计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20150307
53(6):1389-1399,2016

一种基于多元社交信任的协同过滤推荐算法

王瑞琴[1] 蒋云良[1] 李一啸[2] 楼俊钢[1]
[1](湖州师范学院信息工程学院 浙江湖州 313000)
[2](浙江财经大学信息管理与工程学院 杭州 310018)
(angelwrq@163.com)

A Collaborative Filtering Recommendation Algorithm Based on Multiple Social Trusts

Wang Ruiqin[1], Jiang Yunliang[1], Li Yixiao[2], and Lou Jungang[1]
[1](*School of Information Engineering, Huzhou University, Huzhou, Zhejiang* 313000)
[2](*School of Information Management and Engineering, Zhejiang University of Finance & Economics, Hangzhou* 310018)

Abstract Collaborative filtering (CF) is one of the most successful recommendation technologies in the personalized recommendation systems. It can recommend products or information for target user according to the preference information of similar users. However the traditional collaborative filtering algorithms have the disadvantages of low recommendation efficiency and weak capacity of attack-resistance. In order to solve the above problems, a novel collaborative filtering algorithm based on social trusts is proposed. Firstly, referring to the trust generation principle in social psychology, a social trust computation method based on multiple trust elements is presented. In social networking environment, trust elements mainly include credibility, reliability, intimacy and self-orientation. Then specific methods of identifying, extraction and quantification of the trust elements are studied in depth. Finally, the trustworthy neighbors of target user are selected in accordance with the social trust, so as to make trust-based collaborative recommendation. Using the FilmTrust and Epinions as test data sets, the performance of the novel algorithm is compared with that of the traditional CF and the-state-of-art methods, as well as the CF based on single trust element. Experimental results show that compared with the other methods, the proposed algorithm not only improves the recommendation precision and recall, but also has powerful attack-resistance capacity.

Key words collaborative filtering (CF); social network; trust; trust elements; recommendation precision; recall; attack-resistance capacity

摘 要 协同过滤推荐是当前最成功的个性化推荐技术之一,但是传统的协同过滤推荐算法普遍存在推荐性能低和抗攻击能力弱的问题.针对以上问题,提出了一种基于多元化社交信任的协同过滤推荐算法CF-CRIS (collaborative filtering based on credibility, reliability, intimacy and self-orientation). 1)借鉴社会心理学中的信任产生原理,提出基于多个信任要素(可信度、可靠度、亲密度、自我意识导向)的

收稿日期:2015-04-20;修回日期:2015-10-13
基金项目:国家自然科学基金项目(61402336,61370173,61403338);国家教育部科学基金项目(14YJCZH152);浙江省自然科学基金项目(LY15F020018);浙江省科技计划项目(2013C31138,2015C33247)
This work was supported by the National Natural Science Foundation of China (61402336, 61370173, 61403338), the Science Foundation of Ministry of Education of China (14YJCZH152), the Natural Science Foundation of Zhejiang Province of China (LY15F020018), and the Science and Technology Planning Project of Zhejiang Province of China (2013C31138, 2015C33247).

信任度计算方法;2)深入研究社交网络环境中各信任要素的识别、提取和量化方法;3)基于用户间的综合信任度选取可信邻居,完成对目标用户的个性化推荐. 基于通用测试数据集的实验研究结果表明:该算法不但可以极大地提高推荐系统的精确度和召回率,而且表现出良好的抗攻击能力.

关键词 协同过滤;社交网络;信任;信任要素;推荐精度;召回率;抗攻击能力

中图法分类号 TP391

随着计算机和 Internet 技术的快速发展,电子商务技术日趋成熟与壮大,人们的日常生活已离不开它. 在电子商务环境下最重要的是信任(trust),然而信任问题仍然是当前电子商务中的一个未解难题[1]. 在社会科学中,信任被认为是一种作用于个人或团体的依赖关系[2]. 信任具有主观性、非对称性、传播性、可组合性、自我加强性和事件敏感性等特征. 传统的推荐技术通常假设用户是独立和恒等分布的,经常忽略用户间基于社会关系产生的信任. 然而,在朋友推荐和系统推荐之间做出选择时,无论从推荐的质量还是推荐的有效性来看,用户往往更倾向于前者,所以提取和量化用户间的信任关系是改善推荐质量的法宝.

社交网络是人们在线交流的平台,也是信息传播的媒介,这一想法激发了基于信息传播的信任推理方法的研究与发展. 根据推理过程的不同,可以分为基于社交网络结构的信任模型和基于社会交互的信任模型 2 大类[3].

基于社交网络结构的信任模型通常建立在 FOAF(friend-of-a-friend)基础上. Paolo 等人[4]基于 Epinions 数据集中的信任数据建立了预测模型 MoleTrust,以此对社交网络中用户间的信任度进行预测,并基于信任进行推荐. Golbeck[5]提出基于最短信任路径的信任推理模型 TidalTrust,该模型计算复杂度低,具有良好的可扩展性. Zhang 等人[6]对 TidalTrust 模型进行扩展,引入信任评价和可信度因子,使用加权图结构计算实体间的信任度. Kuter 等人[7]根据来自不同信任链的信息,提出基于贝叶斯网络的信任推理模型 SUNNY. Kim[8]根据用户在特定领域的专业程度和用户间的熟悉程度构建信任网络,并进行信任度的计算. Caverlee 等人[9]基于社会关系和用户反馈进行信任计算和推理,提出基于信誉的动态信任推理模型 SocialTrust,该模型具有很好的鲁棒性. Zuo 等人[10]提出利用信任图中的信任链计算信任度的方法,同时考虑了信任的组合问题. Hang 等人[11]提出基于图结构的信任度量方法,并利用信任图的相似性进行节点推荐. Seth 等人[12]分析了社交网络中常见的信任传播模式,提出基于概率图模型的信任推理算法. 邢星[13]在信任网络中定义了串联和并联 2 种信任组合路径,分别定义不同的传递算子来计算用户间的间接信任度. Jiang 等人[14]利用小世界网络特性从大型社交网络中提取基于用户领域的小型信任网络,计算复杂度更低,结果更客观、更稳定.

