软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(5):1266−1284 [doi: 10.13328/j.cnki.jos.004882]

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

基于用户邻域和主题的新颖性 Web 社区推荐方法*

余 骞[2], 彭智勇[1,2], 洪 亮[3], 万言历[2]

[1](软件工程国家重点实验室(武汉大学),湖北 武汉 430072)

[2](武汉大学 计算机学院,湖北 武汉 430072)

[3](武汉大学 信息管理学院,湖北 武汉 430072)

通讯作者: 彭智勇, E-mail: peng@whu.edu.cn; 洪亮, E-mail: hong@whu.edu.cn

摘 要: 社区推荐从海量社区中为用户过滤出有价值的社区,变得越来越重要.新颖性推荐逐渐得到关注,因为单纯追求准确度的推荐结果存在局限性.已有新颖性推荐方法不适用于社区推荐,因其无法处理Web社区特性,包括社区成员用户通过交互形成的关系网络以及社区主题.提出了一种新颖性社区推荐方法 NovelRec,向用户推荐其有潜在兴趣但不知道的社区,旨在拓展用户视野和推动社区发展.NovelRec 基于用户交互网络中的邻域关系,利用用户之间在主题上的关联,计算候选社区对用户的准确度;根据用户与社区在邻域和主题上的关联,提出一种用户社区距离度量方式,并利用该距离计算候选社区的新颖度.在此基础上,NovelRec 最终进行新颖性社区推荐,并兼顾推荐结果的准确性.真实数据集上的对比实验结果表明,NovelRec 方法在新颖性上优于现有方法,同时能够保证推荐结果的准确性.

关键词: 网络社区;新颖性推荐;用户邻域;主题分类

中图法分类号: TP311

中文引用格式: 余骞,彭智勇,洪亮,万言历.基于用户邻域和主题的新颖性 Web 社区推荐方法.软件学报,2016,27(5):1266−1284. http://www.jos.org.cn/1000-9825/4882.htm

英文引用格式: Yu Q, Peng ZY, Hong L, Wan YL. Novel Web community recommendation based on user neighborhood and topic. Ruan Jian Xue Bao/Journal of Software, 2016,27(5):1266−1284 (in Chinese). http://www.jos.org.cn/1000-9825/4882.htm

Novel Web Community Recommendation Based on User Neighborhood and Topic

YU Qian[2], PENG Zhi-Yong[1,2], HONG Liang[3], WAN Yan-Li[2]

[1](State Key Laboratory of Software Engineering (Wuhan University), Wuhan 430072, China)

[2](School of Computer, Wuhan University, Wuhan 430072, China)

[3](School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract: Community recommendation has become increasingly important in sifting valuable communities from massive amounts of communities on the Internet. In recent years novel recommendation is attracting attention, because of the limitation of accurate recommendation which purely pursues accuracy. Existing novel recommendation methods are not suitable for Web community as they fail to utilize unique features of Web community, including the social network established by interactions between users, and the topics of Web community. In this paper, a novel recommendation method, NovelRec, is proposed to suggest communities that users have not seen but are potentially interested in, in order to better broaden users' horizons and improve the development of communities. Specifically, the method explores neighborhood relationships and topical associations from the aforementioned features. First, NovelRec identifies

* 基金项目: 国家自然科学基金(61232002, 61303025); 武汉科技局高新技术产业科技创新团队培养计划(2014070504020237)

Foundation item: National Natural Science Foundation of China (61232002, 61303025); Program for Innovative Research Team of Wuhan of China (2014070504020237)

收稿时间: 2015-01-09; 修改时间: 2015-03-18; 采用时间: 2015-08-05; jos 在线出版时间: 2016-01-11

CNKI 网络优先出版: 2016-01-12 11:22:18, http://www.cnki.net/kcms/detail/11.2560.TP.20160112.1122.001.html

candidate communities for users based on neighborhood relationships between users, and computes accuracy of the candidates using topical associations between users. Next, NovelRec computes novelty of the candidates based on a new metric of user-community distance, and the distance metric is defined by associations between users and communities on both user neighborhood and topic taxonomy. Finally, NovelRec balances novelty with accuracy for the candidates to improve the overall recommendation quality. Experimental results on a real data set of Douban communities show that the proposed method outperforms competitors on the recommendation novelty, and guarantees the recommendation accuracy.

Key words: Web community; novel recommendation; user neighborhood; topic taxonomy

Web 社区作为社交网络的重要组成部分正持续高速增长(http://www.nielsen.com/us/en/insights/reports/2011/social-media-report-q3.html).为解决海量 Web 社区带来的信息过载问题,社区推荐通过信息过滤帮助用户选择有价值的社区加入.已有大多数推荐方法属于准确性推荐[1−6],该类方法旨在提高推荐准确度,认为推荐结果与用户历史偏好越接近,准确度越高、推荐效果越好.然而单纯追求高准确度会降低推荐系统质量[7].在社区推荐场景下,准确性方法存在以下两个方面的问题:(1) 用户可能对准确性推荐结果不满:准确性方法旨在推荐与用户历史偏好接近的社区,同时倾向于推荐大众流行的社区[8],则用户可能因为已知被推荐的社区而产生不满.(2) 社区提供商可能对准确性推荐结果不满:准确性推荐对大众流行社区的倾向,会产生“穷者越穷富者越富”的马太效应,使得小众不流行的社区很难被推荐,而占大多数的小众社区(帕累托法则,即 80/20 法则)却有能力吸引大量用户加入[9],则社区提供商可能因为小众社区很难进入推荐列表而不满.鉴于准确性推荐方法存在的问题,新颖性推荐方法逐渐得到关注[8,10−15].新颖性社区是指用户有潜在兴趣但不知道的社区[16].图 1 通过从豆瓣社区(http://www.douban.com/group/)中抽取的实例,比较了准确性与新颖性社区推荐.

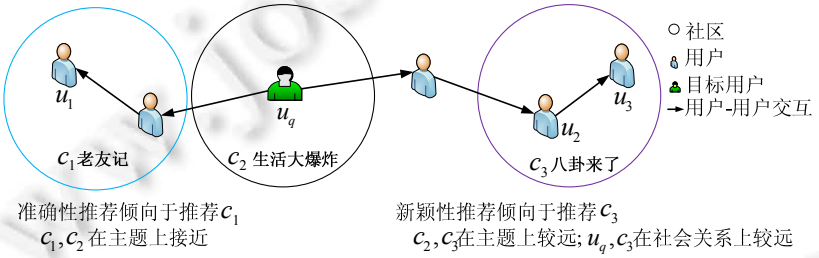


Fig.1 Illustration: Accurate vs. novel community recommendation

图 1 准确性 vs. 新颖性社区推荐示例

图 1 中目标用户 u_q 加入的社区 c_2 为其历史偏好.准确性社区推荐更可能向 u_q 推荐社区 c_1:c_1 与 u_q 的历史偏好 c_2 接近,在主题上两者同为情景喜剧.新颖性社区推荐则更可能向 u_q 推荐社区 c_3:u_q 对 c_3 有潜在兴趣,因与 u_q 存在间接交互的用户加入了该社区;u_q 可能不知道该社区,因 c_3 在主题上距离 c_1 较远,且在交互关系上距离 u_q 较远.图示新颖性社区推荐利用了 Web 社区特性,包括社区成员用户交互所形成的网络以及社区主题.新颖性社区距离目标用户历史偏好较远,更有利于拓展用户视野;新颖性社区推荐更有利于推动社区发展,因其能更多地覆盖占大多数的小众社区,而大量的小众社区有能力吸引大量用户加入.

