基于多重信任的协同过滤推荐算法

于 阳 于洪涛 黄瑞阳
(国家数字交换系统工程技术研究中心 郑州 450002)

摘 要 针对评分数据稀疏性和用户冷启动所导致的协同过滤推荐系统的准确度与覆盖率较低的问题,文中融合显性信任和隐性信任因素,提出了一种基于多重信任的协同过滤推荐算法。首先,依据用户间推荐评分的准确性与可依赖度因子,提出一种改进的均方差(Mean Squared Difference,MSD)信任度量方法,并在此基础上提出基于隐性信任信息的评分模型;其次,以最大信任传播距离为约束,提出一种显性信任信息的关系模型;最后,依据评分相似性与显性信任关系,利用0-1背包组合优化策略选择出目标用户的最优近邻集合,从而进行评分预测。在Epinions数据集上与多种主流算法的对比仿真实验结果表明,该算法通过引入有效评分和显性信任关系,极大地缓解了数据稀疏性和冷启动问题,并且在不牺牲覆盖率的条件下显著提升了推荐准确度。

关键词 协同过滤,稀疏性,冷启动,显性信任,隐性信任,0-1背包问题
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.05.019

Collaborative Filtering Recommendation Algorithm Based on Multiple Trust

YU Yang YU Hong-tao HUANG Rui-yang
(China National Digital Switching System Engineering & Technological R&D Center,Zhengzhou 450002,China)

Abstract Aiming at the reduction of accuracy and coverage for collaborative filtering recommendation system caused by sparseness of scoring data and cold start of users,this paper integratesd the explicit trust and implicit trust factors,and proposed a collaborative filtering recommendation algorithm based on multiple trust. Firstly,an improved Mean Squared Difference(MSD) trust metric method was proposed based on the accuracy and dependability factor of the recommended scores among users. Based on this,a scoring model based on implicit trust information was proposed. Secondly,regarding the maximum trust propagation distance as the constraint,a relational model of explicit trust information was proposed. Finally,based on the similarity between the score and the explicit trust,the optimal neighbor set of the target user was selected by the 0-1backpack combination optimization strategy,and the scoring prediction was carried out. Comparisons of the simulation results with a variety of state-of-the-art algorithms on Epinions dataset demonstrate that the proposed algorithm can greatly alleviate the data sparsity and cold start problems by introducing effective score and explicit trust relationship,and significantly improve the recommendation accuracy while preserving good coverage.

Keywords Collaborative filtering,Sparsity,Cold start,Explicit trust,Iimplicit trust,0-1 knapsack problem

1 引言

在大数据时代下,一股股数据浪潮将人们层层包裹在信息的海洋之中,使之迷失方向,信息过载(Information Overload)问题由此产生。传统的基于分类目录和关键词搜索等普适化信息的过滤技术已经无法满足人们的需求,用户想要在海量数据中获取所需的信息变得尤为困难。与之不同,作为解决信息过载问题而发展起来的产物,推荐系统能够利用用户的兴趣、行为、情景等信息,主动地为用户推荐其可能感兴趣的内容。追溯历史,20世纪90年代初,一句微弱的歌声:“利用网络上数百万人的意见帮助用户发现其可能感兴趣的内容”[1],今已被证实为天籁之音。经过20余年的蜕变,推荐系统已从最初的普适化推荐发展成为如今的个性化推荐,涉及机器学习、数据挖掘、神经网络等诸多研究领域,涵盖电子商务、社交网络、在线内容服务等多种应用场景[2]。

网络中用户对信息的浏览、选择、评价等行为正是用户个体对信息过滤的体现,而协同过滤(Collaborative Filtering)就是将用户个体的行为视为用户群体间的分布式协同工作,利

到稿日期:2017-03-14 返修日期:2017-06-04 本文受国家自然科学基金创新群体项目(61521003),国家自然科学基金资助项目(61171108),国家科技支撑计划(2014BAH30B01)资助。

于 阳(1991-),男,硕士生,主要研究方向为推荐系统、网络大数据分析,E-mail:yang_y9802@163.com;于洪涛(1970-),男,研究员,主要研究方向为网络空间安全、网络大数据分析,E-mail:yht_ndsc@139.com(通信作者);黄瑞阳(1986-),男,副研究员,主要研究方向为数据挖掘、复杂网络。

用群体智慧进行信息过滤。文献[3]阐述了协同过滤算法的基本思想、分类以及应用场景。然而,协同过滤推荐系统虽然已经在某些领域得到广泛应用,但仍存在两个亟待解决的问题:评分数据稀疏性和用户冷启动[4-5]。其中,评分数据稀疏性是指推荐系统中大多数用户往往仅对少许项目进行打分,导致可用的评分信息极其有限;冷启动用户是指系统中存在的仅具有少量评分信息甚至不具有评分信息的用户。文献[6]表明,协同过滤推荐系统通常很难为冷启动用户提供满意的推荐,同时在评分数据较稀疏的情况下也无法表现出良好的性能。

为了缓解协同过滤推荐系统面临的窘境,研究者提出了大量的解决方法,本文主要将其归结为两条解决思路。第一条:综合考虑用户群体的评分信息,旨在全面挖掘用户之间的相关性。按此研究思路,一些学者提出了矩阵填充模型[7]、相似性计算模型[8]等方法。该类模型虽然能够在一定程度上弥补上述缺陷,但是在面对评分数据极剧稀疏的情况时仍无济于事。另一些学者提出了矩阵降维模型[9]、聚类模型[10]等方法,前者的模型训练过程大多在离线环境下进行,导致实时信息无法及时参与训练;后者受以下因素影响:1)邻居用户仅从固定的簇成员中选取;2)推荐性能易受簇质量的影响,往往表现出相对较低的准确度和覆盖率。第二条:引入额外信息源,旨在广泛融合用户之间的相关性。沿此思路,相关研究者[11-12]提出融合网络中用户间的好友关系、成员关系、信任关系等方法。Singla 等[13]提出,相比于其他因素,信任因素不仅具有更高的可靠性,而且与相似性极为相关。同时,在社会化推荐中,信任关系备受关注。一方面,在信任匮乏的时代下,人们更加珍惜彼此之间的信任;另一方面,在现实世界中,人们的购买行为易受信任朋友的影响。鉴于此,信任感知推荐系统[14-15](Trust-aware Recommender Systems,TARS)能够利用社会化网络中的信任关系挖掘出用户间的相关性,为缓解评分数据稀疏性和用户冷启动问题带来了新的契机。

