整的推荐框架.

2.4 完整推荐框架

综合上文所述的 TPMF 模型及其个性化权值算法,本文提出了如图 4 所示的完整推荐框架,以实现特定用户的个性化评分预测.

① 从数据库中读取用户历史评分数据以及历史信任数据;

② 利用评分数据以及信任数据挖掘用户信任特征向量,并以此对用户进行聚类后,利用权重函数得到描

述用户借鉴信任好友评分程度的个性化权值;

③ 利用信任数据通过 PageRank 算法得到用户评分重要性,并结合评分数据以及步骤 2 中得到的个性化权值,通过 TPMF 模型学习出用户的特征向量以及项目的特征向量;

④ 在得到用户的个性化权值、用户的特征向量以及项目的特征向量后,通过评分预测函数预测用户对项目的评分,进而推荐评分最高的 N 个项目.

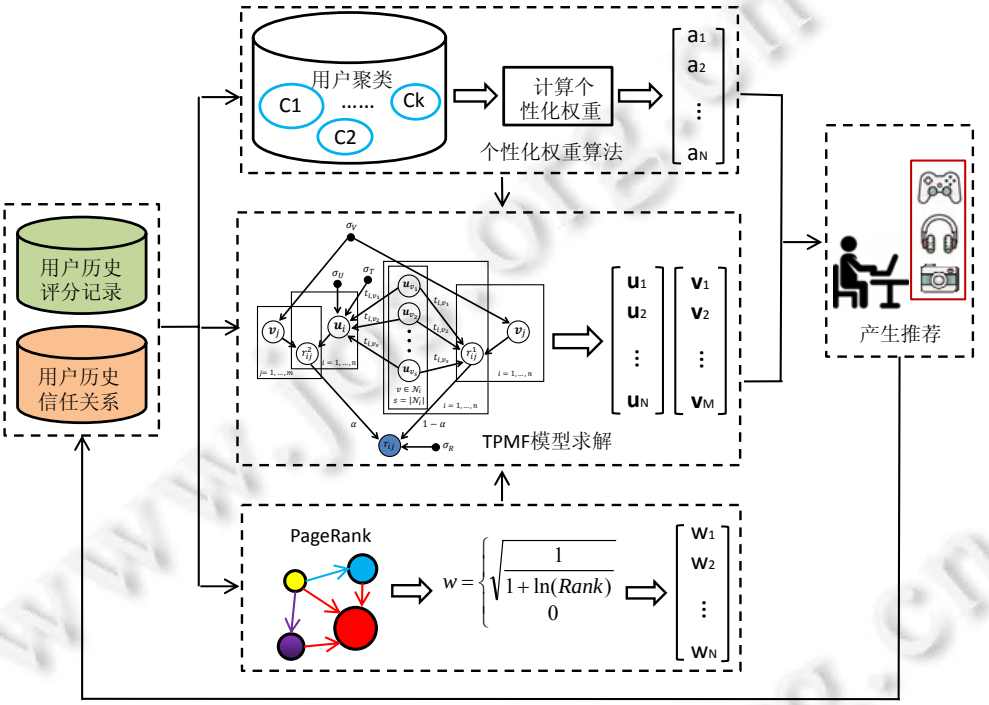


Fig.4 Recommendation framework

图 4 推荐框架

对于大部分社交网络而言,其并没有提供明确的信任关系,因此,在缺少标记数据的情况下预测信任关系并不现实.我们采用的方法是以普通社交关系为基础,结合其他信息如转发、评论等行为进行估计;而对于少数提供明确信任关系但又比较稀疏的社交网络,我们可以利用常见的链接预测算法[21,22],通过已有的信任关系或者再借助其他交互数据预测新的信任关系.因此,本文提出的推荐框架可以推广至一般的社交网络.此外,根据社交网络中的同质性原理,具有相似评分行为,即具有共同偏好的用户对更容易建立信任关系.因此,在给定时间信息的情况下,本文的框架可以进一步动态分析用户偏好的改变以及用户间信任关系的演化过程,从而做出更加准确的推荐.上面的分析表明,本文提出的推荐框架具有良好的通用性与扩展性.

3 实验结果与分析

3.1 实验设置

实验数据集采用公开数据集 Epinions 以及 Ciao.为保证实验质量,同时研究不同特征用户对算法的适用性,本文采用 3 种方法对用户进行筛选.实验删去了其中评分数较少的用户、被评分数较少的项目以及信任用户数较少的用户,从而得到新的数据集.表 2 中,实验采用 Epinions 以及 Ciao 数据集的 3 个子集:Epinions1,Epinions2,Epinions3 以及 Ciao1,Ciao2,Ciao3 作为实验数据集.其中:Epinions1 和 Ciao1 数据集是通过筛选原始数据集评分数和信任用户数均大于 50 的用户以及被评分数大于 50 的项目后得到;Epinions2 和 Ciao2 数据集则通过筛选原

始数据集评分数和信任用户数均大于30的用户以及被评分数大于30的项目后得到;Epinions3和Ciao3数据集则通过筛选原始数据集评分数和信任用户数均大于 10 的用户以及被评分数大于 10 的项目后得到.其中,Epinins3 以及 Ciao3 数据集不仅与实际数据集更加相近,同时也能进一步研究不同筛选方法下,本文提出算法的性能.数据集的详细信息见表 2.