近年来提出的新颖性推荐方法[8,10−15]利用用户与推荐项的交互,即用户-项评分矩阵进行新颖性推荐.然而已有新颖性方法不适用于 Web 社区推荐,原因如下:(1) 已有新颖性方法不能将用户交互网络用于推荐社区,而存在交互关系的用户其行为会相互影响[17,18];如图 1 中 u_2 和 u_3 加入 c_3 会对 u_q 选择社区产生影响.(2) 已有新颖性方法不能利用用户、社区通过社区主题产生的关联,而这些关联会影响用户对社区的选择[19];如图 1 和图 2(见第 2 节)所示,u_q 通过 c_2 以及社区主题(分类)与 c_1 和 c_3 产生间接关联,该关联会影响 u_q 对 c_1 和 c_3 的选择.(3) 已有方法缺少合理的新颖度定义,从而无法客观衡量项的新颖度(详见第 1.2 节).

在真实社会网络中,存在邻域关系(即 1 跳或多跳的社会关系)的用户,其行为会相互影响;而且 Web 社区之间存在主题上的关联.因此,为了进行高质量的新颖性社区推荐,本文提出基于用户邻域和主题的新颖性推荐方法 NovelRec.该方法从用户交互网络中建模用户邻域;利用社区分类和用户-社区交互建模用户之间、用户与社

区之间的主题关联.NovelRec 将邻域用户(与目标用户有邻域关系)加入的社区,作为目标用户的候选社区;通过衡量目标用户与候选社区的距离,计算候选社区的新颖度;根据目标用户与其邻域用户的主题关联,计算候选社区的准确度.在此基础上,该方法最终进行新颖性社区推荐,同时兼顾推荐结果的准确性.本文主要贡献如下:

- 提出新颖性社区推荐方法 NovelRec,通过多阶邻域交互计算,确定目标用户对候选社区的潜在兴趣,并使候选社区尽可能地远离目标用户,从而提高推荐的新颖性;同时利用邻域用户的行为及其与目标用户在主题上的关联,兼顾推荐的准确性.
- 提出一种用户社区距离度量方式,在该距离基础上定义并计算社区的新颖度;该距离考虑用户与社区间的邻域关系和主题关联,能够客观地衡量社区对用户的新颖程度.
- 提出将用户之间在邻域和主题上的关联,离线建模到邻域用户相似度矩阵,旨在减少在线推荐的计算量,并保证在线计算的低复杂度.
- 豆瓣社区数据上的实验结果表明,NovelRec 在新颖性度量标准、包括覆盖率和流行度上均优于已有方法;验证了高质量新颖性推荐需要考虑“三度影响理论”;表明 NovelRec 能够保证推荐结果的准确性.

1 相关工作

大部分已有推荐方法为准确性推荐,而单纯追求推荐准确度会降低推荐系统质量[7],近年来国内外研究者开始关注新颖性推荐,因此本节从准确性推荐和新颖性推荐两个方面阐述相关工作.

1.1 准确性推荐

准确性推荐方法旨在提高推荐准确度,认为推荐结果与用户历史偏好越接近,准确度越高、推荐效果越好.已有大多数推荐方法属于准确性推荐,其中最具代表性的是协同过滤 CF(collaborative filtering).协同过滤一般分为基于记忆(memory-based)[1−3]和基于模型(model-based)[4]两种[20].基于记忆协同过滤分为 3 类:(1) 基于推荐项的协同过滤(item-based CF)向目标用户推荐与其历史偏好相似的项(item)[1];(2) 基于用户的协同过滤(user-based CF)向目标用户推荐其相似用户选择的项[2];(3) 混合型 CF 结合以上两类 CF 思想,向目标用户推荐其相似用户选择的,同时与其历史偏好相似的项[3].基于记忆协同过滤的关键是利用用户-项评分矩阵计算用户相似度或项相似度.基于模型的 CF 在模型计算过程中融入 CF 思想,如文献[4]利用 LDA 模型从用户-社区评分矩阵中计算出用户与社区基于潜在主题的关联,并向目标用户推荐关联度高的社区;此外,该方法从用户-社区评分矩阵中挖掘出社区间关联规则,将与目标用户已评分的社区存在关联的其他社区,根据置信度推荐给目标用户;该方法验证了 LDA 相对于关联规则挖掘能够达到更好的推荐准确性.

社会化推荐利用社会化关系进行准确性推荐,该类方法将 CF 中基于相似度的关联,替代为社会化关系以降低数据稀疏产生的影响[21],从而提高推荐准确度.社会化推荐基本思想是,目标用户会对与其有社会化关系的用户所选择的项感兴趣.文献[5]将推荐数据建模为图:用户、推荐项以及用户为项加的标签均建模为图中节点,节点间相应关联建模为边,利用随机游走算法得到该图上所有节点间的关联度,依据关联度值进行推荐.文献[6]提出将目标用户历史偏好、用户之间的社会化影响两个因素统一建模,基于该模型提出一套矩阵运算方法,旨在融合上述两个因素并进行推荐.孟祥武等人[22]提出使用推土机距离实现跨推荐项的用户相似度计算,并提出融合用户信任关系和项特征的社会化推荐方法.Ma 等人在用户-项评分矩阵基础上加入显式的用户信任关系提高推荐准确度[21],并通过大量实验[23]比较了分别利用显式和隐式社会化关联进行推荐对准确度的影响.

McNee 等人[7]指出,高准确度的推荐结果可能对用户没有用处.推荐结果准确度越高,反映其与目标用户的历史偏好越接近;推荐结果中远离其历史偏好,即对目标用户“新颖”的项,会降低推荐的准确度,但新颖的项可能对用户更有用.Herlocker 等人[16]也提出了新颖性推荐的概念和一些度量标准.准确性推荐方法的目标,即单纯提高推荐准确度可能存在不足,因此近年来国内外研究者开始关注新颖性推荐.