在基于社会交互的信任计算模型方面,Liu 等人[15]利用在线社区中用户的交互行为预测用户的信任度. 清华大学的课题组[16]和武汉大学的课题组[17]利用博弈原理,基于用户间的交互行为研究用户间的信任关系. Adali 等人[18]基于交流信任和传播信任计算用户在社交网络中的信任度. 北京邮电大学的课题组[19]借鉴社会心理学中人与人之间的信任产生原理,基于用户交互图和用户-项目评分进行信任计算. Nepal 等人[20]提出一个基于交互的信任模型 STrust,融合了流行信任度和参与信任度 2 种类型的信任. Kim 等人[21]基于用户评价衡量用户对于相关主题的专业程度和用户的兴趣,进而计算用户间的信任度. Li 等人[22]利用用户评分的相似性和用户间由交互产生的熟悉性综合度量用户间的信任度. Nikolay 等人[23]从评分数据中提取一系列关键特征用于用户间信任度的计算,并使用分类算法度量每个特征在信任计算中的重要性. Emanuel 等人[24]基于多数据源下的交互信息,定义了一系列基于内容和基于网络的用户相似性度量方法进行协同推荐.

综上所述,研究者们在信任计算和基于信任的推荐方面取得了一定的研究成果,然而已有方法主要关注用户之间显式的信任关系,很多有价值的隐式信任关系往往被忽略,对于信任传播和信任融合方面的研究也不够深入. 另外,已有研究几乎都是从技术角度来分析信任问题,缺乏社会心理学方面的理论指导,这些问题都有待深入研究.

本文通过对社会心理学中信任产生原理的学习,提出用户间社交信任的计算方法,并深入研究社交网络环境中各信任要素的提取、量化和集成方法,

进而提出一种基于社交信任的协同过滤(collaborative filtering, CF)推荐算法.本文的创新点主要体现在4个方面:

1)借鉴社会心理学中的信任产生原理,在社交信任计算中综合考虑多个信任要素,具有一定的理论基础;

2)充分利用用户-项目评分数据和用户间的初始信任关系,提取信任要素并对其进行量化,准确度量用户间的综合社交信任度;

3)提出基于多元社交信任的协同过滤推荐算法,利用用户间的综合信任关系选取推荐邻居,通过推荐邻居的评分信息,实现对目标用户的推荐;

4)在FilmTrust和Epinions数据集上进行了实验验证,结果表明,和同类方法相比,本文算法不但提高了推荐的精度和召回率,而且在抗攻击能力方面表现良好.

1 多元社交信任的理论基础

1.1 信任产生原理

在《The Trusted Advisor》一书中,David等人[25]借鉴社会心理学中人们之间信任产生的过程,提出一个商业领域的信任计算公式:

$$TR=(C\times K\times I)/S, \tag{1}$$

其中,TR表示用户间的信任度,C表示用户的可信度(credibility),K表示用户的可靠度(reliability),I表示用户间的亲密度(intimacy),S表示用户的自我意识导向(self-orientation).

需要指出的是,式(1)并不是一个严格的用于信任计算的数学公式,而是指明信任计算中前3个要素C,K,I是正向指标,最后1个要素S是反向指标.

1.2 相关定义

在协同过滤推荐系统中,用户评分数据由m个用户组成的用户集$U=\{u_1,u_2,\cdots,u_m\}$,n个项目组成的项目集$T=\{t_1,t,\cdots,t_n\}$和一个$m\times n$阶评分矩阵$\boldsymbol{R}$组成,其中R_{ij}表示用户u_i对项目t_j的评分值,如果用户u_i对项目t_j没有评价,则$R_{ij}=0$.

一些电子商务社交平台允许人们显式地指定信任关系,信任关系通常用$m\times m$阶的信任矩阵$\boldsymbol{DT}$表示,DT_{ij}表示用户u_i对用户u_j的信任程度,$DT_{ij}\in\{0,1\}$,0表示不信任,1表示信任.信任矩阵也可以使用一个信任网络图$G=(V,E)$来表示,其中节点V表示用户,边E表示用户间具有信任关系,边的权重表示信任的程度.

定义1. 候选邻居用户集H.给定一个历史评分矩阵$\boldsymbol{R}$,当前用户$u_i\in U$,当前项目$t_k\in T$,此时$R_{ik}=0$.如果存在$u_j\in U$,使得$R_{jk}\neq 0$,那么称用户u_j是用户u_i在项目t_k上的一个候选邻居.则u_i在项目t_k上的候选邻居集$H_k(u_i)$表示为

$$H_k(u_i)=\{u_j\mid R_{ik}=0\wedge R_{jk}\neq 0,u_j\in U,t_k\in T\}. \tag{2}$$

定义2. 共评项目集CI.给定用户u_i和用户u_j,如果存在$t_k\in T$,使得$R_{ik}\neq 0$且$R_{jk}\neq 0$,那么就说项目t_k是用户u_i和用户u_j的一个共评项目.则用户u_i和用户u_j的共评项目集CI_{ij}表示为

$$CI_{ij}=\{t_k\mid R_{ik}\neq 0\wedge R_{jk}\neq 0,t_k\in T\}. \tag{3}$$

定义3. 用户的可信度C.给定用户u_i和用户u_j,如果$CI_{ij}\neq\varnothing$,那么可以根据用户u_i和用户u_j在CI_{ij}上的评分相似度$sim(u_i,u_j)$来衡量用户间的可信度.则用户u_i和用户u_j间的可信度表示为

$$C(u_i,u_j)=sim(u_i,u_j),\ \text{if}\ CI_{ij}\neq\varnothing. \tag{4}$$

定义4. 用户的可靠度K.给定用户u_i和用户u_j,如果$CI_{ij}\neq\varnothing$,那么可以使用用户u_j在CI_{ij}上对用户u_i预测评分的推荐准确率$PR(u_i,u_j)$作为用户u_i对用户u_j可靠性的度量.则用户u_i对用户u_j可靠性的度量表示为

$$K(u_i,u_j)=PR(u_i,u_j),\ \text{if}\ CI_{ij}\neq\varnothing. \tag{5}$$

定义5. 用户间的亲密度I.给定一个初始信任矩阵$\boldsymbol{DT}$,用户$u_i\in U$,用户$u_j\in U$,如果$DT_{ij}\neq 0$或者通过信任推理后用户u_i和用户u_j间的间接信任度$IT_{ij}\neq 0$,那么就说用户u_i和用户u_j之间具有一定的社会亲密度.则用户u_i和用户u_j间的亲密度表示为

$$I(u_i,u_j)=\begin{cases}DT(u_i,u_j),\ \text{if}\ DT_{ij}\neq 0,\\ IT(u_i,u_j),\ \text{if}\ DT_{ij}=0\wedge IT_{ij}\neq 0.\end{cases} \tag{6}$$

定义6. 用户的自我意识导向S.给定一个信任网络G,当前用户$u_i\in U$,那么可以使用用户u_i在G中全局信誉度$rep(u_i)$的倒数来衡量u_i的自我意识导向.则用户u_i的自我意识导向表示为

$$S(u_i)=\frac{1}{rep(u_i)}. \tag{7}$$

2 多元社交信任的计算方法

2.1 用户间可信度(C)的计算

可信度指的是人们给出的可以证明自己没有言过其实的种种信号.比如他们的确拥有自己声称拥有的资质,或者他们的职业水平确实如他们自称的一样出色.某人的可信度越高,你就越是可以相信