Table 2 Details for datasets

表 2 数据集详细信息

数据集	用户个数	项目个数	评分记录	信任关系
Epinions1	396	1 797	11 802	7 583
Epinions2	1 100	3 659	32 576	26 367
Epinions3	6 329	13 969	145 842	262 170
Ciao1	425	285	7 697	15 812
Ciao2	773	716	16 182	31 140
Ciao3	1 974	4 387	58 114	69 436

实验采用平均绝对误差(MAE)评估机制来衡量算法预测性能的优劣,该评价方法具体定义如下:

$$MAE=\frac{\sum_{(i,j)\in\Gamma}|r_{i,j}-\hat{r}_{i,j}|}{|\Gamma|} \tag{17}$$

其中,Γ是实验测试数据集,为〈用户,项目〉二元组;$r_{i,j}$表示用户 u_i 对已经评分过的项目 v_j 的实际评分;而 $\hat{r}_{i,j}$ 则表示利用模型预测出的用户 u_i 对项目 v_j 的评分.

为了全面地评价文中提出的 TPMF 算法的推荐效果,本文采用如下 3 种对比算法:(1) 概率矩阵分解(probabilistic matrix factorization,简称 PMF)算法[19];(2) STE 算法[2](用户信任关系只影响评分);(3) SocialMF 算法[3](用户信任关系只影响特征向量).同时,为了更好地验证提出的算法,本文将第 2 节中提出的算法称为 TPMF 算法,并将使用个性化权重方法的 TPMF 算法称为 PTPMF(personalized trust-driven probabilistic matrix factorization)算法.这些模型的各自特点见表 3.

Table 3 Characteristics of the models

表 3 模型的特点

模型	评分信息	社交信息影响评分	社交信息影响特征	个性化权重
PMF	√	×	×	×
STE	√	√	×	×
SocialMF	√	×	√	×
TPMF	√	√	√	×
PTPMF	√	√	√	√

实验将随机得到的 80%评分数据作为训练集,余下的数据作为测试集.实验中,各模型特征维度 D 的改变会影响模型预测性能以及时间复杂度,文献[3,23]对此作了进一步调研.但对同一个特征维度 D 而言,各模型都是公平的.因此本文分别设定 D 为 5 和 10 进行实验[3].

3.2 实验结果

表 4~表 6 给出了本文提出的 TPMF 以及 PTPMF 算法与对比算法 PMF,STE 以及 SocialMF 在 Epinions1, Ciao1,Epinions2,Ciao2,Epinions3 以及 Ciao3 数据集上的 MAE 值.在实验中,各个模型的参数都设置为使模型推荐效果最佳时的参数.如 Epinions1 数据集上:当特征维度为 5 时,TPMF 的参数λ_T取为 1.8,而参数α取为 0.6; PTPMF 在 TPMF 的参数设置上将用户分为 2 类,每一类的权重分别为 0.77 和 0.43;对比算法 SocialMF 的参数 λ_T 设为 3.1,STE 的参数α设为 0.5.

从实验结果中可以看到:在不同的数据集上,本文提出的 TPMF 以及 PTPMF 算法相对于已有的 PMF,STE 以及 SocialMF 算法在推荐效果上都有着不同程度的提升.这是由于已有的这些模型并未有效地对信任用户的影响进行分类和量化,而 PTPMF 充分考虑了用户间的信任关系影响机制,并为每个用户提供了个性化权重.同时,实验结果也充分说明了本文提出的用户评分受信任用户两种不同机制影响的正确性.而 PTPMF 的推荐效果

相对于 TPMF 的进一步提高,也表明考虑用户个性化权重的合理性以及必要性.此外,通过对比 3 种筛选方法得到的数据集及其实验结果可以发现:当信任关系数/评分数越大时,本文提出的算法相对于已有算法具有更大的优势.

Table 4 MAE values of each model in Epinions1 and Ciao1

表 4 Epinions1 及 Ciao1 中各个模型的 MAE 值

	Epinions1		Ciao1	
方法	D=5	D=10	D=5	D=10
PMF	0.947 5	0.935 1	0.886 7	0.892 1
STE	0.906 3	0.907 8	0.873 4	0.880 5
SocialMF	0.904 8	0.904 7	0.874 0	0.875 5
TPMF	0.901 9	0.900 2	0.866 8	0.871 0
PTPMF	**0.900 1**	**0.899 5**	**0.866 0**	**0.869 8**

Table 5 MAE values of each model in Epinions2 and Ciao2

表 5 Epinions2 及 Ciao2 中各个模型的 MAE 值

	Epinions2		Ciao2	
方法	D=5	D=10	D=5	D=10
PMF	0.889 9	0.902 1	0.855 9	0.855 7
STE	0.875 2	0.882 6	0.842 6	0.842 7
SocialMF	0.879 6	0.879 3	0.831 7	0.831 6
TPMF	0.873 2	0.872 9	0.811 3	0.817 8
PTPMF	**0.872 5**	**0.872 2**	**0.805 9**	**0.811 5**