1.2 新颖性推荐

新颖性推荐的概念由 Herlocker 等人[16]首次提出:向目标用户推荐其有潜在兴趣但不知道的项.相对于准确

性推荐,新颖性推荐能够更好地拓展用户兴趣,并使得相对小众不流行却能创造巨大价值的项更多地被推荐.近几年国内外研究者提出了一些新颖性推荐方法.Oh 等人[8]提出将用户-项评分矩阵中用户的评分模式建模为PPT(personal popularity tendency),同时为项建立相应的被评分模式;该方法设计了一个 PPT 匹配算法,项的被评分模式与目标用户的 PPT 差别越大,则该项的新颖度越高.Onuma 等人[10]将用户-项评分矩阵建模为二分图,用户和项为节点,评分关联为边;该方法利用随机游走方法计算出所有节点之间的关联度,基于该关联度定义项节点的"TANGENT"值,该值越高则项的新颖度越高.Nakatsuji 等人[11]结合用户-项评分矩阵和项的分类信息,将项所属分类与用户已评分的分类之间的距离,定义为该项对目标用户的新颖度;该方法根据项的新颖度排序生成推荐列表.文献[14,15]利用项的流行度衡量其新颖度:一个项越流行,用户知道该项的可能性越大,该项的新颖度越低.文献[12]在用户-项评分矩阵中引入评分时间,将较早评分某项的用户视为革新者(innovator),尚未评分该项的用户为其潜在跟随者(follower),认为革新者评分的项对跟随者有高新颖度;该方法视目标用户为跟随者,计算其他用户为其革新者的概率,并根据该概率值将革新者评分的项推荐给目标用户.Zhang 等人[13]构建以项为节点、项相似度为边的图,则用户已评分的项对应该图的子图;该方法将特定项节点加入目标用户的子图,计算该项的聚集因子,该因子值越大,则该项对目标用户的新颖度越高.文献[12,13]所提出的"惊喜性"推荐方法没有体现"用户反馈",而"推荐惊喜度=新颖度+用户反馈"[7],因此仍被归类为新颖性推荐方法.

上述新颖性推荐方法忽略了 Web 社区的特性:(1) 社区成员用户通过交互形成的社会关系网络;(2) 社区的主题特性以及用户、社区基于主题的关联.此外,上述方法都没有合理的新颖度定义:其中文献[8,10,12]没有定义项的新颖度,因此无法针对性衡量新颖度,而文献[11,13,14]对新颖度的定义存在缺陷.Vargas 等人[14]提出单纯用项的流行度衡量其新颖度,会导致每个项对所有用户的新颖度相同,而这种全局值不适用于个性化推荐;Nakatsuji 等人[11]提出的定义,使得属于相同分类的项对目标用户的新颖度相同,导致无法对这些项进行新颖度排序;Zhang 等人[13]提出的定义存在与文献[11]类似的问题:与目标用户子图中同样数目的相同节点存在边的项节点(项之间相似度不为 0 则存在边),其新颖度相同.此外,该 3 种定义均忽略了前述 Web 社区的特性.综上所述,以上新颖性推荐方法不适用于新颖性 Web 社区推荐.

2 问题定义

Web 社区中存在用户、社区和社区分类(taxonomy)3 类对象,以及对象之间的 3 类交互关系.用户通过交互形成的关系网络,其邻接矩阵记为 $\boldsymbol{A}$.用户-社区交互由矩阵 $\boldsymbol{R}$ 记录,$\boldsymbol{R}_{qi}$ 反映用户 u_q 在社区 c_i 内的活跃程度.社区之间通过社区分类(主题)产生的间接交互,由社区分类树 T 记录.

图 2 为 Web 社区示意图:交互网络中标出了目标用户 u_q 的各阶邻域(定义 2);矩阵 $\boldsymbol{R}$ 记录用户-社区交互,用户-社区边的粗细反映矩阵元素值的大小,即用户在社区的活跃程度;社区分类树 T 中社区为叶子节点,分支节点为社区分类,社区之间通过分类节点产生主题上的间接关联.人工分类广泛存在且呈树状结构[19],如 Amazon 为其商品提供的分类(http://www.amazon.com/gp/site-director

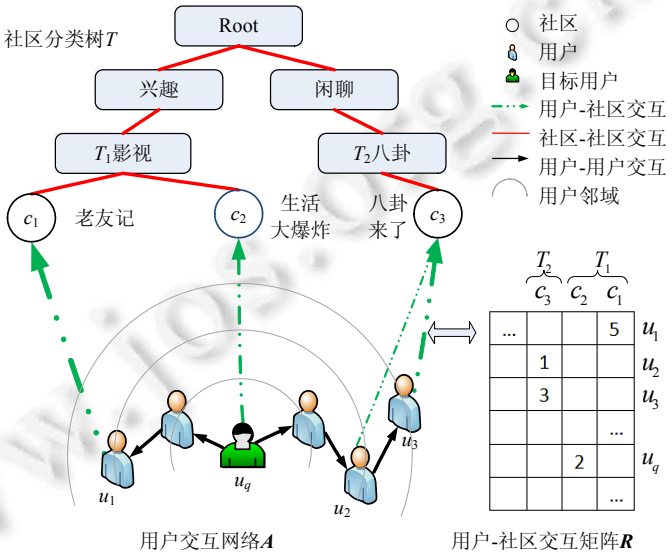


Fig.2 An illustration of Web community

图 2 Web 社区示意图

y/ref=sa_menu_top_fullstore);Web 社区提供商对社区的人工分类蕴含社区的主题信息.本文定义新颖性 Web 社区推荐问题如下:

定义 1(新颖性 Web 社区推荐). 给定用户交互网络、用户-社区交互矩阵以及社区分类树,新颖性 Web 社区推荐旨在向目标用户 u_q 推荐 k 个社区,使得这些社区对 u_q 新颖的同时保证准确性.

3 NovelRec 方法

本节首先给出 NovelRec 的方法框架;然后描述对用户邻域、邻域用户相似度和社区主题距离的建模;在离线建模基础上,描述该方法的在线推荐部分;本节最后分析了 NovelRec 方法的复杂度和用户冷启动问题.

3.1 方法框架

NovelRec 方法包括离线建模和在线推荐计算两个部分,NovelRec 方法框架如图 3 所示.离线部分从用户交互网络中建模用户邻域;结合用户-社区交互矩阵与社区分类树建模用户主题相似度;结合用户邻域和主题相似度建模邻域用户相似度矩阵;通过社区分类树建模社区-社区主题距离.邻域、邻域用户相似度以及社区-社区主题距离将用于在线推荐计算.离线部分将用户之间在邻域和主题上的关联,映射到邻域用户相似度矩阵,使得单个用户的推荐计算量分别与用户数量、社区数量线性相关,从而保证在线方法的低复杂度(详见第 3.4 节).