信任感知推荐系统的基本思想是:用户的信任邻居与其具有相似的兴趣偏爱。该系统的原型如图 1 所示。在推荐过程中,用户间的信任关系通常扮演着两个重要角色:1)作为近邻选择的重要准则;2)用于未知项目的评分预测。

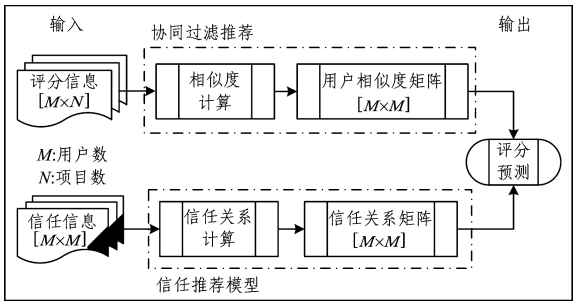


图 1　信任感知推荐系统原型

Fig. 1　Trust-aware recommendation system prototype

Guo[16]将推荐系统中的信任定义为:“信任衡量的是用户之间能够提供有价值评分信息的能力”。随后,Guo[17]全面阐述了信任的特性:非对称性、传递性、动态性和上下文依赖性。依据信任的不同来源,可将其分为显性信任(Explicit Trust)和隐性信任(Implicit Trust)。显性信任是指网络中用户之间主动表达的信任关系,隐性信任通常是指基于评分信息挖掘出的信任关系。在推荐系统中,用显性信任来表征信任关系时具有高度的可靠性与准确性;而隐性信任能够更好地区分信任度,且信任信息易获取。

在基于隐性信任的研究中,相关学者提出了许多有关信任的度量方法。最初,Papagelis 等[18]利用 Pearson 相关性衡量了信任相关性。O'Donovan 等[19]详细阐述了两种信任度量模式,即用户级信任(Profile-levelTrust)和项目级信任(Item-level Trust),并分别描述了特定用户的全局信任和用户对特定项目的全局信任。随后,Hwang 等[20]提出首先仅依据单一邻居的评分信息为目标用户进行评分预测,然后通过用户间共同评分项目的平均预测偏差来表征信任度。依据相同的策略,Shambour 等[21]提出利用用户间的预测评分与真实评分的均方差来表征信任度。然而,上述信任度量方法通常都不具备非对称性,Guo[17]全面剖析了以上信任度量方法的性能。为了设计更加合理的信任计算方法,Bedi 等[22]提出基于蚁群优化的算法来动态更新用户间的信任度。Tang 等[23]全面衡量了全局信任度和局部信任度,从而构建了用户信任模型。尽管上述文献从不同的角度衡量了用户间的信任度并取得了一定成效,但是大多数算法都仅简单地利用信任关系对未知项目进行评分预测,均未衡量预测评分的可靠性。

在基于显性信任的相关研究中,Golbeck[24]最早打开信任研究的大门,将社会化网络中的信任关系引入到协同过滤推荐系统中,提出了基于广度优先搜索策略的 TidalTrust 模型。在此基础上,Massa 等[25]讨论了传统协同过滤推荐系统存在的缺陷,详细阐述了融合信任因素的合理性,并基于深度优先搜索策略设计了与 TidalTrust 类似的 MoleTrust 模型。两种模型都仅使用信任关系来充当相似性关系的角色,实验证明其虽然能够在一定程度上提高覆盖率,但是推荐准确度的提升空间却相对有限。Chowdhury 等[26]针对目标用户的相邻用户中未对目标项目进行评分的用户,用基于信任的推荐模型为其产生预测评分,进而融合更多有效的近邻用户。该算法虽然提高了推荐质量,但是在冷启动用户上的表现却不令人满意。随后,Ray 等[27]设定了相似性阈值,提出若用户间的相似性低于阈值,则在信任网络中移除相应的信任关系,进而重构信任网络。该算法不仅是以牺牲覆盖率为代价来提升推荐准确度,而且无法处理冷启动用户。最近,Guo 等[28]提出了 Merge 模型,该模型首先基于信任邻居对未评分项目进行评分预测;接着考虑信任邻居对目标项目的评分数和评分一致性因素,提出了可靠性量化方法来度量预测评分的可靠性,进而通过融合评分全面表征用户的兴趣偏爱。Moradi 等[29]给出了 RTCF 模型,该模型首先基于显性信任机制为目标项目进行打分,并衡量评分的可靠性;随后设定可靠性阈值,针对评分可靠性低于阈值的用户,通过综合考虑积极因子、消极因子来重构其信任网络。上述文献简述了基于显性信任推荐系统的研究历程,虽然从不同角度缓解了协同过滤推荐系统存在的问题,提升了推荐质量,但是仍存在一个基本问题:大多数算法都仅简单地依据“信任度越大的相邻用

户与其偏好相似性越高”这一假设来选取近邻用户，忽略了目标用户信任但与其偏好相似性较低的用户对推荐系统产生的负面影响。

针对上述问题，本文在已有研究的基础上聚焦于用户间的显性信任关系和隐性信任关系，提出一种基于多重信任的协同过滤推荐算法，其模型如图 2 所示。该模型主要分为 3 个模块：隐性信任评分模块、显性信任关系模块和近邻集合选取模块。首先，依据用户间推荐评分的准确性与可依赖度因子提出一种改进的均方差信任度量方法，并在此基础上提出基于隐性信任信息的评分模型；接着，给出一种置信度评价方法来评测预测评分的可靠性，从而准确地衡量用户间评分的相似性；其次，以最大信任传播距离为约束，提出一种基于显性信任信息的关系模型；最后，依据评分相似性和显性信任关系，利用 0-1 背包组合优化策略选择出目标用户的最优近邻集合，从而进行预测推荐。在 Epinions 数据集上与多种主流算法的对比仿真实验结果表明，该算法通过引入有效评分和显性信任关系，极大地缓解了数据稀疏性和冷启动问题，并且在不牺牲覆盖率的条件下显著提升了推荐准确度。

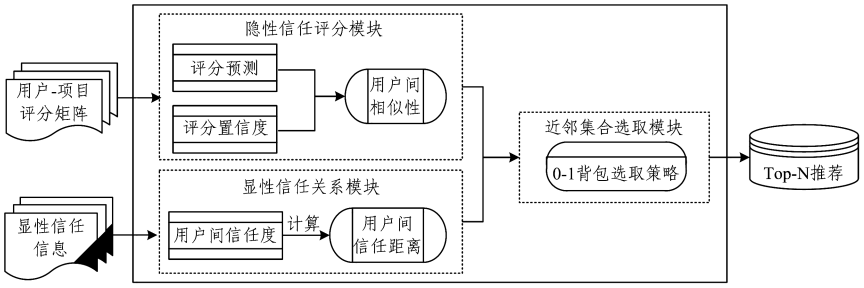


图 2 基于多重信任的推荐模型

Fig. 2 Recommendation model based on multiple trust

本文第 2 节详细阐述所提算法各部分的实现流程；第 3 节介绍实验设计并分析实验结果；最后总结全文。

2 基于多重信任的协同过滤推荐算法

为了更好地处理评分数据稀疏性和用户冷启动问题，本文给出一种基于多重信任的推荐算法。算法分为 4 个部分：1)隐性信任评分模型；2)显性信任关系模型；3)近邻集合选取策略；4)预测推荐。本节将描述各部分的相关计算，并给出算法的整体实现流程。