Table 6 MAE values of each model in Epinions3 and Ciao3

表 6 Epinions3 及 Ciao3 中各个模型的 MAE 值

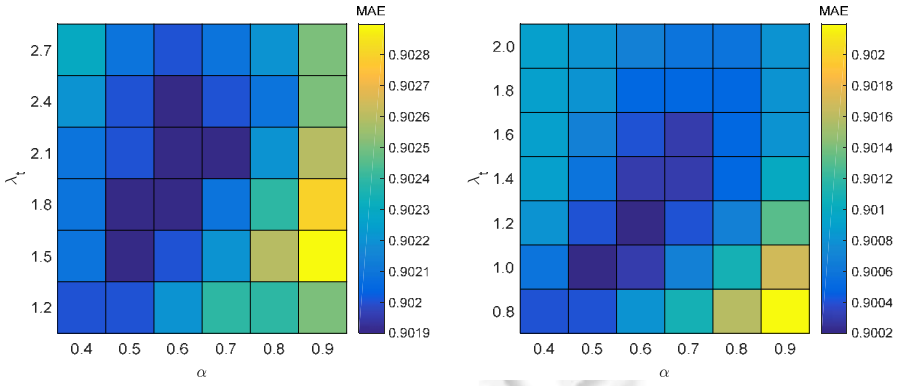
	Epinions3		Ciao3	
方法	D=5	D=10	D=5	D=10
PMF	0.878 0	0.870 7	0.811 0	0.809 7
STE	0.862 2	0.859 4	0.794 2	0.794 0
SocialMF	0.857 0	0.859 7	0.780 8	0.781 2
TPMF	0.850 6	0.849 3	0.775 5	0.776 3
PTPMF	**0.848 9**	**0.848 5**	**0.772 9**	**0.772 5**

对于每个 Epinions1 测试集中的用户项目对$\langle u_i,v_j\rangle$,将 PTPMF 算法(D=5)求得的用户 u_i 对项目 v_j 的评分记为 P_{ij},同时将 SocialMF 算法(D=5)求得的结果记为 S_{ij}.令 $\boldsymbol{p}$ 是所有用户项目对$\langle u_i,v_j\rangle$对应的预测评分 P_{ij} 组成的向量,$\boldsymbol{s}$ 是对应的预测评分 S_{ij} 组成的向量.下面对 $\boldsymbol{p}$ 和 $\boldsymbol{s}$ 进行 t 检验,其中,原假设和备择假设分别为:H_0:$\boldsymbol{p}=\boldsymbol{s}$,$H_1$:$\boldsymbol{p}\neq\boldsymbol{s}$.在显著性水平为 0.05 的情况下,Epinions1 数据集(D=5)中所得的 p 值为 0.02<0.05.因此拒绝原假设,即,PTPMF 算法得到的预测评分和 SocialMF 算法得到的预测评分之间存在显著性差异.在其他数据集以及不同特征维度时的 t 检验结果也类似,在此不一一赘述.

3.3 参数α以及λ_T对推荐性能的影响

为了更好地研究信任用户的评分影响以及特征影响,在参数α以及λ_T 变化的情况下对推荐性能进行研究.从图 5~图 8 中可以看到:参数α和参数λ_T 会相互影响,仅当α和λ_T 呈某种正相关性时,模型会取得较好的推荐效果.这是由于λ_T对应信任用户的特征影响,其值越大,信任用户的影响效果越明显;而α对应信任用户的评分影响,其越小,信任用户的影响效果就越强.所以当α偏小而λ_T 过大时,用户会严重受到信任用户的影响;而当偏α大,λ_T 过小时,用户几乎完全只受自身特征影响.这两种情况都不利于提升推荐性能.因此,当α和λ_T 同时增大或者减小时,其推荐效果会得到明显的提升.从图中可以看到,在对角线附近的α和λ_T 能够获得好的推荐效果.这是由于用户充分结合了自身兴趣偏好以及信任好友的特征影响和评分影响.

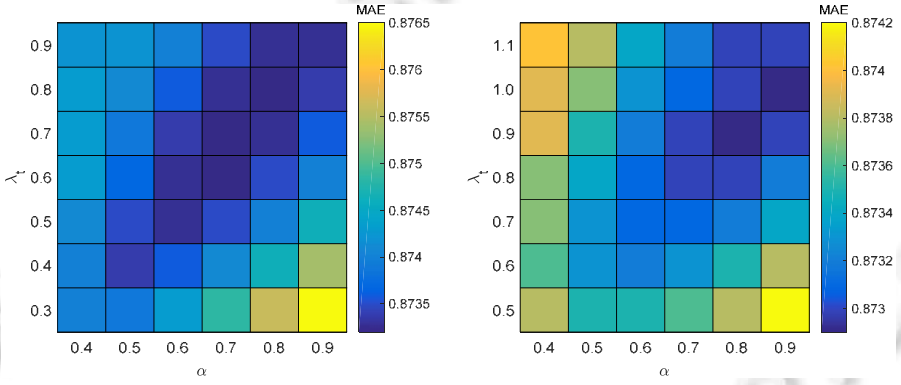
最后,通过观察图中的每一行以及每一列的推荐效果可以发现:对于同一个参数λ_T,其推荐效果随着α的增大先提高后降低;而对于同一个参数α,其推荐效果随着λ_T 的增大先提升后又降低.这些结果证实了本文提出的两种信任用户影响机制的正确性.