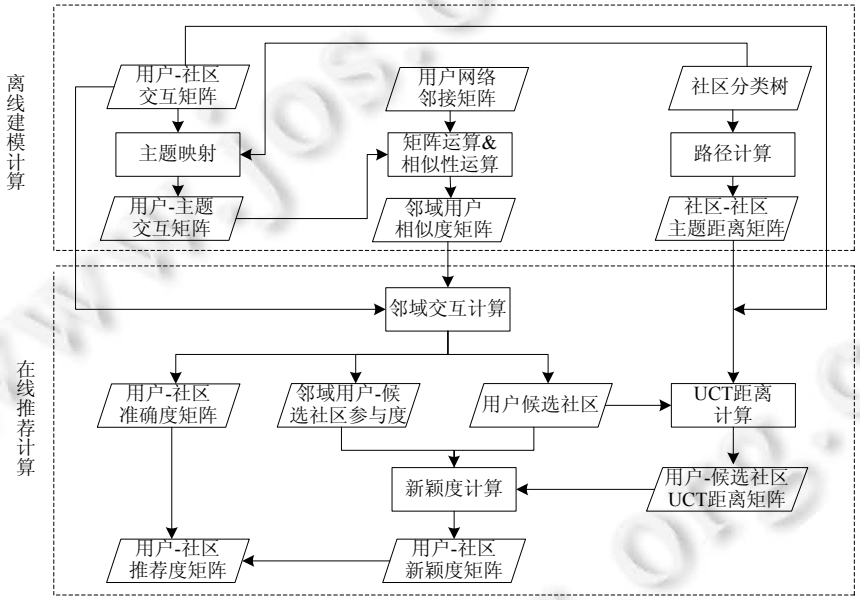


Fig.3 Framework of NovelRec

图 3 NovelRec 框架

NovelRec 在线部分首先利用邻域用户相似度和用户-社区交互矩阵进行邻域交互计算.邻域交互计算过程中,在线推荐算法确定目标用户的候选社区、邻域用户在候选社区中的参与度,以及候选社区的推荐准确度;同时利用社区-社区主题距离和用户-社区交互,分别计算目标用户与候选社区在主题和邻域上的距离,并将主题和邻域上的较小距离作为两者之间的距离.在此基础之上,在线推荐算法利用该距离与前述参与度计算候选社区的推荐新颖度,最终结合准确度和新颖度计算候选社区的推荐度.表 1 对本文用到的主要符号进行描述.

Table 1 Description of main symbols

表 1 主要符号描述

符号	描述	符号	描述
U	用户集合,$\|U\|=m$	$\boldsymbol{R}'$	用户-主题交互矩阵
C	社区集合,$\|C\|=n$	$\boldsymbol{NS}^h$	h 阶邻域用户相似度矩阵
$\boldsymbol{A}$	用户交互网络邻接矩阵	$\boldsymbol{CD}$	社区-社区主题距离矩阵
$\boldsymbol{R}$	用户-社区交互矩阵	C_q	目标用户 u_q 的候选社区集合
T	社区分类树	$\boldsymbol{AR}$	用户-社区准确度矩阵
$N^h(u_q)$	用户 u_q 的 h 阶邻域	$\boldsymbol{NR}$	用户-社区新颖度矩阵
$\boldsymbol{N}^h$	h 阶邻域矩阵	$\boldsymbol{UCTR}$	用户-社区推荐度矩阵

3.2 离线建模计算

3.2.1 用户邻域建模

本节利用用户交互网络邻接矩阵 $\boldsymbol{A}$ 建模用户邻域,直观上与用户存在交互且保持特定距离的其他用户组成该用户的邻域.Christakis 等人[17,18]提出如果用户间存在直接或间接的交互,其行为和兴趣等会相互影响.用户邻域反映用户间的交互关系,因此本文根据邻域将目标用户有潜在兴趣的社区作为其候选社区.

邻接矩阵 $\boldsymbol{A}$ 为布尔矩阵,如果用户 u_q 与 u_i 之间存在交互则 $\boldsymbol{A}_{qi}=1$, 否则 $\boldsymbol{A}_{qi}=0$. 用 $\boldsymbol{A}^{(h)}$ 表示布尔运算下 $\boldsymbol{A}$ 的 h 次方: $\boldsymbol{A}^{(1)}=\boldsymbol{A}$; ${\boldsymbol{A}^{(h)}}_{qi}=1$ 表示交互网络中存在从 u_q 到 u_i 长度为 h 的路径, ${\boldsymbol{A}^{(h)}}_{qi}=0$ 则表示不存在该路径.

定义 2(h 阶邻域). 用户交互网络中用户 u_q 的 h 阶邻域定义为用户集合 $N^h(u_q)=\left\{u_i \middle| {\boldsymbol{N}^h}_{qi}=1, u_i\in U\right\}$, 其中,$h$ 阶邻域矩阵 $\boldsymbol{N}^h=\begin{cases}\boldsymbol{A}, & h=1\\ \boldsymbol{A}^{(h)}\wedge\neg\bigvee\limits_{k=1}^{h-1}\boldsymbol{N}^k, & h=2,3,...,h_{\max}\end{cases}$.

定义 2 中 $\boldsymbol{N}^h$ 为 m 阶方阵,m 为用户数量(见表 1),$\boldsymbol{N}^h$ 非零元素记录所有用户的 h 阶邻域用户,若 $u_i\in N^h(u_q)$, 则 ${\boldsymbol{N}^h}_{qi}=1$, 否则, ${\boldsymbol{N}^h}_{qi}=0$. 当 h=2,3,…,$h_{\max}$ 时, ${\boldsymbol{N}^h}_{qi}=1$ 当且仅当 ${\boldsymbol{A}^{(h)}}_{qi}=1$ 同时 $\bigvee\limits_{k=1}^{h-1}{\boldsymbol{N}^k}_{qi}=0$:u_i 属于 u_q 的 h 阶邻域,当且仅当用户交互网络中存在从 u_q 到 u_i 长度为 h 的路径,同时不存在长度小于 h 的路径,即 u_i 只能属于 u_q 的某一个特定邻域.结合图 2,有 $u_1\in N^2(u_q)$, ${\boldsymbol{N}^2}_{q1}=1$; $u_2\in N^2(u_q)$, ${\boldsymbol{N}^2}_{q2}=1$; $u_3\in N^3(u_q)$, ${\boldsymbol{N}^3}_{q3}=1$. 如果 $u_i\in N^h(u_q)$, 本文称 u_i 为 u_q 的 h 阶邻域用户,例如 u_1 和 u_2 均为 u_q 的 2 阶邻域用户.