在详细阐述算法之前，为便于读者的阅读和后文描述的简洁，首先简要概述本文所涉及的基本概念和相关符号。

定义 1(用户-项目评分矩阵 R) 依据评分信息，由 m 个用户组成的集合 $U=\{u_1,u_2,\cdots,u_m\}$ 和 n 个项目组成的集合 $I=\{i_1,i_2,\cdots,i_n\}$ 构成 $m\times n$ 维用户-项目评分矩阵 R。同时，$r_{u,i}(\{(u,i)\mid 1\leqslant u\leqslant m,1\leqslant i\leqslant n\})$ 表示用户 u 对项目 i 的评分，$I_u=\{i\mid r_{u,i}\in R,i\in I\}$ 表示用户 u 已评分的项目集合，$U_i=\{u\mid r_{u,i}\in R,u\in U\}$ 表示评价过项目 i 的用户集合。

表 1 详细列出了本文所用到的主要符号。

表 1 符号的定义及含义

Table 1 Definition of symbols and corresponding meanings

符号	含义
U,I,R	用户集合、项目集合、用户-项目评分矩阵
I_u,I_v	用户 u 和用户 v 已评分的项目集合
$r_{u,i},r_{v,i}$	用户 u 和用户 v 对项目 i 的评分
$\bar{r}_u,\bar{r}_v$	用户 u 和用户 v 的评分均值
$r_{\max},r_{\min}$	评分范围(通常为[1,5])的最大值、最小值
$tr^{im}(u,v)$	用户 u 对用户 v 的隐性信任度
$C_{u,i},C_{v,i}$	用户 u 和用户 v 对项目 i 的评分置信度
$sim(u,v)$	用户 u 和用户 v 之间的评分相似性
$tr^{ex}(u,v)$	用户 u 对用户 v 的显性信任度
$d_t(u,v)$	用户 u 对用户 v 的信任距离
$w(u,v)$	用户 v 对用户 u 的推荐权重
$P_{u,i}$	用户 u 对项目 i 的预测评分

2.1 隐性信任评分模型

2.1.1 基于信任的评分预测

用户群体的评分信息中隐藏着用户间的多种关系，而信任作为其中最重要的一种关系，能够直接影响用户未来的行为决策。因此，从评分数据中有效地挖掘出用户间的隐性信任关系是提高推荐质量的关键。用户间的信任关系受多方面因素的影响，但其本质是：信任主体(Trustor)依据与信任客体(Trustee)的历史交互信息，对其未来行为决策的主观期望。

于是，本节首先基于用户间推荐评分的准确性与可依赖度因子提出了一种隐性信任度量方法，随后依据信任邻居的评分信息对目标用户进行评分填充，旨在全面表征用户的兴趣偏爱。用户之间的隐性信任网络如图 3 所示。

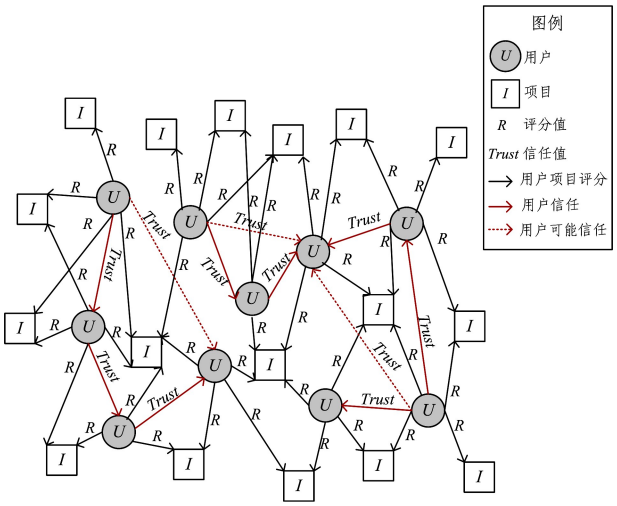


图 3 隐性信任网络

Fig. 3 Implicit trust network

MSD 是一种衡量预测评分与真实评分之间的贴近程度的有效方法。最近，文献[21]提出了一种基于 MSD 的隐性

信任度量方法,但此计算方法通常不具备信任的非对称性。于是,本节在 Shambour[21] 的基础上提出一种改进的 MSD 隐性信任度量方法。首先,通过 MSD 衡量基于用户 v 对用户 u 进行预测评分时的准确性(Accuracy),计算公式如式(1)所示:

$$Accuracy(u,v)=1-\frac{1}{|I_{uv}|}\sum_{i\in I_{uv}}\left(\frac{p_{u,i}-r_{u,i}}{r_{\max}}\right)^2 \tag{1}$$

其中,$I_{uv}=I_u\cap I_v$ 表示用户 u 与用户 v 共同评分的项目集合,$p_{u,i}$ 表示基于单一用户的评分信息计算出的用户 u 对项目 i 的预测评分,计算公式如式(2)所示:

$$p_{u,i}=\bar{r}_u+(r_{v,i}-\bar{r}_v) \tag{2}$$

其次,受评分数据稀疏性的影响,当用户间的共同评分项目数较少时,准确性下降。同时,又如第 2 节所述,信任具有非对称性。因此,本节给出一种依赖性(Dependability)因子来衡量基于用户 v 对用户 u 进行预测评分时的可依赖度,计算公式如式(3)所示:

$$Dependability(u,v)=\frac{|I_u\cap I_v|}{|I_v|} \tag{3}$$

最后,综合考虑用户间推荐评分的准确性与可依赖度因素,给出隐性信任的计算公式:

$$\begin{aligned}tr^{im}(u,v)&=Dependability\times Accuracy\\&=\frac{|I_u\cap I_v|}{|I_v|}\cdot\left[1-\frac{1}{|I_{uv}|}\sum_{i\in I_{uv}}\left(\frac{p_{u,i}-r_{u,i}}{r_{\max}}\right)^2\right]\\&=\frac{|I_{uv}|-\sum_{i\in I_{uv}}\left(\frac{p_{u,i}-r_{u,i}}{r_{\max}}\right)^2}{|I_v|}\end{aligned} \tag{4}$$