(a) Epinion1 (D=5) (b) Epinion1 (D=10)

Fig.5 Influence of parameters' variation on MAE in Epinions1

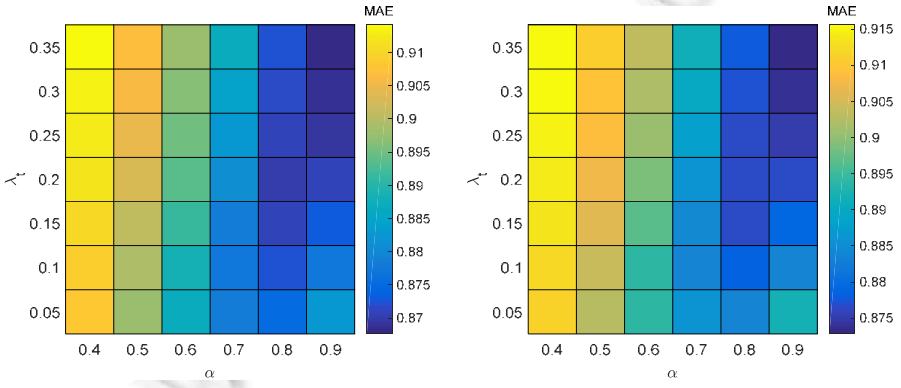
图 5 Epinion1 上参数变化对 MAE 的影响图



(a) Epinion2 (D=5) (b) Epinion2 (D=10)

Fig.6 Influence of parameters' variation on MAE in Epinions2

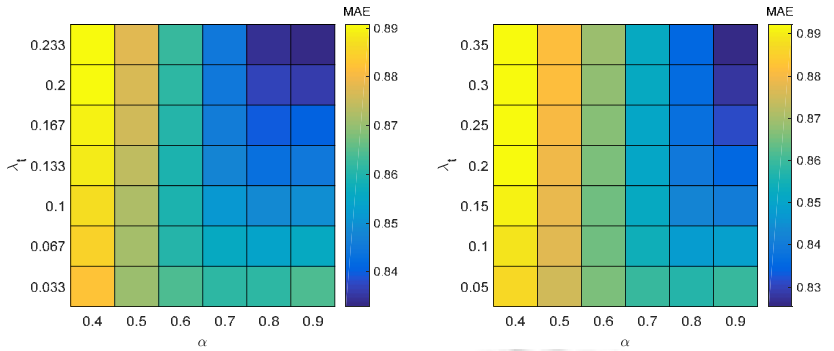
图 6 Epinion2 上参数变化对 MAE 的影响图



(a) Ciao1 (D=5) (b) Ciao1 (D=10)

Fig.7 Influence of parameters' variation on MAE in Ciao1

图 7 Ciao1 上参数变化对 MAE 的影响图



(a) Ciao2 (D=5) (b) Ciao2 (D=10)

Fig.8 Influence of parameters' variation on MAE in Ciao2

图 8 Ciao2 上参数变化对 MAE 的影响图

3.4 不同比例训练集下的推荐结果

为了衡量本文提出的算法在不同比例的训练集下的推荐效果,在 Epinions 1,Epinions 2 以及 Ciao1,Ciao2 这 4 个数据集上分别对 40%~80%的训练样本进行实验,并取特征维度为 5.实验结果如图 9 和图 10 所示.

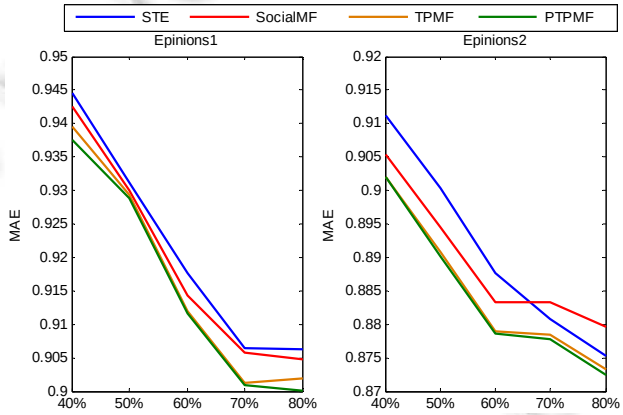


Fig.9 Recommendation effect in different proportion of Epinions training data

图 9 不同比例 Epinions 训练集下的推荐效果

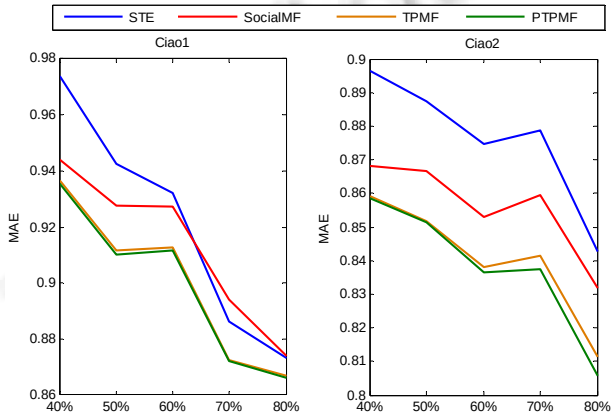


Fig.10 Recommendation effect in different proportion of Ciao training data

图 10 不同比例 Ciao 训练集下的推荐效果

从图 9 和图 10 中可以看到,在两个数据集中,各个算法的推荐效果随着训练集比例的增加而提高.这是由于训练样本的增大一定程度上减小了评分数据稀疏性的影响.此外可以看到,本文提出的 TPMF 以及 PTPMF 算法在各个比例的训练集下都展现出最佳的推荐性能.这是由于这两个模型在已有的信任数据中进一步挖掘了信任用户对待评分用户的影响机制,从而更好地提升了推荐性能.