他.现实世界中“物以类聚、人以群分”的自然选择规律在虚拟的社交网络中同样适用,研究表明,兴趣爱好越相似的用户之间的信任程度越高,我们称之为相似信任.因此,本文基于用户间的评分相似度来衡量用户间的可信度,用户间的评分相似度可以采用Pearson相关系数来度量:

$$s(u_i,u_j)=\frac{\sum_{t_k\in CI_{ij}}(R_{ik}-\bar{R}_i)\times(R_{jk}-\bar{R}_j)}{\sqrt{\sum_{t_k\in CI_{ij}}(R_{ik}-\bar{R}_i)^2}\times\sqrt{\sum_{t_k\in CI_{ij}}(R_{jk}-\bar{R}_j)^2}},\quad(8)$$

其中,R_{ik}和R_{jk}分别表示用户u_i和用户u_j对项目t_k的评分值,$\bar{R}_i$和$\bar{R}_j$分别表示用户u_i和用户u_j在所有项目上的评分均值,$CI_{i,j}$表示用户u_i和用户u_j的共评项目集.

假设$s(u_i,u_j)=s(u_i,u_k)$,但是$|CI_{ij}|>|CI_{ik}|$,即用户u_i和用户u_j之间的共评项目数大于用户u_i和用户u_k之间的共评项目数,显然此时用户u_i和用户u_j间的评分相似度应该比用户u_i和用户u_k间的评分相似度大.下面利用用户间的共评项目数$|CI_{i,j}|$对评分相似度的计算公式进行优化:

$$sim(u_i,u_j)=s(u_i,u_j)\times\frac{1}{1+e^{-\frac{|CI_{ij}|}{2}}},\quad(9)$$

式(9)使用指数函数避免$|CI_{i,j}|$过大对相似度计算结果造成的影响,使得用户相似度落在[0,1]区间内.当$|CI_{i,j}|$足够大时,式(9)的右项值趋于1;对于很小的$|CI_{i,j}|$,该项的值约为0.6;当$|CI_{i,j}|>5$时,该项的值大于0.9.

2.2 用户可靠度(*K*)的计算

简单地讲,可靠度指一个人做事情的靠谱程度.在电子商务推荐系统中,用户的可靠度就是用户推荐的准确度,比如他们越是经常性地向你推荐你喜欢的商品,那么你越是有理由相信该用户未来的推荐也是可靠的.因此,本文通过计算用户的推荐准确率来评估其可靠性.

将用户$u_j\in H_k(u_i)$作为用户u_i的唯一推荐用户,对于$t_k\in CI_{ij}$,根据以下公式对目标用户u_i进行评分预测:

$$P_{ik}=\bar{R}_i+\frac{(R_{jk}-\bar{R}_j)\times sim(u_i,u_j)}{|sim(u_i,u_j)|},\quad(10)$$

其中,P_{ik}表示用户u_i对项目t_k的预测评分,R_{jk}表示用户u_j对t_k的真实评分,$\bar{R}_i$和$\bar{R}_j$分别表示用户u_i和用户u_j的评分均值,$sim(u_i,u_j)$表示用户u_i和用户u_j之间的评分相似度.

根据实际评分值与预测评分值之间的差异程度,得到用户u_i对用户u_j推荐能力的计算公式如下:

$$pr_{ij}^k=1-\frac{|P_{ik}-R_{ik}|}{P_{\max}},\quad(11)$$

其中,pr_{ij}^k表示目标用户u_i对推荐用户u_j在项目t_k上的推荐能力的估计值,P_{ik}表示用户u_i对项目t_k的预测评分,R_{ik}表示用户u_i对t_k的实际评分,$P_{\max}$表示预测评分与实际评分差异的极大值.

目标用户u_i对推荐用户u_j推荐准确率的计算为

$$PR(u_i,u_j)=\frac{\sum_{k=1}^{|CI_{ij}|}pr_{ij}^k}{|CI_{ij}|},\quad(12)$$

其中,$|CI_{ij}|$表示用户u_i和用户u_j的共评项目数.

2.3 用户间亲密度(*I*)的计算

社交网络的主体是用户,用户能创建和维护与其他用户之间的朋友关系,具有朋友关系的用户之间具有较强的亲密度.朋友关系具有传递性,所谓“friend of a friend is a friend”就是这个道理.亲密度是信任中最强有力的情感因素之一,是信任计算中不可忽视的一个组成部分.

如定义5所述,社会亲密度由直接信任度或间接信任度来表征,这里借鉴文献[19]中提出的方法计算用户间的间接信任度.给定用户间的初始信任网络G,如图1所示,要计算当前用户u_i和其他用户间的间接信任度.首先以用户u_i为起点,将与用户u_i有直接信任关系的所有用户排列在用户u_i的周围;再将这些用户直接信任的用户排在以用户u_i为圆心的第2层,以此类推,形成一系列以用户u_i为圆心的同心圆;为了获取最短信任路径,只保留不同层节点之间的连接边,得到目标节点的信任网络图G',如图2所示,此时第1层节点(u_1,u_2,u_3,u_4)为用户u_i的朋友,第2层节点(u_5,u_6,u_7,u_8)为用户u_i的朋友(u_1,u_2,u_3,u_4)的朋友,以此类推.

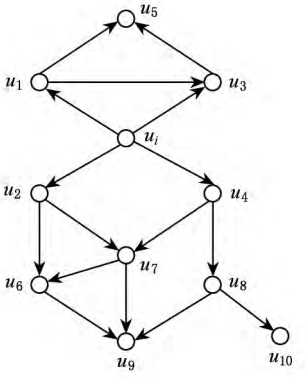

Fig. 1 Initial trust network of users.

图1 用户间的初始信任网络

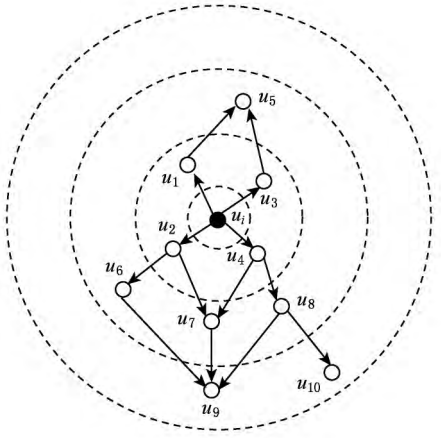

Fig. 2 Trust network of the target user node.