考虑到用户总是信任自己,设定 $tr^{im}(u,u)=1$。

用户的评分信息反映了用户对项目的兴趣偏爱,但受评分数据稀疏性的影响,可用的信息偏少。然而,通过上述讨论可知,用户未来的行为决策易受信任朋友的影响。于是,本节利用信任邻居的评分信息来对目标用户进行评分预测。

定义 2(隐性信任邻居集合 TN^{im})　给定目标用户 $u(u\in U)$,若 $\exists v\in U$,使得 $tr^{im}(u,v)>0$ 成立,则称用户 v 为用户 u 的隐性信任邻居。因此,目标用户 u 的隐性信任邻居集合表述为:$TN_u^{im}=\{v|tr^{im}(u,v)>0,v\in U\}$。

定义 3(候选评分项目集合 $\bar{I}$)　给定目标用户 $u(u\in U)$,若 $\exists v\in TN_u^{im}$ 且 $\exists i\in(I-I_u)$,使得 $r_{v,i}\in R$ 成立,则可以称项目 i 为用户 u 的候选评分项目。因此,目标用户 u 的候选评分项目集合表述为:$\bar{I}_u=\{i|r_{v,i}\in R,v\in TN_u^{im}\text{ 且 }i\in(I-I_u)\}$。

于是,目标用户 u 对其候选评分项目 $i(i\in\bar{I}_u)$ 的预测评分 $\tilde{r}_{u,i}$ 可通过式(5)计算得出:

$$\tilde{r}_{u,i}=\frac{\sum_{v\in TN_u^{im}}tr^{im}(u,v)r_{v,i}}{\sum_{v\in TN_u^{im}}tr^{im}(u,v)} \tag{5}$$

基于原始评分信息能够准确体现用户对项目的偏爱程度的事实,经评分填充后,目标用户 u 的评分向量表述为:$I_u'=\{r'_{u,1},\cdots,r'_{u,i},\cdots,r'_{u,n}\}$,其中 $r'_{u,i}=\begin{cases}r_{u,i}, & \text{if } i\in I_u\\ \tilde{r}_{u,i}, & \text{else if } i\in\bar{I}_u\\ 0, & \text{else}\end{cases}$。

2.1.2　评分置信度

尽管用户对其候选评分项目的预测评分可以通过式(5)获得,但预测评分的置信度(Confidence)是未知的。如果忽略评分置信度而对其进行盲目使用,那么往往会适得其反。针对原始评分数据较多的活跃用户来说,此做法或许会带来更大的负作用。因此,有效地衡量预测评分的可靠性变得至关重要。最近,Hernando 等[30] 基于相似性关系提出了一种评分可靠性评价方法,该方法显著提高了推荐质量。本节在 Hernando 的基础上,结合信任关系,给出一种合理的计算评分置信度的方法。

设目标用户 u 对其候选评分项目 i 的预测评分为 $\tilde{r}_{u,i}$,则用户 u 对项目 i 的评分置信度 $C_{u,i}$ 被定义为:

$$C_{u,i}=\left[f_s(S_{u,i})\cdot f_v(V_{u,i})^{f_s(S_{u,i})}\right]^{\frac{1}{1+f_s(S_{u,i})}} \tag{6}$$

其中,$f_s(S_{u,i})$ 和 $f_v(V_{u,i})$ 分别表示积极算子函数和消极算子函数。评分置信度 $C_{u,i}$ 的取值范围为(0,1),$C_{u,i}$ 值越大表示预测评分越可靠。同时,鉴于 2.1.1 节提及的"事实",设定 $C_{u,j}=1(j\in I_u)$。因此,目标用户 u 对全体项目 i 的评分置信度被描述为:$C_{u,i}=\begin{cases}1, & \text{if } i\in I_u\\ C_{u,i}, & \text{else if } i\in\bar{I}_u\\ 0, & \text{else}\end{cases}$。

1)积极算子函数

$$f_s(S_{u,i})=1-\frac{\bar{S}}{\bar{S}+S_{u,i}} \tag{7}$$

$$S_{u,i}=\sum_{v\in TN_{u,i}^{im}}tr^{im}(u,v) \tag{8}$$

$$TN_{u,i}^{im}=\{v|r_{v,i}\in R,v\in TN_u^{im}\text{ 且 }i\in\bar{I}_u\} \tag{9}$$

$$\bar{S}=\sum_{i\in\bar{I}_u}S_{u,i} \tag{10}$$

其中,积极算子 $S_{u,i}$ 的定义为式(8);积极算子的平均值 $\bar{S}$ 的定义为式(10);$TN_{u,i}^{im}$ 表示 TN_u^{im} 中评价过项目 $i(i\in\bar{I}_u)$ 的用户集合,计算公式为式(9)。

2)消极算子函数

$$f_v(V_{u,i})=\left(\frac{r_{\max}-r_{\min}-V_{u,i}}{r_{\max}-r_{\min}}\right)^{\gamma} \tag{11}$$

$$\gamma=\frac{\ln 0.5}{\ln\frac{r_{\max}-r_{\min}-\bar{V}}{r_{\max}-r_{\min}}} \tag{12}$$

$$V_{u,i}=\frac{\sum_{v\in TN_{u,i}^{im}}tr^{im}(u,v)\cdot(r_{v,i}-\bar{r}_v-\tilde{r}_{u,i}+\bar{r}_u)}{\sum_{v\in TN_{u,i}^{im}}tr^{im}(u,v)} \tag{13}$$

$$\bar{V}=\sum_{i\in\bar{I}_u}V_{u,i} \tag{14}$$

其中,消极算子 $V_{u,i}$ 的定义为式(13);消极算子的平均值 $\bar{V}$ 的定义为式(14)。

2.1.3　用户评分相似度

协同过滤推荐系统中,准确地衡量用户间的评分相似性是保障推荐质量的关键。于是,基于上述讨论,考虑到信任邻居对目标用户的融合评分和评分置信度因素,本节基于皮尔森相关系数(Pearson Correlation Coefficient,PCC)的原型[1]

给出相似性计算方法。

设目标用户 u 与用户 v 的评分集合分别为 I'_u($I'_u=\{i|r'_{u,i}\in R,i\in I\}$)和 I'_v($I'_v=\{j|r'_{v,j}\in R,j\in I\}$),则用户 u 和用户 v 之间的相似性定义为:

$$sim(u,v)=\frac{\sum_{i\in I'_{uv}}C_{u,i}(r'_{u,i}-\bar{r}'_{u,i})C_{v,i}(r'_{v,i}-\bar{r}'_{v,i})}{\sqrt{\sum_{i\in I'_{uv}}C^2_{u,i}(r'_{u,i}-\bar{r}'_{u,i})^2}\sqrt{\sum_{i\in I'_{uv}}C^2_{v,i}(r'_{v,i}-\bar{r}'_{v,i})^2}} \tag{15}$$