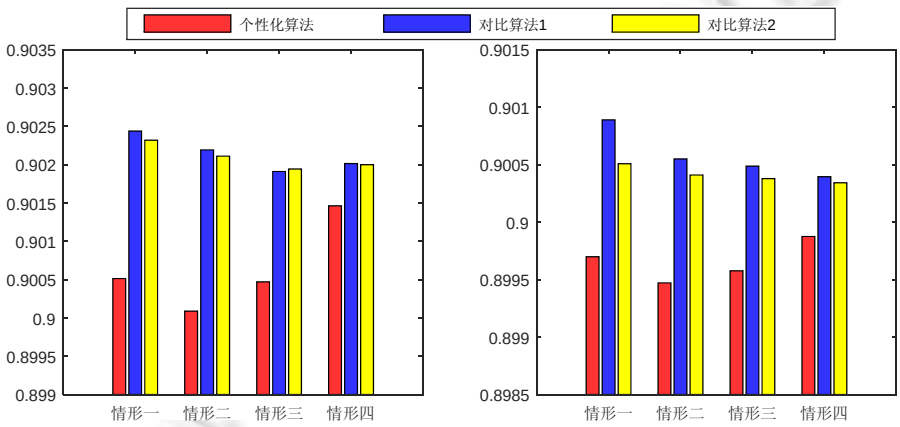
3.5 个性化权重的影响

为了更全面地比较个性化权重算法的实验效果,本文在 Epinions1,Epinions2 以及 Ciao1,Ciao2 数据集上分别对 PTPMF 模型当 D=5 以及 D=10 这两种情况进行验证.此外,为了直观地显示个性化权重算法的实验效果,不妨将聚类后所得类别数设为 2.具体的 PTPMF 模型个性化权重的取值以及对比算法的权重取值情况见表 7.比如,第 2 行表示各个模型在 Epinions1 数据集上的参数取值,其中,PTPMF 模型以及对比模型 TPMF 的特征维度 D 均取为 5.情形 1 即α=0.85/0.35,表示个性化权重算法(PTPMF)将用户聚类后分成 2 类,每一类的个性化权重α分别为 0.85 和 0.35(即,其对应的 TPMF 模型最佳固定权值α为 0.60,取参数ε=0.25,从而得到个性化权值 0.85 以及 0.35);而对比算法 1(TPMF)固定权重α为 0.85,对比算法 2(TPMF)固定权重α为 0.35.情形 2~情形 4 类似类似于情形 1,只是将个性化权重算法(PTPMF)分成的两个类的权重差异逐渐减小(情形 1~情形 4 的参数ε取值依次沿着 0.25,0.17,0.10,0.05 递减).实验分别针对表 7 中的不同情形进行验证,具体结果如图 11~图 14 所示.而表 7 中的最后一列表示使得 TPMF 取得最佳效果的α值.

Table 7 Values for PTPMF model

表 7 PTPMF 模型取值

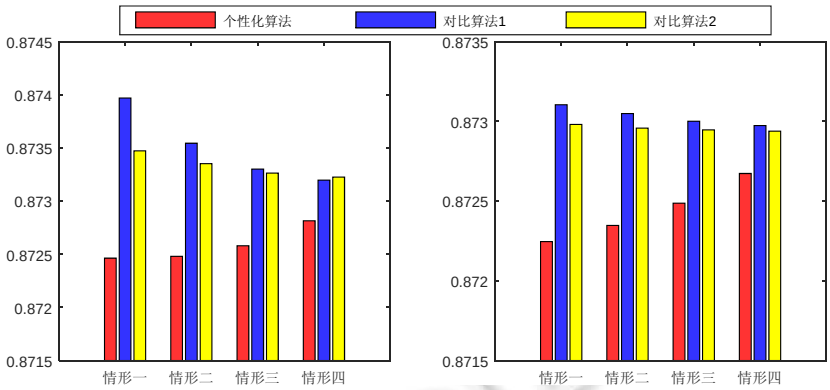
数据集	情形 1	情形 2	情形 3	情形 4	TPMF 取值
Epininos1 (D=5)	α=0.85/0.35	α=**0.77/0.43**	α=0.70/0.50	α=0.65/0.55	α=0.60
Epininos1 (D=10)	α=0.80/0.40	α=**0.73/0.47**	α=0.70/0.50	α=0.65/0.55	α=0.60
Epininos2 (D=5)	α=**1.00/0.40**	α=0.93/0.47	α=0.85/0.55	α=0.77/0.63	α=0.70
Epininos2 (D=10)	α=**1.00/0.60**	α=0.95/0.65	α=0.90/0.70	α=0.85/0.75	α=0.80
Ciao1 (D=5)	α=1.00/0.88	α=0.99/0.89	α=0.98/0.90	α=**0.97/0.91**	α=0.94
Ciao1 (D=10)	α=1.00/0.88	α=**0.98/0.90**	α=0.97/0.91	α=0.96/0.92	α=0.94
Ciao2 (D=5)	α=**1.00/0.88**	α=0.99/0.89	α=0.98/0.90	α=0.97/0.91	α=0.94
Ciao2 (D=10)	α=**1.00/0.90**	α=0.99/0.91	α=0.98/0.92	α=0.97/0.93	α=0.95



(a) Epinion1 (D=5) (b) Epinion1 (D=10)