用户邻域由邻接矩阵 $\boldsymbol{A}$ 决定,$\boldsymbol{A}$ 的小规模变动不会显著影响 $\boldsymbol{N}^h$.不妨设邻接矩阵发生小规模变动:元素 $\boldsymbol{A}_{qi}$ 由 0 变为 1,即 u_i 变为 u_q 的 1 阶邻域用户,则仅邻域包含 u_q 的用户 u_x 受到影响: $u_q\in N^h(u_x)\rightarrow u_i\in N^{h+1}(u_x)$, 即如果 u_q 为 u_x 的 h 阶邻域用户,则 u_i 会变为 u_x 的 h+1 阶邻域用户.邻域不包含 u_q 的用户不受影响.同理,新用户邻域如果发生明显变化,只会对邻域包含该新用户的其他用户产生影响.

邻接矩阵 $\boldsymbol{A}$ 的变动在一定时间(Δt)内累积,会造成 $\boldsymbol{N}^h$ 的剧烈变动.要保证 NovelRec 方法有效,需要估算Δt 内 $\boldsymbol{A}$ 的变化程度,从而判断是否需要更新 $\boldsymbol{N}^h$:例如,Δt 时间内如果 $\boldsymbol{A}$ 的每一行元素均发生变化,则需要更新 $\boldsymbol{N}^h$. 邻接矩阵随时间的变化可根据密度幂律分布(densification power law)[24]估算.根据该分布有 $e(t)\propto n(t)^{\alpha}$, $e(t)$为 t 时刻图的边数,$n(t)$为节点数,则 t 时刻 $\alpha=\log_{n(t)}(e(t))$. 用 $n(t+\Delta t)$表示 $t+\Delta t$ 时刻的节点数,则该时刻边数为 $n(t+\Delta t)^{\alpha}$. 令 $\Delta avgIncdgr=\left(n(t+\Delta t)^{\alpha}-n(t)^{\alpha}\right)\Big/n(t+\Delta t)$, $\Delta avgIncdgr$ 为 $t+\Delta t$ 时刻节点度数的平均增加值,例如 $\Delta avgIncdgr$=1 表明宏观上图中每个节点的度平均加 1、邻接矩阵中 $n(t+\Delta t)$个元素发生变化.因此通过 $\Delta avgIncdgr$ 可估算 $\boldsymbol{A}$ 的变化并判断是否需要更新 $\boldsymbol{N}^h$,且该判断仅需观测节点数量的变化.

3.2.2 邻域用户主题相似度建模

本节利用社区分类树、用户-社区交互以及用户邻域,建模邻域用户相似度矩阵,该矩阵对保证在线推荐计算的低复杂度有重要意义(见第 3.4 节);与传统相似度建模相比,该建模能够提高相似度的有效性,并减少计算开销.

数据稀疏问题,导致基于共同评分项的传统建模方法[1−3]不能准确地反映用户之间的相似度[21]:实际系统中用户-推荐项评分矩阵密度(非零元素所占比例)通常小于1%[21],如本文所用数据集中用户-社区交互矩阵 $\boldsymbol{R}$ 密度仅为0.79%,从而共同评分项所占比重较小,会导致大量用户间的相似度无法衡量.本节利用用户之间以社区分类树为桥梁的主题关联,建模用户主题相似度,首先定义用户-主题交互矩阵如下:

定义 3(用户-主题交互矩阵 $\boldsymbol{R}'$). 定义用户-主题交互矩阵 $\boldsymbol{R}'$记录用户与社区分类的交互,其矩阵元素 $\boldsymbol{R}'_{ql}=\sum \boldsymbol{R}_{qi}, c_i \in T_l$,其中,$c_i \in T_l$表示该社区直属于分类节点 T_l,T_l在社区分类树 T 中处于 $h(T)$−1 层.

其中,$h(T)$表示分类树的高度,不妨约定根节点处于第1层,叶子节点即社区处于 $h(T)$层,则 $h(T)$−1 层分类节点反映最具体的社区主题,如图2所示的 T_1 和 T_2.定义3将用户与社区的交互,映射为用户与主题的交互.

实际应用中推荐项的数量远大于其分类的数量[19],因此相对于传统方法,基于 $\boldsymbol{R}'$建模用户主题相似度有两点优势:(1) 主题相似度的有效性更高,因为 $\boldsymbol{R}'$密度大于 $\boldsymbol{R}$ 则 $\boldsymbol{R}'$中的共同评分项多于 $\boldsymbol{R}$,如本文数据集中 $\boldsymbol{R}'$密度为8.6%而 $\boldsymbol{R}$ 密度仅为0.79%;(2) 主题相似度计算开销更小,因为开销由矩阵列的数量决定,如本文数据集中 $\boldsymbol{R}$ 列数量为1 041而 $\boldsymbol{R}'$仅为67,则利用 $\boldsymbol{R}'$能够减少约94%的计算开销.

在用户-主题交互矩阵 $\boldsymbol{R}'$基础上,本节建模邻域用户主题相似度如下:

定义4(邻域用户主题相似度). 用户与其 h 阶邻域用户之间基于社区主题的相似度由矩阵 $\boldsymbol{NS}^h$ 记录,矩阵元素 $\boldsymbol{NS}^h{}_{qi}=sim(\boldsymbol{R}'_q,\boldsymbol{R}'_i)$ 表示用户 u_q 与 u_i 之间的主题相似度,其中,$u_i \in N^h(u_q), h=1,2,...,h_{\max}$, $\boldsymbol{R}'_q$ 和 $\boldsymbol{R}'_i$ 分别为矩阵 $\boldsymbol{R}'$中 u_q 和 u_i 对应的行向量,$sim(\boldsymbol{R}'_q,\boldsymbol{R}'_i)$ 表示两向量间的相似度.

因为针对有邻域关系的用户进行相似度建模,定义 4 能够节省部分计算开销:传统方法针对全体用户计算相似度,需 $m(m-1)/2$ 次相似性计算,m 为用户数量;而邻域用户所需计算次数为所有 $\boldsymbol{N}^h$ 的非零元素个数之和,本文数据集中,当 $h_{\max}$=3时邻域用户相似度计算次数减少约20% .实验部分(见第4.1.2节)将详细讨论 $h_{\max}$的取值.

3.2.3 社区主题距离建模

Web 社区通过社区分类树的分类节点产生主题上的间接关联,本节利用该主题关联建模社区之间基于主题的距离.社区之间的主题距离将用于度量用户与社区之间的距离,以及计算社区新颖度.

任意两个社区通过社区分类树均产生间接关联.本节基于这种主题关联建模社区之间的主题距离.

定义5(社区-社区主题距离). $\forall c_i,c_j \in C$,若 $c_i \in T_i, c_j \in T_j$,c_i 与 c_j 的主题距离 $\boldsymbol{CD}_{ij}=2^{(h(T)-l-1)}$,即 T_i 与 T_j 在社区分类树 T 中的距离,T_i 与 T_j 在 T 的第 l 层拥有共同祖先节点.矩阵 $\boldsymbol{CD}$ 记录 C 中所有社区间的主题距离.