其中,$I'_{uv}=I_u'\cap I_v'$ 表示经评分填充处理后用户 u 和用户 v 共同评分的项目集合;其余符号表示的意义与上文相似,在此不再赘述。

2.2 显性信任关系模型

社会化网络中,用户之间不仅存在隐性信任关系,而且可以通过主动地表达信任意愿来产生显性信任关系。尽管依据这种显性信任可以准确地构建信任网络,但由于隐私保护问题,真实数据集中包含的信任信息偏少,造成推荐过程中所含的相邻信任邻居较少。鉴于信任具有传递性(如果用户 A 信任用户 B 且用户 B 信任用户 C,那么用户 A 也可能信任用户 C),基于信任的传播与聚合规则可以有效地度量非相邻用户之间的信任关系。同时,考虑到在真实数据集中通常使用二分值来表征用户之间的信任度(即:1 表示信任,0 表示不信任),经典的信任传播算法 TidalTrust[24] 和 MoleTrust[25] 虽然能够度量非相邻用户间的信任度,但是均无法区分信任距离。鉴于此,本节给出式(16)来合理地度量用户间的显性信任关系:

$$tr^{ex}(u,v)=\frac{d_{\max}-d_{u,v}+1}{d_{\max}} \tag{16}$$

其中,$d_{u,v}$ 表示信任网络中用户 u 到用户 v 的最短信任距离,$d_{\max}$ 表示信任网络的最大信任传播距离。

目前,多数相关研究都仅简单地依据六度分离(Six Degree of Separation)理论主观地决定 $d_{\max}$ 的取值。事实上,选择一个合适的 $d_{\max}$ 尤为重要。$d_{\max}$ 值越大,推荐过程中就能包含越多的信任邻居,从而提高推荐覆盖率,但随之也会造成更大的信任噪声与计算复杂度;$d_{\max}$ 值越小,推荐过程中就能越准确地计算用户之间的信任度,进而提高推荐准确度,但随之也会牺牲覆盖率。然而,文献[31]证明了真实的信任网络通常属于小世界网络,且它的小世界拓扑结构不受网络动态性的影响。因此,基于小世界网络特性客观地给出 $d_{\max}$ 的计算公式,如式(17)所示:

$$d_{\max}=L^R=\frac{\ln N}{\ln\langle k\rangle} \tag{17}$$

其中,L^R 表示与信任网络具有相同规模和密度的 ER 随机网络的平均路径长度;$\langle k\rangle$ 表示信任网络的平均度,其计算公式如式(18)所示:

$$\langle k\rangle=\frac{M}{N} \tag{18}$$

其中,M 和 N 分别表示信任网络的边数和节点数。

于是,设目标用户 u 对用户 v 的显性信任度为 $tr^{ex}(u,v)$,则用户 u 对用户 v 的信任距离定义为:

$$d_t(u,v)=1-tr^{ex}(u,v) \tag{19}$$

2.3 近邻集合选取策略

为了去伪存真,减小目标用户信任但与其偏好相似性较低的用户对推荐系统产生的影响,本节综合考虑用户间的相似性和信任距离两种准则,基于 0-1 背包问题(0-1 Knapsack-Problem)提出一种合理的近邻集合选取策略。

0-1 背包问题是一种典型的组合优化问题,并被广泛应用于诸多领域。该问题可以简单地表述为:给定 N 种物品(每种物品有且只有一件)和一个容量为 C 的背包,其中任意物品 i 的质量和价值分别为 w_i 和 v_i,则在总质量不超过背包容量的条件下,如何使得所选物品的总价值达到最大。该问题被形式化描述为:给定 $N>0,C>0$,对于 $\forall i(1\leqslant i\leqslant N)$,$\exists w_i>0,v_i>0$,如何在满足约束的条件下使得目标函数达到最大值。

$$\text{约束条件:}\begin{cases}\sum_{i=1}^{N}w_ix_i\leqslant C\\ x_i\in\{0,1\},1\leqslant i\leqslant N\end{cases}$$

$$\text{目标函数:}\sum_{i=1}^{N}v_ix_i$$

然而,尽管“组合爆炸”,使得背包问题属于 NP 完全问题,但是许多基于动态规划的求解方法已经将其时间复杂度和空间复杂度优化到 O(NC)。

定义 4(候选邻居集合 CNS) 给定目标用户 $u(u\in U)$,若 $\exists v\in U$,使得 $sim(u,v)>0$ 且 $tr^{ex}(u,v)>0$ 同时成立,则称用户 v 为用户 u 的候选邻居用户,因此目标用户 u 的候选邻居集合表述为:$CNS_u=\{v|sim(u,v)>0$ 且 $tr^{ex}(u,v)>0,v\in U\}$。

在 MTCF 算法中,将 CNS_u 中的用户数定义为物品总数 N,将目标用户 u 与任意用户 v($v\in CNS_u$)的 $sim(u,v)$ 和 $d_t(u,v)$ 分别定义为物品的价值和质量,同时,依据式(20)定义背包容量 C,则目标用户 u 的最近邻居集合(Nearest-neighbors Set,NNS)被描述为式(21):

$$C=mean(d_t(u))\times K \tag{20}$$

$$NNS_u=\{v\mid \text{Knapsack}(sim(u,v),d_t(u,v),C),v\in CNS_u\} \tag{21}$$

其中,K 表示选取的最近邻居数,$mean(d_t(u))$ 表示用户 u 到其他用户的信任距离的平均值。

2.4 预测推荐

预测推荐过程包含两部分:1)评分预测;2)TOP-N 推荐。在评分预测过程中,主要基于目标用户 u 的近邻用户对目标项目 i 的评分信息,通过用户间的推荐权重(Weight)加权计算得出 u 对 i 的预测评分 $P_{u,i}$,计算公式如下:

$$P_{u,i}=\bar{r}_u+\frac{\sum_{v\in NNS_u}[w(u,v)(r_{v,i}-\bar{r}_v)]}{\sum_{v\in NNS_u}w(u,v)} \tag{22}$$

其中,$w(u,v)$ 表示近邻用户 v 对目标用户 u 的推荐权重。考虑到推荐权重受用户间相似性和信任关系两方面因素的影响,本节基于调和平均函数给出 $w(u,v)$ 的计算公式,如式(23)所示:

$$w(u,v)=\frac{2\times sim(u,v)\times tr^{ex}(u,v)}{sim(u,v)+tr^{ex}(u,v)} \tag{23}$$