Fig.11 Recommendation results of personalized algorithm in Epinions1

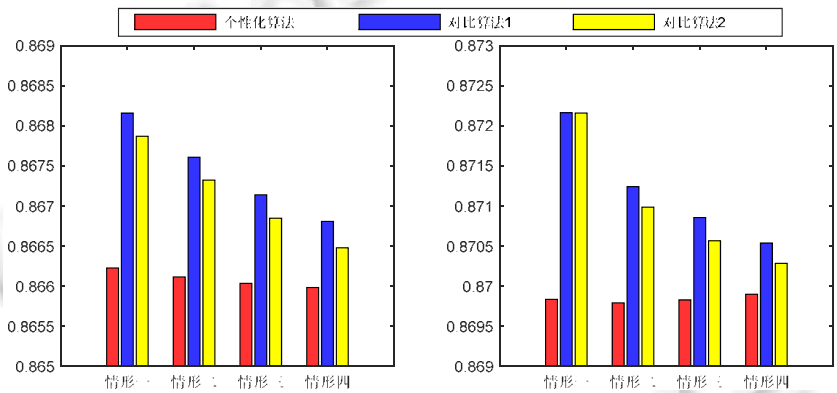
图 11 Epinions1 上个性化算法推荐结果



(a) Epinion2 (D=5) (b) Epinion2 (D=10)

Fig.12 Recommendation results of personalized algorithm in Epinions2

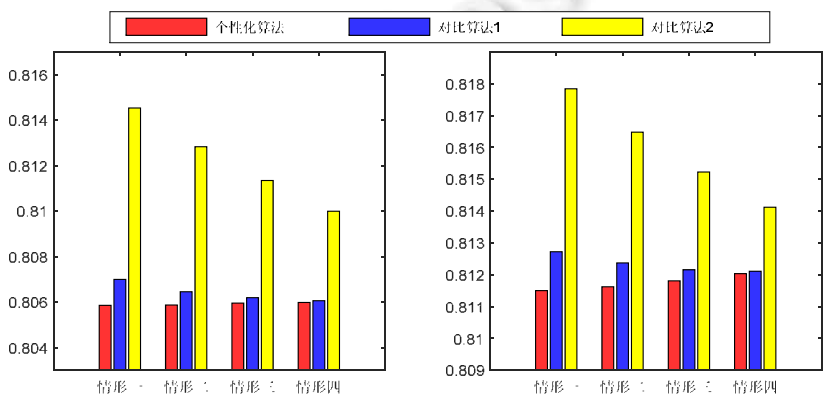
图 12 Epinions2 上个性化算法推荐结果



(a) Ciao1 (D=5) (b) Ciao1 (D=10)

Fig.13 Recommendation results of personalized algorithm in Ciao1

图 13 Ciao1 上个性化算法推荐结果



(a) Ciao2 (D=5) (b) Ciao2 (D=10)

Fig.14 Recommendation results of personalized algorithm in Ciao2

图 14 Ciao2 上个性化算法推荐结果

从图 11~图 14 的实验效果图中可以看到,PTPMF 算法在将用户聚类并对每一类赋予不同权重后,所得 *MAE* 值不仅均优于固定权重时的情形,同时也优于最佳 TPMF 算法的效果.此外,观察各个数据集中最优 PTPMF 算法的参数α时可以发现:对于同一个数据集,在模型维度改变时,最佳参数α均比较接近(如 Epinions1 数据集中,D=5 时,最佳α=0.77/0.43;D=10 时,最佳α=0.73/0.47).这是由于同一个数据集中的用户有着确定的最佳α值,它们不会随着模型的改变而发生变化;而在不同数据集中,最佳参数α之间差异比较明显(如 D=5 时,Epinions1 数据集中最佳α=0.77/0.43,而 Epinions2 以及 Ciao1 数据集中最佳α分别为 1.00/0.40,0.97/0.91).这是由于不同数据集中用户偏好往往不同,其最佳α分布也不同.

上述结果证明了不同用户对这两种信任机制的权重是不同的,同时也说明了本文提出算法的正确性.

4 相关工作

4.1 传统的推荐算法

传统的推荐算法[5,6]主要有基于内容的推荐算法[5]、基于协同过滤的推荐算法[6]以及混合推荐算法[24,25].其中,基于内容的推荐算法利用用户历史信息构建用户配置文件,进而结合项目内容信息为用户推荐与其最相似的项目.而基于协同过滤的推荐算法分为基于内存的协同过滤算法和基于模型的协同过滤算法.

- 基于内存的协同过滤算法借助用户对项目的评分信息,寻找与待推荐用户最相似的邻居用户集,并将这些邻居用户对项目的评分作为待推荐用户对项目评分的参考,常见的方法有基于用户的协同过滤算法、基于项目的协同过滤算法等;
- 基于模型的协同过滤算法通过训练集获得模型参数,从而利用该模型为用户进行推荐.常见的方法有矩阵分解模型[19]、聚类模型[26]、概率模型[27]等.其中,Salakhutdinov 等人[19]提出的 PMF 模型对用户特征、项目特征以及用户评分赋予高斯先验,并利用低维矩阵分解方法获得用户偏好以及项目属性,从而实现了推荐效果的有效性以及推荐算法的高效性.因此,PMF 模型得到了广泛的关注和应用.