图 2 目标用户节点信任网络示意图

采用以下公式计算当前用户 u_i 对处于 G' 中第 2 层及以上的用户 u_j 的间接信任度:

$$IT(u_i,u_j)=\frac{1}{2^{L_j-1}}\times\frac{1}{1+e^{-\frac{n}{2}}}, \tag{13}$$

其中,$IT(u_i,u_j)$ 表示用户 u_i 与用户 u_j 的间接信任度,L_j 为用户 u_j 所在的层,n 表示从用户 u_i 到用户 u_j 共有 n 条路径. 以上信任推理过程同时考虑了信任路径的长度和多信任路径的组合问题.

例如对于用户节点 u_7,它处于第 2 层,所以 $L_7=2$;从 u_i 到 u_7 有 2 条路径($u_i\rightarrow u_2\rightarrow u_7$ 和 $u_i\rightarrow u_4\rightarrow u_7$),所以 $n=2$;则用户 u_i 与用户 u_7 的间接信任度为 $(1/2^1)(1/(1+e^{-1}))\approx 0.37$.

2.4 用户自我意识导向(S)的计算

人际网络建设的核心要素在于乐于助人,构建不以交换原则为基础的新型人际关系. 社交网络亦是如此,使用自己的网络来解决问题,通过给他人提供商业机会来构筑杠杆原理. 所以说,自我意识导向是信任中的负面因素,一个人的自我意识导向越是强烈,人们越是无法信任此人. 比如一个只对自己感兴趣、全然不在意他人的感受,这样的人就是自我意识导向强烈者的范例之一;而愿意推荐竞争对手的好产品而不是坚持自家产品垄断,这样的人就有着较低的自我意识导向.

本文基于用户在社交信任网络中的全局信誉度(reputation)来衡量用户自我意识导向的强度,用户信誉度越高,其自我意识导向程度越弱,越值得信任. 给定信任网络 G,目标用户 $u_i\in U$,用户 u_i 的信誉度 $rep(u_i)$ 与 G 中信任用户 u_i 的用户数和这些用户自身的信誉度密切有关,本文采用 PageRank 算法来计算用户的信誉度:

$$rep(u_i)=\frac{1-q}{m}+q\times\sum_{u_j\in TU(u_i)}\frac{rep(u_j)}{|TN(u_j)|}, \tag{14}$$

其中,m 是信任网络中的用户数量,$TU(u_i)$ 是信任用户 u_i 的用户集,$rep(u_j)$ 是用户 u_j 的信誉度,$|TN(u_j)|$ 是用户 u_j 的信任用户数,q 是调和因子.

2.5 基于信任四要素的社交信任计算方法

完成所有信任要素的提取和量化之后,首先进行各信任要素的归一化操作,然后采用线性合并的方法(测试多种合并方式后发现此方法性能最好)得到用户间的综合社交信任程度:

$$TR(u_i,u_j)=w_1\times C(u_i,u_j)+w_2\times K(u_i,u_j)+w_3\times I(u_i,u_j)-w_4\times S(u_j), \tag{15}$$

其中,$TR(u_i,u_j)$ 表示用户 u_i 对用户 u_j 的综合社交信任度;$C(u_i,u_j)$ 表示用户 u_i 对用户 u_j 的可信度的度量值;$K(u_i,u_j)$ 表示用户 u_i 对用户 u_j 的可靠度的度量值;$I(u_i,u_j)$ 表示用户 u_i 和用户 u_j 之间的亲密度的度量值;$S(u_j)$ 表示用户 u_j 的自我意识导向的度量值;w_1,w_2,w_3,w_4 为各信任要素的权重因子,满足 $w_1+w_2+w_3+w_4=1$,具体权重分配采用实验训练方法得到.

3 基于多元社交信任的协同推荐算法

传统协同推荐方法基于相似用户对目标项目的历史评分来估计当前用户对目标项目的喜好程度,用户间的相似度基于历史评分矩阵计算得到. 在实际应用中,随着系统规模的不断扩大,用户-项目评分矩阵会变得越来越稀疏,导致用户相似性的计算结果很不准确,从而影响推荐质量. 另外,在面对用户概貌注入攻击(profile injection attacks)时,基于用户相似度的协同过滤算法抗攻击能力较差. 针对上述问题,本文提出一种基于多元社交信任的协同推荐算法 CF-CRIS (collaborative filtering based on credibility, reliability, intimacy and self-orientation),其核心思想如下:

1) 针对目标项目 t_k,选取目标用户 u_i 的候选邻居集 $H_k(u_i)$;

2) 分别利用式(4)~(7)计算每个候选邻居的可信度 C、可靠度 K、亲密度 I 和自我意识导向 S,然后利用式(15)将以上 4 个信任要素进行合并,得到目标用户对候选用户的综合信任度 TR;

3) 按照综合信任度对候选用户进行降序排列,选取前 k 个信任度最大的用户作为目标用户的推荐邻居 $knn(u_i)$;

4) 根据推荐邻居对目标项目 t_k 的评分信息,采用基于社交信任的协同过滤方法计算目标用户 u_i 对目标项目 t_k 的预测评分:

$$P_{ik}=\bar{R}_i+\frac{\sum_{u_j\in knn(u_i)}(R_{jk}-\bar{R}_j)\times TR(u_i,u_j)}{\sum_{u_j\in knn(u_i)}TR(u_i,u_j)},\quad(16)$$

其中,P_{ik}表示目标用户u_i对目标项目t_k的预测评分,$knn(u_i)$表示目标用户u_i的Top-k推荐邻居集,R_{jk}表示推荐邻居u_j对目标项目t_k的评分,$\bar{R}_i$和$\bar{R}_j$分别表示用户u_i和邻居u_j的评分均值,$TR(u_i,u_j)$表示目标用户u_i和信任邻居u_j之间的综合信任度.

根据以上算法思想,给出算法CF-CRIS的伪代码描述如下:

算法1. CF-CRIS.

输入:评分矩阵$\boldsymbol{R}$、初始信任矩阵$\boldsymbol{DT}$、目标用户$u_i\in U$、目标项目$t_k\in T$;

输出:用户u_i对项目t_k的预测评分P_{ik}.