在此基础上,TOP-N 推荐主要依据预测评分$\{P_{u,1}, P_{u,2}, \cdots, P_{u,j}\}$,选取评分值最高的前 N 个项目来组成目标用户的推荐列表。

2.5 算法流程

综上所述,本节直接给出算法的实现流程,如算法 1 所列。

算法 1　基于多重信任的协同过滤推荐算法

Input:用户-项目评分矩阵 $R_{m\times n}$,显性信任矩阵 $T_{m\times m}$,最近邻居数 K

Output:目标用户 a 对目标项目 q 的预测评分 $P_{a,q}$

1. $TN_u^{im}, CNS_u, NNS_u \leftarrow \varnothing, d_{max} \leftarrow \ln N/\ln\langle k\rangle, T_{m\times m}$
2. For each user $u\in U$ do
3. For each user $v\in U$ do
4. $Accuracy(u,v) \leftarrow p_{u,i}(i\in I_{uv}), R_{m\times n}$　　▷using Eq. (1)
5. $Dependability(u,v) \leftarrow |I_u \cap I_v|/|I_v|$
6. $tr^{im}(u,v) \leftarrow Dependability(u,v), Accuracy(u,v)$
7. If $tr^{im}(u,v)>0$ then $TN_u^{im} \leftarrow v$
8. End for
9. For each item $i\in I$ do
10. If $i\in \bar{I}_u$ then
11. $\tilde{r}_{u,i} \leftarrow tr^{im}(u,v), R_{m\times n}$　　▷using Eq. (5)
12. $C_{u,i} \leftarrow tr^{im}(u,v), \tilde{r}_{u,i}, R_{m\times n}$　　▷using Eq. (6)
13. Else if $i\in I_u$ then $C_{u,i} \leftarrow 1$
14. Else $C_{u,i} \leftarrow 0$
15. End for
16. End for
17. For each user $p\in U$ do
18. $sim(u,p) \leftarrow PCC, C_u, C_p$　　▷using Eq. (15)
19. $tr^{ex}(u,p) \leftarrow (d_{max}-d_{u,p}+1)/d_{max}, T_{m\times m}$
20. $d_t(u,p) \leftarrow 1-tr^{ex}(u,p)$
21. If $sim(u,p)>0$ & $tr^{ex}(u,p)>0$ then $CNS_u \leftarrow p$
22. End for
23. $C \leftarrow mean(d_t(u))\times K$
24. $NNS_u \leftarrow Knapsack(sim(u,p), d_t(u,p), C), CNS_u$
25. $w(u,p) \leftarrow sim(u,p), tr^{ex}(u,p), NNS_u$　　▷using Eq. (23)
26. $P_{u,q} \leftarrow w(u,p), NNS_u, R_{m\times n}$　　▷using Eq. (22)

该算法的流程主要分为 4 个部分。第一部分为第 1 行,用于初始化变量。第二部分为第 2－16 行,主要计算用户之间的信任关系,利用隐性信任评分模型来填充评分矩阵,并衡量预测评分的可靠性。其中,第 3－8 行依据单一评分信息来衡量用户之间的隐性信任度,第 9－15 行主要完成评分预测与评分置信度的度量。第三部分为第 17－22 行,用于计算目标用户 u 与其他用户 $p(p\in U)$的相似性和显性信任度,选取候选邻居集合 CNS_u。其中,第 18－20 行分别用于计算用户 u 对用户 p 的相似性、显性信任度以及信任距离。第四部分为第 23－26 行,主要是基于 0-1 背包组合优化策略选取最近邻居集合 NNS_u,并计算推荐权重,进而依据最近邻居用户的评分信息完成对目标项目 i 的评分预测。

3 实验设计与性能分析

3.1 数据集

Epinions.com 是一个在线服务网站,用户在该网站上既可以为物品打分,也可以主动地对用户进行信任评价。本文实验使用的 Epinions 数据集是 Massa 等人在 2003 年 11－12 月间耗时 5 周从“Epinions.com”爬取的。该数据集包含 49289 个用户对 139738 个物品的 664824 条评分信息,评分范围为[1,5]区间内的整数值,评分密度为 0.0118%。同时,数据集还包含用户之间的 487183 条信任信息,其取值为{0,1},其中“1”表示信任,“0”表示不信任。

3.2 评价指标

为了全面评价算法,本文使用预测评分准确度和覆盖率两类评价指标[32]。预测评分准确度用于衡量预测评分与实际评分之间的贴近度,平均绝对误差(Mean Absolute Error, MAE)是最流行的评价指标。覆盖率则主要用于衡量算法挖掘长尾项目的能力,其中最基本的指标为评分覆盖率(Rating Coverage, RC)。

1)平均绝对误差 *MAE*

MAE 定义为预测评分与实际评分之间绝对偏差的平均值,其计算公式如下:

$$MAE=\frac{\sum_u\sum_i |P_{u,i}-r_{u,i}|}{N_t} \tag{24}$$

其中,N_t 表示测试集中的评分总数。然而,*MAE* 值越小,表示系统的预测评分准确度越高。

2)评分覆盖率 *RC*

RC 定义为系统可预测的评分数占所有评分数的比例,其计算公式如下:

$$RC=\frac{N_r}{N} \tag{25}$$

其中,N_r 和 N 分别表示算法可预测的评分数和系统评分总数。

3.3 实验设计

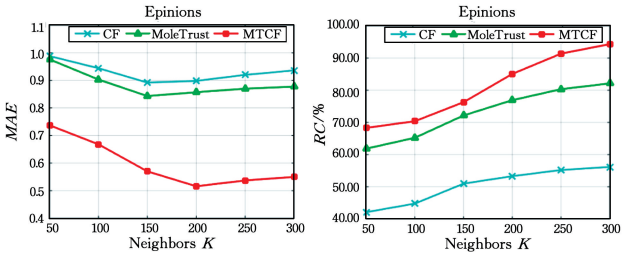
在基于 Java 的 Eclipse 仿真环境下进行五折交叉实验。为了佐证算法的性能,将基于多重信任的协同过滤推荐算法(MTCF)与 6 种主流算法进行比较,这 6 种算法分别是传统的基于用户的协同过滤算法 CF(采用 PCC 计算相似性),MoleTrust[25],MSD[21],TCF[26],Merge[28] 和 RTCF[29]。依据相应文献的描述,实验将每种算法的参数设置如下:MoleTrust 和 TCF 算法的最大信任传播距离 d_{max} 分别为 3 和 2;MSD 算法的信任阈值为 0;Merge 算法的评分相似性权重 $\alpha=0.2$,信任关系权重 $\beta=0.4$,相似性阈值 $\theta_s=0$,最大信任传播距离 $d_{max}=3$;RTCF 算法的积极因子阈值 $\alpha=0.6$,消极因子阈值 $\beta=0.5$,最近邻居数 $K=200$。在具体实验中,首先从全体用户(All users)的角度比较 CF,MoleTrust 和 MTCF 3 种算法的准确度与覆盖率,并分析最近邻居数 K 对算法性能的影响。随后,从全体用户和冷启动用户(Cold users)的角度将 MTCF 与上述 6 种算法进行综合对比分析,以验证本文算法的有效性。其中,冷启动用户的定义为:评分项目数少于 5 的用户[25]。