定义5中,$c_i \in T_i$ 表示 T 中社区 c_i 直属于分类节点 T_i,$h(T)$为 T 的高度.不妨约定 $\forall 1 \leqslant i \leqslant n, \boldsymbol{CD}_{ii}=0$,则 n 阶方阵 CD 为对角线元素均为0的对称矩阵,因此只需建模该矩阵的“上三角”部分.结合图2,$h(T)$=4,社区 c_1 和 c_2 在 T 的第3层拥有共同祖先节点 T_1,则 $\boldsymbol{CD}_{12}=2^{(4-3-1)}=1$; 同理,$\boldsymbol{CD}_{13}=2^{(4-1-1)}=4, \boldsymbol{CD}_{23}=2^{(4-1-1)}=4$.

3.3 在线推荐计算

本节通过邻域交互计算确定目标用户候选社区的推荐准确度(见第3.3.1节);在邻域交互计算中结合社区-社区距离矩阵,确定候选社区的推荐新颖度(见第3.3.2节);最终进行新颖性社区推荐,同时兼顾准确性(见第3.3.3节).

3.3.1 社区准确度计算

本节通过邻域交互计算,即邻域用户相似度矩阵与用户-社区交互矩阵的乘法,确定目标用户的候选社区,并计算候选社区的准确度.

本节依据 Fowler 和 Singla 等人[17,18]提出的用户行为理论确定候选社区:目标用户会对与其有直接或间接交互的用户所加入的社区感兴趣,因此 NovelRec 通过邻域交互计算将邻域用户加入,且目标用户未加入的社区作为其候选社区.用户候选社区的定义如下:

定义6(用户候选社区). 目标用户 u_q 的候选社区集合 $C_q=\{c_i \mid \exists 1 \leqslant k \leqslant m, \boldsymbol{R}_{ki} \neq 0 \wedge \boldsymbol{NS}^{\boldsymbol{h}}{}_{qk} \neq 0 \wedge \boldsymbol{R}_{qi}=0\}$,其中,$1 \leqslant q,k \leqslant m, m=|U|, c_i \in C, h=1,2,\ldots,h_{\max}$.

定义 6 中 $\boldsymbol{R}_{ki}\neq 0$ 表示用户 u_k 已加入社区 c_i(存在交互); $\boldsymbol{NS}^{\boldsymbol{h}}{}_{qk}\neq 0$ 表示 u_k 是 u_q 的直接或间接交互用户(邻域用户),且两用户的主题相似度不为 0; $\boldsymbol{R}_{qi}=0$ 表示 u_q 未加入社区 c_i.定义 6 表明,确定候选社区同时考虑邻域关系和主题关联.假设 u_q 的全体邻域用户中仅 u_k 加入 c_i,但 u_k 与 u_q 无主题关联(两者相似度为 0),则 c_i 不会成为 u_q 的候选社区,因为 c_i 与 u_q 无主题关联其准确性无法保证,根据本文问题定义 c_i 不应被推荐.

邻域交互计算通过 1 次矩阵乘法 $\boldsymbol{NS}^{\boldsymbol{h}}\cdot\boldsymbol{R}$,可确定由 h 阶邻域决定的用户候选社区,命题如下:

命题 1. $\forall 1\leqslant q\leqslant m,1\leqslant i\leqslant n$, 如果 $\boldsymbol{R}_{qi}=0$, 则 $(\boldsymbol{NS}^{\boldsymbol{h}}\cdot\boldsymbol{R})_{qi}\neq 0\rightarrow c_i\in C_q$.

证明: $\forall 1\leqslant q\leqslant m,1\leqslant i\leqslant n,(\boldsymbol{NS}^{\boldsymbol{h}}\cdot\boldsymbol{R})_{qi}=\sum_{k=1}^{m}\boldsymbol{NS}^{\boldsymbol{h}}{}_{qk}\boldsymbol{R}_{ki}$, $(\boldsymbol{NS}^{\boldsymbol{h}}\cdot\boldsymbol{R})_{qi}\neq 0\rightarrow\exists 1\leqslant k\leqslant m,\boldsymbol{NS}^{\boldsymbol{h}}{}_{qk}\neq 0\wedge\boldsymbol{R}_{ki}\neq 0$. 给定 $\boldsymbol{R}_{qi}=0$, 则 $(\boldsymbol{NS}^{\boldsymbol{h}}\cdot\boldsymbol{R})_{qi}\neq 0\rightarrow\exists 1\leqslant k\leqslant m,\boldsymbol{R}_{ki}\neq 0\wedge\boldsymbol{NS}^{\boldsymbol{h}}{}_{qk}\neq 0\wedge\boldsymbol{R}_{qi}=0$, 有 $(\boldsymbol{NS}^{\boldsymbol{h}}\cdot\boldsymbol{R})_{qi}\neq 0\rightarrow c_i\in C_q$. □

在确定用户候选社区后,利用协同过滤[2]以及社会化推荐[21]思想进行准确度计算:u_q 的邻域用户与其相似度越大,且邻域用户在候选社区 c_i 中越活跃,则 c_i 对 u_q 的准确度越高.同时考虑到 u_q 对 c_i 的兴趣来源于其邻域用户的影响,而影响力会随着 h (距离)的增加而降低[17].综合以上因素,定义用户-社区准确度如下:

定义 7(用户-社区准确度). 定义候选社区 c_i 对目标用户 u_q 的准确度 $\boldsymbol{AR}_{qi}=\sum_{h=1}^{h_{\max}}\sum_{u_k\in N^h(u_q)}\boldsymbol{NS}^{\boldsymbol{h}}{}_{qk}\boldsymbol{R}_{ki}\big/2^{(h-1)}$, 用户-社区准确度矩阵 $\boldsymbol{AR}=\sum_{h=1}^{h_{\max}}\boldsymbol{NS}^{\boldsymbol{h}}\cdot\boldsymbol{R}/2^{(h-1)},\forall 1\leqslant q\leqslant m,1\leqslant i\leqslant n,\boldsymbol{R}_{qi}\neq 0\rightarrow\boldsymbol{AR}_{qi}=0$.