① $knn(u_i)=\varnothing$;

② $H_k(u_i)\leftarrow\{u_j\mid R_{ik}=0\wedge R_{jk}\neq 0,u_j\in U\}$;

③ for each $u_j\in H_k(u_i)$ do

④ $C(u_i,u_j)=sim(u_i,u_j)$;

⑤ $K(u_i,u_j)=P(u_i,u_j)$;

⑥ if $DT_{ij}\neq 0$ then

⑦ $I(u_i,u_j)=DT(u_i,u_j)$;

⑧ else if $IT_{ij}\neq 0$ then

⑨ $I(u_i,u_j)=IT(u_i,u_j)$;

⑩ end if

⑪ $S(u_j)=1/rep(u_j)$;

⑫ $TR(u_i,u_j)=w_1\times C(u_i,u_j)+w_2\times K(u_i,u_j)+w_3\times I(u_i,u_j)-w_4\times S(u_j)$;

⑬ end for

⑭ 根据TR降序排列u_j;

⑮ $knn(u_i)\leftarrow\{TR$最大的前k个用户$\}$;

⑯ $P_{ik}=\bar{R}_i+\frac{\sum_{u_j\in knn(u_i)}(R_{jk}-\bar{R}_j)\times TR(u_i,u_j)}{\sum_{u_j\in knn(u_i)}TR(u_i,u_j)}$;

⑰ return P_{ik}.

算法1主要包括3个阶段:阶段1完成变量的初始化,并选取目标用户的候选邻居用户集H,对应行①~②;阶段2完成目标用户对候选邻居用户综合信任度TR的计算,对应行③~⑬;阶段3根据综合信任度选取Top-k信任邻居,利用信任邻居的评分信息预测目标用户对目标项目的评分值P_{ik},对应行⑭~⑰.

4 实验结果与分析

4.1 测试数据集

本文实验采用2个数据集:

1) FilmTrust站点[①]提供的数据集.该数据集是一个电影评分数据集,其中包括1 508个用户对2 071部电影的35 497次评分,评分范围为0.5~4,评分数据的稀疏度为98.86%;此外,该数据集还包括1 642个用户之间的1 853个显式信任关系,信任数据的稀疏度为99.93%.

2) Epinions站点[②]提供的数据集.该数据集是一个大众消费者点评数据集,其中包括49 290个用户对139 738件商品的664 824次评分,评分范围为1~5,评分数据的稀疏度为99.99%;此外,该数据集还包括49 290个用户之间的487 181个显式信任关系,信任数据的稀疏度为99.98%.

实验中,我们采用了5-cross交叉验证法,首先将原始评分数据集划分为互不相交的5组;然后对于每组数据,随机选取其中的10%作为训练集,进行参数w_1,w_2,w_3,w_4的估计,其余90%作为测试集,采用leave-one out方法进行评分预测;最后取5组测试的平均值作为实验结果.

4.2 性能评价指标

1) 平均绝对误差(mean absolute error, MAE). MAE是一个广泛用于评估推荐算法性能的重要参数,通过计算实际评分值与预测评分值之间的偏差得到. MAE的值越低说明推荐算法的精度越高.

$$MAE=\frac{\sum_{k=1}^{n}|P_k-R_k|}{n},\quad(17)$$

其中,P_k表示预测评分值,R_k表示真实评分值,n表示评分预测的次数.

2) 召回率(recall, RL). RL也叫查全率,指通过算法可以预测出来的评分数与所有待测评分数之间的比值.

$$RL=\frac{m}{n},\quad(18)$$

其中,m表示通过算法可以得到的预测评分数,n表示测试集中待测评分数.

① http://trust.mindswap.org

② http://www.epinions.com

3) **平均预测偏差**(average prediction shift, *APS*). *APS* 用来评价推荐算法的抗攻击能力,描述推荐算法受攻击前后预测性能的差异程度,*APS* 越小说明推荐算法的抗攻击能力越强. 单个项目的 *APS* 定义为

$$APS_k=\frac{1}{|U|}\times\sum_{u_i\in U}(P'_{ik}-P_{ik}), \tag{19}$$

其中,P_{ik} 和 P'_{ik} 分别表示受攻击前和受攻击后用户 u_i 对项目 t_k 的预测评分,U 表示用户集合. 在此基础上继续定义所有项目的 *APS* 均值为

$$\overline{APS}=\frac{1}{|T|}\times\sum_{t_k\in T}APS_k, \tag{20}$$

其中,T 表示项目集合.

4.3 推荐性能的比较

文献[26]提出一种基于双重邻居选取策略的协同过滤推荐算法 CF-DNC,与本文方法有类似之处. 该算法基于评分相似度选择目标用户的兴趣相似用户集,然后利用 leave-one-out 方法计算目标用户对兴趣相似用户的信任程度,以此作为选取可信邻居用户的依据. 与本文提出的 CF-CRIS 算法相比,CF-DNC 算法只考虑了信任的前 2 个要素,没有考虑用户间的亲密度和用户的自我意识导向.

为了评价推荐算法的精度,在同样的实验环境下,将本文提出的推荐算法(CF-CRIS)与传统的协同过滤推荐算法(CF)和 CF-DNC 算法进行实验比较. 另外,我们还和采用单一信任要素的协同推荐算法进行了对比,包括基于可信度的协同推荐算法(CF-C)、基于可靠度的协同推荐算法(CF-R)、基于亲密度的协同推荐算法(CF-I)和基于自我意识导向的协同推荐算法(CF-S). 采用 FilmTrust 和 Epinions 数据集,分别为目标用户选取不同的信任邻居个数(k)得到的推荐精度(*MAE*)对比结果如表 1 和表 2 所示.

从表 1 和表 2 中可以看出,在信任用户数 $k=15$ 和 $k=25$ 时,采用 FilmTrust 和 Epinions 数据集的推荐方法达到最佳推荐效果. 同时,不论采用哪种数据集,CF-CRIS 的推荐 *MAE* 值都明显小于 CF 算法和 CF-DNC 算法,以及基于单一信任要素的推荐算法 CF-C,CF-R,CF-I,CF-S. 这不仅说明基于多元社交信任的协同推荐算法可以改善推荐质量,而且表明本文提出的社交信任度量方法是可取的,因为该方法在信任计算中综合考虑了多个信任要素,所以推荐邻居的选择更加准确,从而获得了更高的推荐精度.