3.4 实验结果及性能分析

3.4.1 准确度与覆盖率的仿真分析

图 4 分别展示了 CF,MoleTrust 和 MTCF 3 种算法的平

均绝对误差 *MAE* 和评分覆盖率 *RC* 随最近邻居数 *K* 变化的趋势。



(a)平均绝对误差 *MAE* (b)评分覆盖率 *RC*

图 4 *MAE*/*RC* 的对比图

Fig. 4 Contrast of *MAE*/*RC*

如图 4(a)所示,随着 *K* 的增加,3 种算法的 *MAE* 均先呈现出下降的趋势,且在 *K*=150 时前两种算法的 *MAE* 值达到最低,而在 *K*=200 时 MTCF 算法的 *MAE* 值才降为最低;随后,3 种算法的 *MAE* 又均呈现出略微上升的趋势,并最终趋于稳定。同时,图 4(b)清晰地表明:随着 *K* 的增加,3 种算法的 *RC* 一直增大,但增速逐渐减缓。其原因是:随着 *K* 逐渐增大,近邻用户能够评分预测的项目数增多,使得评分覆盖率逐渐增长,但是近邻集合中与目标用户兴趣相近的用户占有的比例却呈现出先增大后减小的趋势,算法的 *MAE* 值的变化曲线如图 4(a)所示。

综合对比算法的 *MAE* 和 *RC* 可知:整体上 MoleTrust 的性能高于 CF,表明引入信任信息有助于提升推荐质量。同时,图 4 也清晰地表明,本文所提出的算法在 MAE 和 RC 两种评价指标上的性能均明显优于另两种算法,佐证了 MTCF 算法的有效性。

3.4.2 综合对比分析

为了充分证明算法的性能,本节分别从 All users 和 Cold users 角度基于 MAE 和 RC 两种评价指标综合对比分析了 CF,MoleTrust,MSD,TCF,Merge,RTCF 和 MTCF 7 种算法的性能,实验结果如表 2 所列。由表 2 可知,基于隐性信任机制的 MSD 和基于显性信任机制的 MoleTrust 算法的性能均优于 CF,并且 MoleTrust 在处理冷启动用户上更胜一筹。如表 2 所列,在冷启动用户上,MoleTrust 的 *MAE*(*RC*)为0.882(41.78%),而 MSD 和 CF 算法的 *MAE*(*RC*)分别为0.963(8.37%)和 1.089(3.22%)。结果表明,传统上基于用户的协同过滤算法因受评分数据稀疏性和用户冷启动问题的影响,推荐质量较差,但是引入信任机制后可以有效缓解此问题。同时,针对所有用户,TCF 能够表现出较高的 *RC*,但是在冷启动用户上的表现却不令人满意。其原因是:TCF 算法在融入信任信息之前仍是基于 CF 方法来选取近邻用户,因此受冷启动问题的束缚。

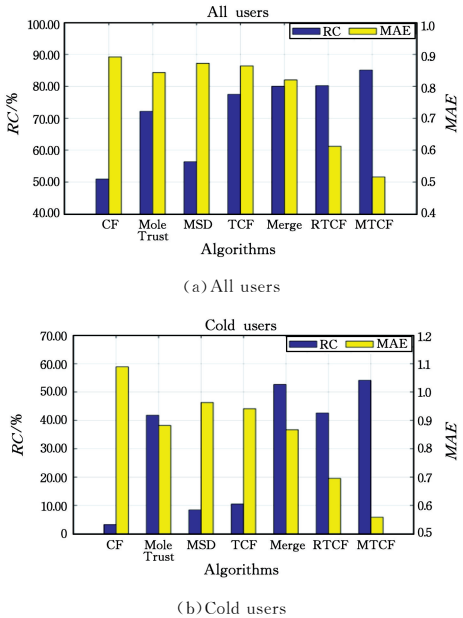
表 2 算法的综合性能

Table 2 Comprehensive performance of algorithms

	评价指标	算法						
	MAE/RC	CF	MoleTrust	MSD	TCF	Merge	RTCF	MTCF
所有用户	*MAE*	0.892	0.843	0.872	0.864	0.820	0.612	0.516
	RC/%	50.96	72.15	56.36	77.48	80.02	80.19	85.06
冷启动用户	*MAE*	1.089	0.882	0.963	0.941	0.867	0.695	0.558
	RC/%	3.22	41.78	8.37	10.45	52.66	42.57	54.12

此外,最近提出的 Merge,RTCF 和 MTCF 算法的性能在整体上都优于前 4 种算法,此现象佐证了衡量预测评分可靠性对提升算法性能产生的积极作用。同时,MTCF 算法表现出最高的预测评分准确度。如表 2 所列,针对 All users 和 Cold users,MTCF 算法的 *MAE* 为 0.516(0.558),而 Merge 和 RTCF 算法的 *MAE* 分别为 0.820(0.867)和 0.612(0.695)。但是,与 Merge 相比,MTCF 算法的 *RC* 提升较小。其原因是:在近邻选择环节中,MTCF 通过 0-1 背包组合优化策略尽可能地选取目标用户信任且与其偏好相似性较高的用户,使得预测准确度显著提升,但同时也移除了目标用户信任但与其偏好相似性较低的伪近邻,限制了 RC 的性能。

图 5 分别从 All users 和 Cold users 两个角度直观地对比了上述算法。其中,图 5(a)展示了算法在 All users 上的性能,而图 5(b)展示了算法在 Cold users 上的性能。从图 5 中可以明确地得出:本文所提出的 MTCF 算法在 All users 和 Cold users 上均能表现出最优的性能,从而充分地证明了该算法能够有效地处理评分数据稀疏性和用户冷启动问题。



(a)All users

(b)Cold users

图 5 算法综合性能的对比

Fig. 5 Comprehensive performance contrast of algorithms

结束语 本文在已有研究的基础之上,通过融合显性信

任和隐性信任各自的优势,提出了一种基于多重信任的协同过滤推荐算法,旨在更好地解决评分数据稀疏性和用户冷启动问题。在真实网络的 Epinions 数据集上,设计两组实验对本文所提算法与 6 种主流算法进行了综合对比分析。结果表明:该算法通过引入有效评分和显性信任关系,极大地缓解了数据稀疏性和冷启动问题,并且在不牺牲覆盖率的条件下显著提升了推荐准确度。同时,本文所提出的算法仍存在不足:推荐覆盖率未得到明显提升。如何更好地利用评分信息来优化算法模型,以达到同时提升推荐准确度与覆盖率的效果,是下一步研究工作的重点。

参 考 文 献

[1] JANNACH D, ZANKER M, FELFERNIG A, et al. 推荐系统[M]. 北京:人民邮电出版社,2013.