然而,传统的推荐算法只考虑了用户对项目的评分信息,忽略了用户间的信任关系对提高推荐性能的巨大作用,许多学者以此为契机,对基于社交信任的推荐算法进行了相应的研究.

4.2 基于社交信任的推荐算法

基于社交信任的推荐算法通过在原有的推荐模型中加入社交网络中的信任信息,使得用户间的兴趣偏好以及相互影响关系得以进一步被挖掘,从而提高了推荐效果[28].

Ma 等人[10]将因子分析方法引入到 PMF 模型中,认为用户间的信任关系可以通过用户特征向量与潜在因子特征向量的乘积得到.该作者又将好友信任关系的影响扩展到评分影响上,进而提出了 STE 模型[2].作者认为信任用户对项目的评分会影响用户对该项目的最终评分,从而综合考虑用户自身评分以及信任用户的评分后给出最终评分.实验证明,该模型优于当时已有的算法.Jamali 等人[3]指出,文献[2]中提出的 STE 模型所考虑的社交信任关系对待推荐用户的影响仅仅是直接的评分影响,而这种影响并不具备传播性.事实上,信任用户之间的影响可以通过影响用户特征从而间接地影响整个社交网络.作者基于这方面的考虑,提出了 SocialMF 模型.实验结果表明,该模型相比于 STE 模型有着明显的提升.Yang 等人[14]在文献[3]的基础上,认为用户在不同场景下对好友的信任程度并不相同,在该场景中,更为专业的好友会得到用户更多的信任.基于此,作者提出了多种在不同场景下计算用户间信任关系的方法,并将新的信任权重应用到 SocialMF 模型中.Xu 等人[29]将社交关系引入推荐用户建模,直接将用户属性表征为好友之间所具有的共同偏好,从而更好地建模不同社交场景下用户所体现出的不同属性.该作者在此基础之上,进一步考虑了用户社交关系的主题敏感性与局部稠密性,从而借助快速捕捉社交互动行为中的稠密子结构,提升了算法的性能[30].

此外,许多工作如在推荐算法中加入时间[8]等信息,也都是在基于社交信息的推荐算法的基础之上展开的,这在另一方面也印证了研究基于社交信任的推荐算法的重要性.

与上述相关工作不同,本文通过结合信任用户间接影响用户偏好和信任用户直接影响评分结果这两种不

同机制得到 TPMF 模型,并针对不同用户受两种机制影响权重不同的问题,实现了用户模型参数的个性化选择,进而提出了新的推荐框架.

5 结束语

本文基于概率矩阵分解模型,从社交关系影响属性的间接影响和社交关系影响决策的直接影响这两种机制入手,深入研究了用户的信任关系对其评分行为产生的影响,从而提出了基于用户信任关系的概率矩阵分解模型.进而,本文针对上述基本模型对所有用户均使用固定权值融合信任关系的两种不同影响机制这一缺陷,通过分析用户的评分相关性实现用户聚类,并设计个性化权值算法,最终提出了完整的个性化用户评分预测系统框架.通过 Epinions 以及 Ciao 公开数据集上的大量实验表明,本文提出的基于用户信任关系的概率矩阵分解模型能够较好地提升推荐准确率.同时,实验与讨论证实,本文所研究的几个因素(评分影响、特征影响以及个性化权值)均对最终的推荐结果起到了关键作用.

由于用户间的信任关系及用户的兴趣、倾向均随时间发生演变,未来工作将主要围绕如何捕捉、建模这一演变规律并实现准确预测展开.此外,借助同质性原理可以发现,用户间的评分行为(兴趣偏好)会影响用户间的信任关系.因此,借助时间信息同样可以刻画用户兴趣偏好以及用户间的信任关系动态演化的过程,这也是今后工作的核心内容.同时,未来工作还将研究如何引入用户信任关系在不同领域下的不同权重,从而进一步完善本文所提出的基于信任关系的用户评分预测技术.

References:

[1] Liu Q, Zeng X, Zhu HS, *et al*. Mining indecisiveness in customer behaviors. In: Proc. of the 2015 IEEE Int'l Conf. on Data Mining (ICDM). IEEE, 2015. 281−290.

[2] Ma H, King I, Lyu MR. Learning to recommend with social trust ensemble. In: Proc. of the 32nd Int'l ACM SIGIR Conf. on Researchand Development in Information Retrieval. New York: ACM Press, 2009. 203−210.

[3] Jamali M, Ester M. A transitivity aware matrix factorization model for recommendation in social networks. In: Proc. of the 22th Int'l Joint Conf. on Artificial Intelligence. Burlington: Morgan Kaufmann Publishers, 2011. 2644−2649.

[4] Liu Q, Ge Y, Li Z, *et al*. Personalized travel package recommendation. In: Proc. of the 2011 IEEE 11th Int'l Conf. on Data Mining (ICDM). IEEE, 2011. 407−416.

[5] Lops P, De Gemmis M, Semeraro G. Content-Based Recommender Systems: State of the Art and Trends. Springer-Verlag, 2011. 73−105.

[6] Schafer J, Frankowski D, Herlocker J, *et al*. Collaborative filtering recommender systems. The Adaptive Web, 2007. 291−324.

[7] Ricci F, Rokach L, Shapira B. Introduction to Recommender Systems Handbook. Springer-Verlag, 2011.