Table 1 Comparison of Recommendation Precision (*MAE*) with FilmTrust Dataset

表 1 采用 FilmTrust 数据集的推荐精度(*MAE*)对比

k	CF	CF-DNC	CF-C	CF-R	CF-I	CF-S	CF-CRIS
5	0.8456	0.6995	0.6894	0.6954	0.8954	0.7745	0.5812
10	0.8321	0.6847	0.6874	0.6845	0.8854	0.7658	0.5728
15	0.8012	0.6714	0.6670	0.6758	0.8651	0.7462	0.5643
20	0.8154	0.6875	0.6858	0.6854	0.8714	0.7541	0.5827
25	0.8256	0.6987	0.6912	0.7014	0.8987	0.7689	0.5904

Table 2 Comparison of Recommendation Precision (*MAE*) with Epinions Dataset

表 2 采用 Epinions 数据集的推荐精度(*MAE*)对比

k	CF	CF-DNC	CF-C	CF-R	CF-I	CF-S	CF-CRIS
10	0.8982	0.7412	0.7660	0.7213	0.9577	0.8419	0.6410
15	0.8798	0.7334	0.7584	0.7214	0.9451	0.8365	0.6446
20	0.8654	0.7254	0.7154	0.6989	0.9265	0.8275	0.6455
25	0.8362	0.7121	0.7110	0.6813	0.9177	0.8019	0.6431
30	0.8521	0.7321	0.7354	0.6987	0.9321	0.8254	0.6457

例如,采用 FilmTrust 数据集的实验,当推荐邻居数 $k=15$ 时,CF-CRIS 的推荐精度比 CF 和 CF-DNC 算法分别提高了大约 29%和 16%,比基于单一信任要素的推荐算法 CF-C,CF-R,CF-I,CF-S 分别提高了大约 15%,16%,35%,24%. 此外,CF-C 的推荐精度比 CF 算法提高了大约 17%,因为在用户相似度的计算过程中,CF-C 考虑了用户间的共评项目数,由此可见,共评项目数是度量用户偏好相似度的重要指标之一. 为了进一步评价推荐算法的召回率,将本文提出的 CF-CRIS 算法与传统的 CF 算法和 CF-DNC 算法,以及基于单一信任要素的推荐算法的召回率进行了实验比较,对比结果如图 3 和图 4 所示:

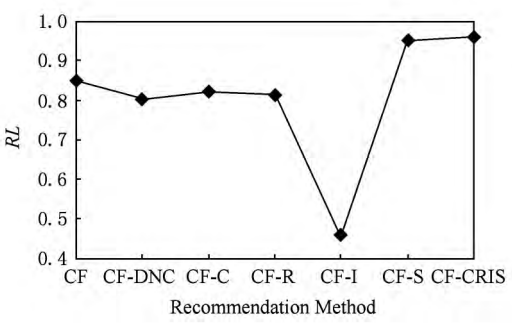

Fig. 3 Comparison of recall (*RL*) using FilmTrust dataset.

图 3 采用 FilmTrust 数据集召回率(*RL*)对比

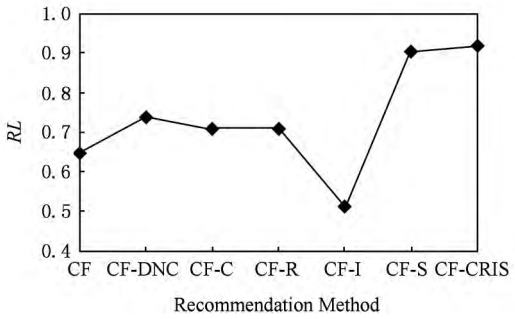

Fig. 4 Comparison of recall (*RL*) using Epinions dataset.

图 4 采用 Epinions 数据集召回率(*RL*)对比

从图 3 和图 4 可以看出,不论采用哪种数据集,算法 CF-CRIS 的召回率与 CF 算法、CF-DNC 算法以及基于单一信任要素的推荐算法 CF-C, CF-R, CF-I 和 CF-S 相比,性能相当或更好. 由此可见,在数据集极端稀疏的情形下,本文提出的推荐算法在提高推荐精度的同时,也获得了较好的召回率. 究其原因在于,社交信任的计算中综合考虑了多个信任要素,所以有效避免了数据稀疏性问题. 此外,不论采用哪种数据集,基于用户亲密度的推荐算法 CF-I 的 *RL* 值都非常低,这一方面是由于用户声明的信任关系非常稀疏,另一方面也说明本文提出的信任推理算法有待进一步优化.

4.4 抗攻击能力的比较

由于推荐系统固有的开放性和对用户信息的敏感性,使其非常容易受用户概貌注入型攻击的影响,从而影响推荐的质量. 为了对比本文算法 CF-CRIS 和传统推荐算法 CF 以及 CF-DNC 算法在抗攻击能力方面的性能,采用混合攻击方式人为地向原始数据集中注入恶意用户概貌信息. 推荐过程中,选取填充规模(filling size)为 1%, 3%, 5%, 10%, 攻击规模(attack size)为 1%, 2%, 3%, 5%, 在不同填充规模和攻击规模下, 3 种推荐算法的推荐精度对比结果如表 3 和表 4 所示:

Table 3 Comparison of Recommendation Precision (*MAE*) for Hybrid Attack when Using FilmTrust Dataset

表 3 混合攻击下基于 FilmTrust 数据集的推荐精度(*MAE*)对比

Filling Size	Attack Size											
	1%			2%			3%			5%		
	CF	CF-DNC	CF-CRIS	CF	CF-DNC	CF-CRIS	CF	CF-DNC	CF-CRIS	CF	CF-DNC	CF-CRIS
1%	0.8612	0.7241	0.6054	0.8753	0.7354	0.6142	0.8885	0.7485	0.6245	0.9015	0.7545	0.6341
3%	0.8684	0.7285	0.6095	0.8765	0.7325	0.6124	0.8855	0.7388	0.6212	0.9111	0.7588	0.6382
5%	0.8748	0.7314	0.6115	0.8845	0.7425	0.6201	0.8947	0.7511	0.6344	0.9188	0.7614	0.6464
10%	0.8785	0.7321	0.6254	0.8975	0.7514	0.6314	0.9074	0.7617	0.6412	0.9221	0.7812	0.6512

Table 4 Comparison of Recommendation Precision (*MAE*) for Hybrid Attack when Using Epinions Dataset

表 4 混合攻击下基于 Epinions 数据集的推荐精度(*MAE*)对比

Filling Size	Attack Size											
	1%			2%			3%			5%		
	CF	CF-DNC	CF-CRIS	CF	CF-DNC	CF-CRIS	CF	CF-DNC	CF-CRIS	CF	CF-DNC	CF-CRIS
1%	0.8817	0.7554	0.7065	0.8945	0.7625	0.7154	0.9054	0.7784	0.7254	0.9155	0.7888	0.7386
3%	0.8898	0.7581	0.7098	0.8998	0.7681	0.7198	0.9045	0.7754	0.7245	0.9176	0.7857	0.7388
5%	0.8912	0.7621	0.7154	0.9012	0.7721	0.7221	0.9144	0.7877	0.7358	0.9277	0.7978	0.7487
10%	0.8988	0.7699	0.7189	0.9110	0.7811	0.7278	0.9223	0.7954	0.7377	0.9387	0.8057	0.7498

从表 3 和表 4 可以看出,在同一填充规模下,随着攻击规模的不断增大, 3 种推荐算法的 *MAE* 都有上升趋势,由此可见,随着攻击用户的增多,系统的推荐精度逐渐下降. 此外,无论在何种攻击规模和填充规模下, CF-CRIS 的推荐 *MAE* 值都明显小于 CF 推荐算法和 CF-DNC 算法,而且 CF-CRIS 因攻击产生的预测偏差也比 CF 算法和 CF-DNC 算法要小. 以采用 FilmTrust 数据集的实验结果为例,在受到平均攻击的情况下, CF-CRIS 算法的推荐精度比 CF 算法和 CF-DNC 算法分别提高了大约 30%和 16%,由此可见,本文算法具有良好的抗攻击能力.