[2] AGGARWAL C C. Neighborhood-Based Collaborative Filtering[M]// Recommender Systems. Springer International Publishing,2016:29-70.

[3] YU Y, YU H T, HUANG R Y. Collaborative filtering recommendation algorithm based on entropy optimization nearest-neighbor selection[J]. Application Research of Computers, 2017,34(9):2618-2623. (in Chinese)
于阳,于洪涛,黄瑞阳. 基于熵优化近邻选择的协同过滤推荐算法[J]. 计算机应用研究,2017,34(9):2618-2623.

[4] DONG Y, ZHAO C, CHENG W, et al. A Personalized Recommendation Algorithm with User Trust in Social Network[C]// International Conference of Young Computer Scientists, Engineers and Educators. Singapore: Springer, 2016:63-76.

[5] FERNÁNDEZ-TOBÍAS I, BRAUNHOFER M, ELAHI M, et al. Alleviating the new user problem in collaborative filtering by exploiting personality information[J]. User Modeling and User-Adapted Interaction, 2016, 26(2/3):221-255.

[6] PEREIRA A L V, HRUSCHKA E R. Simultaneous co-clustering and learning to address the cold start problem in recommender systems[J]. Knowledge-Based Systems, 2015, 82(C):11-19.

[7] JIE Q, LEI C, HUI P. A solution of missing value in collaborative filtering recommendation algorithm[C]// Chinese Automation Congress(CAC), 2015. IEEE, 2015:2184-2187.

[8] CHOI K, SUH Y, YOO D. Extended Collaborative Filtering Technique for Mitigating the SparsityProblem[J]. International Journal of Computers, Communications & Control, 2016, 11(5): 631.

[9] OCEPEK U, RUGELJ J, BOSNIČ Z. Improving matrix factorization recommendations for examples in cold start[J]. Expert Systems with Applications, 2015, 42(19):6784-6794.

[10] ZHANG J, Lin Y, LIN M, et al. An effective collaborative filtering algorithm based on user preference clustering[J]. Applied Intelligence, 2016, 45(2):230-240.

[11] KONSTAS I, STATHOPOULOS V, JOSE J M. On social networks and collaborative recommendation[C]// 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2009:195-202.

[12] 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.

[13] SINGLA P, RICHARDSON M. Yes, there is a correlation:-from social networks to personal behavior on the web[C]// 17th International Conference on World Wide Web. ACM, 2008:655-664.

[14] SUN Z, HAN L, HUANG W, et al. Recommender systems based on social networks[J]. Journal of Systems and Software, 2015, 99(C):109-119.

[15] GAO P, MIAO H, BARAS J S, et al. Star:semiring trust inference for trust-aware social recommenders[C]// 10th ACM Conference on Recommender Systems. ACM, 2016:301-308.

[16] GUO G. Integrating trust and similarity to ameliorate the data sparsity and cold start for recommender systems[C]// 7th ACM Conference on Recommender Systems. ACM, 2013:451-454.

[17] GUO G, ZHANG J, THALMANN D, et al. From ratings to trust:an empirical study of implicit trust in recommender systems[C]// Proceedings of the 29th Annual ACM Symposium on Applied Computing. ACM, 2014:248-253.

[18] PAPAGELIS M, PLEXOUSAKIS D, KUTSURAS T. Alleviating the sparsity problem of collaborative filtering using trust inferences[C]// International Conference on Trust Management. Berlin: Springer, 2005:224-239.

[19] O'DONOVAN J, SMYTH B. Trust in recommender systems[C]// 10th International Conference on Intelligent User Interfaces. ACM, 2005:167-174.

[20] HWANG C S, CHEN Y P. Using trust in collaborative filtering recommendation[C]// International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Berlin: Springer, 2007:1052-1060.

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

[22] BEDI P, SHARMA R. Trust based recommender system using ant colony for trust computation[J]. Expert Systems with Applications, 2012, 39(1):1183-1190.

[23] TANG J, HU X, GAO H, et al. Exploiting Local and Global Social Context for Recommendation[C]// IJCAI. 2013:264-269.

[24] GOLBECK J A. Computing and applying trust in web-based social networks[M]. University of Maryland at College Park, 2005.

[25] MASSA P, AVESANI P. Trust-aware recommender systems[C]// 2007 ACM Conference on Recommender Systems. ACM, 2007:17-24.

[26] CHOWDHURY M, THOMO A, WADGE W W. Trust-Based Infinitesimals for Enhanced Collaborative Filtering[C]// COMAD. 2009.

[27] RAY S, MAHANTI A. Improving prediction accuracy in trust-aware recommender systems[C]// 2010 43rd Hawaii International Conference on System Sciences(HICSS). IEEE, 2010:1-9.

[28] GUO G, ZHANG J, THALMANN D. Merging trust in collaborative filtering to alleviate data sparsity and cold start[J]. Knowledge-Based Systems, 2014, 57(2):57-68.

[29] MORADI P, AHMADIAN S. A reliability-based recommendation method to improve trust-aware recommender systems[J]. Expert Systems with Applications, 2015, 42(21):7386-7398.

[30] HERNANDO A, BOBADILLA J S, ORTEGA F, et al. Incorporating reliability measurements into the predictions of a recommender system[J]. Information Sciences, 2013, 218(218):1-16.

[31] YUAN W, GUAN D, LEE Y K, et al. The small-world trust network[J]. Applied Intelligence, 2011, 35(3):399-410.

[32] ZHU Y X, LV L Y. Evaluation Metrics for Recommender Systems[J]. Journal of University of Electronics Science and Technology of China, 2012, 41(2):163-175. (in Chinese)
朱郁筱,吕琳媛. 推荐系统评价指标综述[J]. 电子科技大学学报,2012,41(2):163-175.