[8] Guo L, Ma J, Chen ZM. Trust strength aware social recommendation method. Journal of Computer Research and Development, 2013,50(9):1805−1813 (in Chinese with English abstract).

[9] Li YS, Song MN, Hai-Hong E, *et al*. Social recommendation algorithm fusing user interest social network. Journal of China Universities of Posts & Telecommunications, 2014,21(14):26−33.

[10] Ma H, Yang H, Lyu MR, *et al*. Sorec: Social recommendation using probabilistic matrix factorization. In: Proc. of the 17th ACM Conf. on Information and Knowledge Management. New York: ACM Press, 2008. 931−940.

[11] Tang J, Gao H, Hu X, *et al*. Exploiting homophily effect for trust prediction. In: Proc. of the ACM Int'l Conf. on Web Search and Data Mining. 2013. 53−62.

[12] Xu T, Zhong H, Zhu HS, *et al*. Exploring the impact of dynamic mutual influence on social event participation. In: Proc. of the 2015 SIAM Int'l Conf. on Data Mining. Philadelphia: SIAM, 2015. 262−270.

[13] Xu T, Zhu HS, Zhao X, *et al*. Taxi driving behavior analysis in latent vehicle-to-vehicle networks: A social influence perspective. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2016. 1285−1294.

[14] Yang XW, Steck H, Liu Y. Circle-Based recommendation in online social networks. In: Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2012. 1267−1275.

[15] Tang J, Hu X, Gao H, *et al*. Exploiting local and global social context for recommendation. 2013.

[16] Massa P. A survey of trust use and modeling in real online systems. Trust E-Services: Technologies, Practices and Challenges. Idea Group Inc, 2007. 51−83.

[17] Leskovec J, Huttenlocher D, Kleinberg J. Predicting positive and negative links in online social networks. In: Proc. of the Int'l Conf. on World Wide Web. ACM Press, 2010. 641−650.

[18] Page L. The PageRank Citation Ranking: Bringing Order to the Web. Stanford InfoLab, 1998. 1−14.

[19] Mnih A, Salakhutdinov RR. Probabilistic matrix factorization. In: Proc. of the Advances in Neural Information Processing Systems. 2008. 1257−1264.

[20] Oestreicher-Singer G, Sundararajan A. Recommendation networks and the long tail of electronic commerce. 2010.

[21] Grover A, Leskovec J. node2vec: Scalable feature learning for networks. In: Proc. of the ACM SIGKDD Int'l Conf. 2016. 855−864.

[22] Beigi G, Tang J, Wang S, *et al*. Exploiting emotional information for trust/distrust prediction. In: Proc. of the Siam Int'l Conf. on Data Mining. 2016.

[23] Jiang M, Cui P, Liu R, *et al*. Social contextual recommendation. In: Proc. of the 21st Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2012. 45−54.

[24] Moreno A, Ariza-Porras C, Lago P, *et al*. Hybrid Model Rating Prediction with Linked Open Data for Recommender Systems Semantic Web Evaluation Challenge. Springer Int'l Publishing, 2014. 193−198.

[25] Son LH. HU-FCF: A hybrid user-based fuzzy collaborative filtering method in recommender systems. Expert Systems with Applications: An Int'l Journal, 2014,41(15):6861−6870.

[26] Gao M, Cao F, Huang JZ. A cross cluster-based collaborative filtering method for recommendation. In: Proc. of the 2013 IEEE Int'l Conf. on Information and Automation. Piscataway: IEEE, 2013. 447−452.

[27] Getoor L, Sahami M. Using probabilistic relational models for collaborative filtering. In: Proc. of the Workshop on Web Usage Analysis and User Profiling (WEBKDD'99). 1999.

[28] Meng XW, Liu SD, Zhang YJ, *et al*. Research on social recommender systems. Ruan Jian Xue Bao/Journal of Software, 2015, 26(6):1356−1372 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4831.htm [doi: 10.13328/j.cnki.jos.004831]

[29] Xu T, Liu D, Chen EH, *et al*. Towards annotating media contents through social diffusion analysis. In: Proc. of the 12th Int'l Conf. on Data Mining (ICDM 2012). Brussels, 2012. 1158−1163.

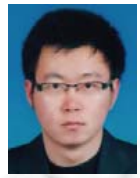
[30] Xu T, Zhu HS, Chen EH, *et al*. Learning to annotate via social interaction analytics. Knowledge and Information Systems, 2014, 41(2):251−276.

附中文参考文献:

[8] 郭磊,马军,陈竹敏.一种信任关系强度敏感的社会化推荐算法.计算机研究与发展,2013,50(9):1805−1813.

[28] 孟祥武,刘树栋,张玉洁,等.社会化推荐系统研究.软件学报,2015,26(6):1356−1372. http://www.jos.org.cn/1000-9825/4831.htm [doi: 10.13328/j.cnki.jos.004831]

杜东舫(1992－),男,浙江宁波人,硕士,主要研究领域为推荐系统,社交网络.



管楚(1989－),男,博士,主要研究领域为用户数据分析,机器学习.

徐童(1988－),男,博士,副研究员,CCF 专业会员,主要研究领域为数据挖掘.

刘淇(1986－),男,博士,副教授,CCF 专业会员,主要研究领域为数据挖掘与知识发现,机器学习方法与应用.

鲁亚男(1996－),女,硕士,主要研究领域为计算机视觉,数据挖掘.

陈恩红(1968－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为数据挖掘,机器学习.