在混合攻击方式下,分别采用 FilmTrust 和 Epinions 数据集,当用户概貌信息的填充规模为 3%, 5%, 10%时,采用 3 种推荐算法的平均预测偏差

(*APS*)对比结果如图 5～7 所示. 从图 5～7 可以看出,在同一填充规模下,无论采用哪个数据集,3 种推荐算法的 *APS* 值都随攻击规模的增大而增大,由此可见,攻击用户数越多推荐质量越差. 在同样的填充规模和攻击规模下,CF-CRIS 算法比 CF 算法预测偏差要小很多,比 CF-DNC 算法的预测偏差也要略小一些,由此可见,CF-CRIS 算法对于用户概貌攻击具有较强的抵抗能力.

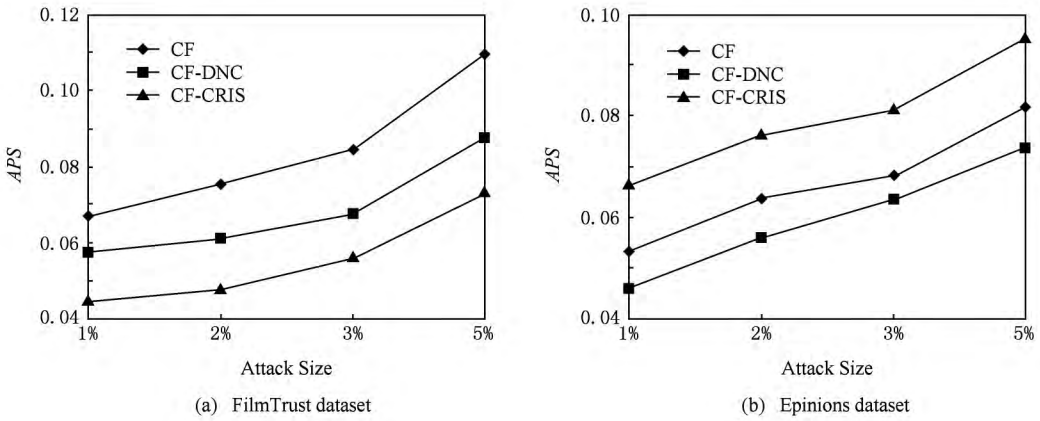

Fig. 5 Comparison of *APS* with 3% filling size.

图 5 3%填充规模时平均预测偏差(*APS*)对比

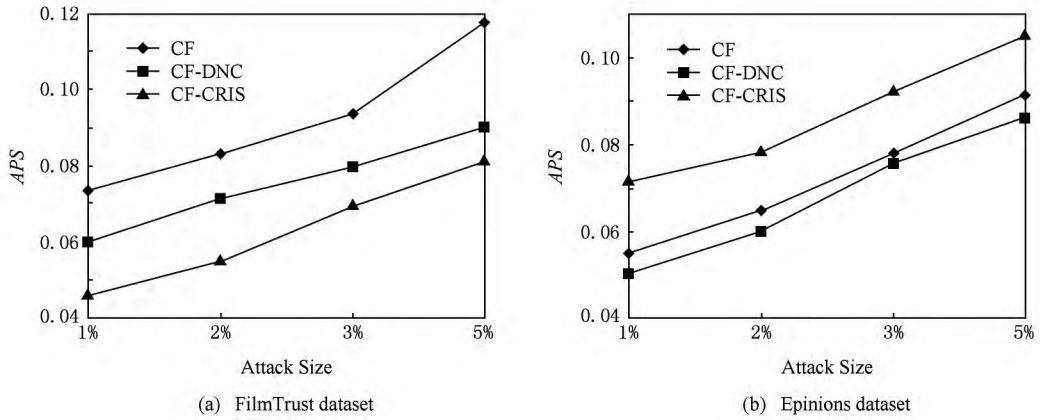

Fig. 6 Comparison of *APS* with 5% filling size.

图 6 5%填充规模时平均预测偏差(*APS*)对比

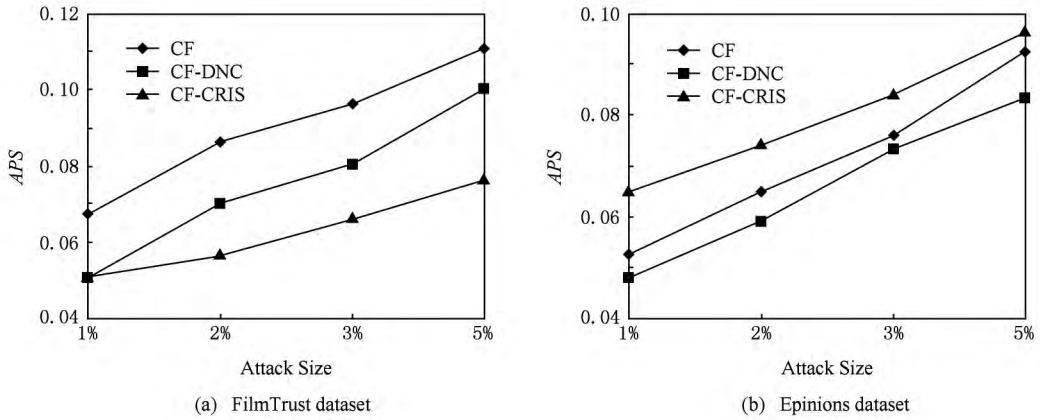

Fig. 7 Comparison of *APS* with 10% filling size.

图 7 10%填充规模时平均预测偏差(*APS*)对比

5 结束语

随着个性化推荐技术在电子商务系统中的广泛应用,关于推荐系统的推荐精度、召回率及抗攻击能力方面的研究越来越引起人们的关注.本文借鉴社会心理学中的信任产生原理,综合考虑多种信任要素在社交信任度量中的作用,提出一种基于多元社交信任的协同过滤推荐算法 CF-CRIS.