定义 7 中 $\boldsymbol{AR}_{qi}$ 的公式反映前述协同过滤思想:邻域用户 u_k 与 u_q 相似度即 $\boldsymbol{NS}^{\boldsymbol{h}}_{qk}$ 越大、u_k 在候选社区 c_i 中越活跃即 $\boldsymbol{R}_{ki}$ 越大,则候选社区准确度 $\boldsymbol{AR}_{qi}$ 越大; $\sum_{h=1}^{h_{\max}}$ 表示考虑 u_q 的 1–$h_{\max}$ 阶邻域,不同阶邻域用户行为对准确度的影响随衰减因子 $1/2^{(h-1)}$[17]变化.用户-社区准确度矩阵 $\boldsymbol{AR}$ 为 $m\times n$ 矩阵,由邻域交互计算 $\boldsymbol{NS}^{\boldsymbol{h}}\cdot\boldsymbol{R}$ 得到;对 $\boldsymbol{AR}$ 的约束 $\forall 1\leqslant q\leqslant m,1\leqslant i\leqslant n,\boldsymbol{R}_{qi}\neq 0\rightarrow\boldsymbol{AR}_{qi}=0$, 旨在满足命题 1 的条件 $\boldsymbol{R}_{qi}=0$, 从而保证 $\boldsymbol{AR}$ 中所有非零元素均为候选社区对用户的推荐准确度,换言之,用户与其已加入社区对应的矩阵元素值必为 0.定义 7 对 $\boldsymbol{AR}_{qj}$ 的定义与矩阵 $\boldsymbol{AR}$ 定义保持一致:根据 $\boldsymbol{AR}$ 定义, $\boldsymbol{AR}_{qi}=\sum_{h=1}^{h_{\max}}\sum_{k'=1}^{m}\boldsymbol{NS}^{\boldsymbol{h}}{}_{qk'}\boldsymbol{R}_{k'i}\big/2^{(h-1)}$, 而 $\boldsymbol{NS}^{\boldsymbol{h}}{}_{qk'}=1\rightarrow u_{k'}\in N^h(u_q)$, 则 $\sum_{k'=1}^{m}\boldsymbol{NS}^{\boldsymbol{h}}{}_{qk'}\boldsymbol{R}_{k'i}=\sum_{u_k\in N^h(u_q)}\boldsymbol{NS}^{\boldsymbol{h}}{}_{qk}\boldsymbol{R}_{ki}$, 即 $\boldsymbol{AR}_{qi}=\sum_{h=1}^{h_{\max}}\sum_{u_k\in N^h(u_q)}\boldsymbol{NS}^{\boldsymbol{h}}{}_{qk}\boldsymbol{R}_{ki}\big/2^{(h-1)}$.

用户-社区准确度计算过程详见第 3.3.3 节中算法 2.

3.3.2 社区新颖度计算

本节计算候选社区对目标用户的推荐新颖度:根据用户与社区之间在邻域和主题上的关联,提出一种用户社区距离度量方式;在该距离基础上计算候选社区对目标用户的新颖度.

新颖性社区指用户有潜在兴趣但不知道的社区[16].目标用户对邻域交互计算所确定的候选社区(见定义 6)有潜在兴趣,因此新颖度计算从以下 3 个方面衡量目标用户不知道候选社区的可能性:(1) 目标用户与候选社区的距离;(2) 邻域用户在候选社区中的参与度;(3) 候选社区的流行度.新颖度计算假设候选社区与目标用户距离越远、加入社区的邻域用户数量越少(参与度低)、社区越不流行,目标用户不知道候选社区的可能性越大,候选社区对目标用户的新颖度就越高.

不妨假设图 2 中 c_3 为 u_q 的候选社区,u_q 既能通过已加入的社区 c_2 经社区分类节点到达 c_3,也可以通过其邻域用户 u_2 和 u_3 到达 c_3.依托社区分类树,u_q 与 c_3 存在主题距离;基于用户邻域,u_q 与 c_3 存在邻域距离.

定义 8(用户-社区主题距离). 用户 u_q 与其候选社区 c_i 的主题距离 TD_{qi},定义为社区 c_i 与 u_q 已加入社区的最小主题距离, $TD_{qi}=\min\{\boldsymbol{CD}_{ik}\big|\forall 1\leqslant k\leqslant n,\boldsymbol{R}_{qk}\neq 0\}$.

社区-社区主题距离矩阵 $\boldsymbol{CD}$(见第 3.2.3 节)记录 C 中所有社区之间的主题距离.如图 2 所示,u_q 加入社区 c_2, 则 u_q 与社区 c_1 的主题距离 $TD_{q1}=\boldsymbol{CD}_{12}=1$, u_q 与 c_3 的主题距离 $TD_{q3}=\boldsymbol{CD}_{32}=4$.

定义 9(用户-社区邻域距离). 用户 u_q 与其候选社区 c_i 的邻域距离 ND_{qi}, 定义为 u_q 所有加入社区 c_i 的邻域用户与 u_q 的最短距离, $ND_{qi}=\min\{h\big|u_k\in N^h(u_q)\wedge\boldsymbol{R}_{ki}\neq 0,h=1,2,...,h_{\max}\}$.

其中, $\boldsymbol{R}_{ki}\neq 0$ 表示 u_k 与社区 c_i 存在交互.以图 2 为例,u_q 加入 c_1 的邻域用户为 u_1 且 $u_1\in N^2(u_q)$, 则 $ND_{q1}=2$; u_q

加入 c_3 的邻域用户为 u_2 和 u_3 且 $u_2 \in N^2(u_q), u_3 \in N^3(u_q)$, 则 $ND_{q3} = \min\{2,3\} = 2$.

目标用户与其候选社区通过社区分类树和用户邻域分别可达,即同时存在主题距离与邻域距离,结合社区主题与用户邻域定义用户-社区 UCT(user-community-taxonomy)距离如下:

定义 10(用户-社区 UCT 距离). 用户 u_q 与其候选社区 c_i 的 UCT 距离 $\boldsymbol{D}_{qi}$, 定义为两者主题距离与邻域距离的最小值, $\boldsymbol{D}_{qi} = \min\{TD_{qi}, ND_{qi}\}$, 矩阵 $\boldsymbol{D}$ 记录 U 中所有用户与其候选社区的 UCT 距离.

仍以图 2 为例,对 c_1 而言,如果仅考虑邻域关系该社区与 u_q 距离为 2,而 c_1 与 u_q 的主题距离为 1,则 $\boldsymbol{D}_{q1} = \min\{TD_{q1}, ND_{q1}\} = 1$. 对 c_3 而言,该社区与 u_q 主题距离为 4,而邻域距离为 2,则 $\boldsymbol{D}_{q3} = \min\{TD_{q3}, ND_{q3}\} = 2$.

给定通过 h 阶邻域交互计算确定社区 c_j 为用户 u_i 的候选社区,算法 1 计算两者之间的 UCT 距离,该算法将被第 3.3.3 节的算法 2 调用.算法 1 第 1 步如果 $\boldsymbol{D}_{ij} = 0$, 说明 $\boldsymbol{D}_{ij}$ 尚未计算,因为任意用户与社区的 UCT 距离不小于 1,而 0 为 $\boldsymbol{D}$ 在算法 2 中的初始值;第 2 步~第 4 步根据定义 10 计算 $\boldsymbol{D}_{ij}$.第 1 步如果 $\boldsymbol{D}_{ij} \neq 0$, 说明该值在 $h'(h'<h)$ 阶邻域已经计算,且 $\boldsymbol{D}_{ij} = \min\{TD_{ij}, h'\}$, 因为 $h'<h$,所以无需再次计算 $\boldsymbol{D}_{ij}$,跳过第 2 步~第 4 步,保留已经计算的 $\boldsymbol{D}_{ij}$ 值.