CF-CRIS 算法利用用户-项目评分数据度量用户的可信度和可靠性,基于用户显式声明的信任关系推理用户间的隐式信任和用户的信誉度,综合以上信任要素进行协同推荐.该算法的推荐精度和召回率都较传统方法和现有方法具有大幅度提高,并表现出良好的抗攻击能力.

CF-CRIS 算法综合利用用户评分相似度和用户信任关系选取推荐邻居,可以避免传统推荐中的数据稀疏性问题,而且信任推理和全局信任度的引入可以有效解决协同推荐中的冷启动问题.进一步在实际应用环境中检测本文方法的性能是下一步的研究工作.

参 考 文 献

[1] Jones K, Leonard L. Trust in consumer-to-consumer electronic commerce [J]. Information Management, 2008, 45(2): 88-95

[2] Gambetta D. Trust [M]. Oxford, UK: Oxford University Press, 1990: 213-237

[3] Wanita S, Suryan N, Cecile P. A survey of trust in social networks [J]. ACM Computing Surveys, 2013, 45(4): 1-33

[4] Paolo M, Paolo A. Trust-aware collaborative filtering for recommender systems [G] //LNCS 3290: Proc of the Int Conf on CoopIS, DOA, and ODBASE. Berlin: Springer, 2004: 492-508

[5] Golbeck J A. Computing and applying trust in Web-based social networks [D]. College Park, Maryland: University of Maryland, 2005

[6] Zhang Y, Chen H, Andwu Z. A social network-based trust model for the semantic Web [G] //LNCS 4158: Proc of the Int Conf on Autonomic and Trusted Computing. Berlin: Springer, 2006: 183-192

[7] Kuter U, Golbeck J. SUNNY: A new algorithm for trust inference in social networks, using probabilistic confidence models [C] //Proc of the 22nd Int Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2007: 1377-1382

[8] Kim Y A. Building a Web of trust without explicit trust ratings [C] //Proc of IEEE ICDE'08. Piscataway, NJ: IEEE, 2008: 531-536

[9] Caverlee J, Liu L, Webb S. Towards robust trust establishment in Web-based social networks with SocialTrust [C] //Proc of the Int Conf on World Wide Web. New York: ACM, 2008: 1163-1164

[10] Zuo Y, Hu W C, O'Keefe T. Trust computing for social networking [C] //Proc of IEEE ICITNG'09. Piscataway, NJ: IEEE, 2009: 1534-1539

[11] Hang C W, Munindar P S. Trust based recommendation based on graph similarities [C] //Proc of the IEEE AAMAS'10. Piscataway, NJ: IEEE, 2010: 1-11

[12] Seth A, Myers, Zhu C G, et al. Information diffusion and external influence in networks [C] //Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2012: 33-41

[13] Xing Xing. Research on recommendation methods in social networks [D]. Dalian: Dalian Maritime University, 2013 (in Chinese)
(邢星. 社交网络个性化推荐方法研究[D]. 大连:大连海事大学, 2013)

[14] Jiang W J, Wang G J, Wu J. SWTrust: Generating trusted graphs for trust evaluation in online social networks [C] // Proc of the IEEE TrustCom'10. Piscataway, NJ: IEEE, 2011: 320-327

[15] Liu H, Lim E P, Lauw H W, et al. Predicting trusts among users of online communities: An Epinions case study [C] // Proc of the 9th ACM Conf on Electronic Commerce. New York: ACM, 2008: 310-319

[16] Tian Liqin, Lin Chuang. A kind of Game-Theoretic control mechanism of user behavior trust based on prediction in trustworthy network [J]. Chinese Journal of Computers, 2007, 30(11): 1930-1939 (in Chinese)
(田立勤, 林闯. 可信网络中一种基于行为信任预测的博弈控制机制[J]. 计算机学报, 2007, 30(11): 1930-1939)

[17] Chen Jing, Du Ruiying, Wang Lina, et al. A trust game method basing on probability model in networks [J]. Acta Electronic Sinica, 2010, 38(2): 427-433 (in Chinese)
(陈晶, 杜瑞颖, 王丽娜, 等. 网络环境下一种基于概率密度的信任博弈模型[J]. 电子学报, 2010, 38(2): 427-433)

[18] Adali S, Escriva R, Goldberg M K, et al. Measuring behavioral trust in social networks [C] //Proc of IEEE ISI'10. Piscataway, NJ: IEEE, 2010: 150-152

[19] Qiao Xiuquan, Yang Chun, Li Xiaofeng, et al. A trust calculation algorithm based on social networking service users' context [J]. Chinese Journal of Computers, 2011, 34 (12): 2043-2052 (in Chinese)
(乔秀全, 杨春, 李晓峰, 等. 社交网络服务中一种基于用户上下文的信任度计算方法[J]. 计算机学报, 2011, 34(12): 2043-2052)

[20] Nepal S, Sherchan W, Patis C. STrust: A trust model for social networks [C] //Proc of the IEEE TrustCom'11. Piscataway, NJ: IEEE, 2011: 841-846

[21] Kim Y A, Phalak R. A trust prediction framework in rating-based experience sharing social networks without a Web of trust [J]. Information Sciences, 2012, 191: 128-145

[22] Li Y M, Shiu Y L. A diffusion mechanism for social advertising over microblogs [J]. Decision Support Systems, 2012, 54(1): 9-22

[23] Nikolay K, Alex T. Trust prediction from user-item ratings [J]. Social Network Analysis and Mining, 2013, 3(3): 749-759

[24] Emanuel L, Dominik K, Lukas E. Utilizing online social network and location-based data to recommend products and categories in online marketplaces [G] //LNCS 8940: Proc of the 4th Int Workshops on MUSE, Berlin: Springer, 2013: 96-115

[25] David H M, Charles H G, Rebert M G. The Trusted Advisor [M]. New York: Free Press, 2000

[26] Jia Dongyan, Zhang Fuzhi. A collaborative filtering recommendation algorithm based on double neighbor choosing strategy [J]. Journal of Computer Research and Development, 2013, 50(5): 1076-1084 (in Chinese)
(贾冬艳, 张付志. 基于双重邻居选择策略的协同过滤推荐算法[J]. 计算机研究与发展, 2013, 50(5): 1076-1084)

Wang Ruiqin, born in 1979. PhD, lecturer. Member of China Computer Federation. Her main research interests include data mining, knowledge services and social recommendation.

Jiang Yunliang, born in 1967. PhD, professor. His main research interests include artificial intelligence and data integration (jylsy@hutc.zj.cn).

Li Yixiao, born in 1982. PhD, associate professor. His main research interests include complex system and complex network (yixiao_li@126.com).

Lou Jungang, born in 1981. PhD, associate professor. His main research interests include dependable computing and software reliability evaluation (ljg@hutc.zj.cn).