算法 1. 用户-社区 UCT 距离算法 UCTDistance(u_i,c_j,h).

输入: 社区 c_j 为用户 u_i 的候选矩阵,邻域的阶 h.

输出: 用户 u_i 与社区 c_j 的 UCT 距离 $\boldsymbol{D}_{ij}$.

1. if $\boldsymbol{D}_{ij} = 0$
2. | for k where $\boldsymbol{R}_{ik} \neq 0$
3. | | $TD_{ij} = \min\{\boldsymbol{CD}_{jk}\}$
4. | $\boldsymbol{D}_{ij} = \min\{TD_{ij}, h\}$

UCT 距离综合考虑用户与社区之间在邻域和主题上的关联,能够客观地反映目标用户与候选社区之间的距离.衡量候选社区新颖度的另外两个方面是社区流行度,以及邻域用户在候选社区中的参与度.社区流行度由社区成员数决定.邻域用户在目标用户候选社区中的参与度由矩阵 $\boldsymbol{NH}$ 记录, $\boldsymbol{NH}_{qi}$ 表示加入社区 c_i 的 u_q 的邻域用户数量,即 $\boldsymbol{NH}_{qi} = |Set_{qi}|$, 而 $Set_{qi} = \{u_k \mid \boldsymbol{NS}^{\boldsymbol{h}}_{qk} \neq 0 \wedge \boldsymbol{R}_{ki} \neq 0, h = 1,2,...,h_{\max}\}$, Set_{qi} 为加入 c_i($\boldsymbol{R}_{ki} \neq 0$)且属于 u_q 邻域的用户集合, Set_{qi} 中的用户同时需要保证与 u_q 存在主题关联 ($\boldsymbol{NS}^{\boldsymbol{h}}_{qk} \neq 0$).

约定参与度矩阵 $\boldsymbol{NH}$ 归一化后的矩阵表示为 $\overline{\boldsymbol{NH}}$,即 $\forall 1 \leqslant q \leqslant m, 1 \leqslant i \leqslant n, \overline{\boldsymbol{NH}}_{qi} = \boldsymbol{NH}_{qi} \Big/ \sum_{k\in[1,n]} \boldsymbol{NH}_{qk}$; 同理,UCT 距离矩阵 D 经过行归一化后的矩阵为 $\overline{\boldsymbol{D}}, \overline{\boldsymbol{D}}_{qi} = \boldsymbol{D}_{qi} \Big/ \sum_{k\in[1,n]} \boldsymbol{D}_{qk}$; 约定$|c_i|$表示该社区成员数,对$|c_i|$归一化得到 $\overline{|c_i|} = c_i \Big/ \sum_{k\in[1,n]} c_k$. 在前述问题基础上,本节定义用户-社区新颖度如下:

定义 11(用户-社区新颖度). 定义候选社区 c_i 对目标用户 u_q 的新颖度 $\boldsymbol{NR}_{qi} = -\log_2\left(\overline{\boldsymbol{NH}}_{qi} \cdot \overline{|c_i|} \Big/ \overline{\boldsymbol{D}}_{qi}\right)$, 用户-社区新颖度矩阵 $\boldsymbol{NR}$ 记录 U 中所有用户与其候选社区之间的推荐新颖度.

因为 $\boldsymbol{NH}_{qi}$, $|c_i|$和 $\boldsymbol{D}_{qi}$ 的取值范围不同,所以对其进行不改变值分布的归一化处理.新颖度公式使用熵的自信息形式($-\log_2 X$),旨在体现对新颖度的假设:候选社区与目标用户距离越远(即 $\overline{\boldsymbol{D}}_{qi}$ 越大)、加入社区的邻域用户数量越少(即 $\overline{\boldsymbol{NH}}_{qi}$ 越小)、社区越不流行(即 $\overline{|c_i|}$ 越小),目标用户不知道候选社区的可能性越大,社区新颖度就越高.

在邻域交互计算(矩阵乘法 $\boldsymbol{NS}^{\boldsymbol{h}} \cdot \boldsymbol{R}$)过程中,增加 1 次加法运算即可得到邻域用户-候选社区参与度;用户-社区 UCT 距离由算法 1 得到;社区流行度为已知信息.用户-社区新颖度的计算过程见第 3.3.3 节算法 2.

3.3.3 社区推荐度计算

本节计算候选社区对目标用户的推荐度,该值融合准确度和新颖度,使得候选社区对目标用户新颖的同时保证准确性.候选社区的推荐度 UCTR 定义如下:

定义 12(用户-社区推荐度 UCTR). 定义候选社区 c_i 对目标用户 u_q 的推荐度 $\boldsymbol{UCTR}_{qi} = \overline{\boldsymbol{NR}}_{qi} \Big/ \overline{\boldsymbol{AR}}_{qi}$, 其中,

$\overline{\boldsymbol{AR}}_{qi}=\boldsymbol{AR}_{qi}\big/\sum_{k\in[1,n]}\boldsymbol{AR}_{qk}$, $\overline{\boldsymbol{NR}}_{qi}=\boldsymbol{NR}_{qi}\big/\sum_{k\in[1,n]}\boldsymbol{NR}_{qk}$, 矩阵 $\boldsymbol{UCTR}$ 记录所有用户与其候选社区之间的推荐度.

根据社区推荐度定义,如果 c_i 新颖度较高则 *UCTR* 值较高,准确度较高则 *UCTR* 值较低.该定义使用除法融合准确度和新颖度;此外,准确度和新颖度均经过归一化处理,以保证计算在相同数值范围内进行.

NovelRec 在线推荐计算的核心是邻域交互计算 $\boldsymbol{NS}^{\boldsymbol{h}}\cdot\boldsymbol{R}$, 候选社区的确定,候选社区的准确度、新颖度以及推荐度计算都基于该稀疏矩阵乘法.下面给出 NovelRec 在线推荐计算算法.该算法分两部分:(1) 第 1 步~第 8 步为邻域交互计算 $\boldsymbol{NS}^{\boldsymbol{h}}\cdot\boldsymbol{R}$, 即在各阶邻域进行 1 次矩阵乘法;(2) 第 9 步~第 12 步根据邻域交互计算的结果,计算候选社区的新颖度和推荐度.

算法第 1 步表示对用户交互网络中 1~$h_{\max}$ 阶邻域进行邻域交互计算;算法第 3 步~第 6 步为稀疏矩阵乘法运算.

第 3 步按行扫描矩阵 $\boldsymbol{NS}^{\boldsymbol{h}}$; 第 4 步定位 $\boldsymbol{NS}^{\boldsymbol{h}}$ 第 i 行中非零元素的列号 k;第 5 步将列号 k 作为矩阵 $\boldsymbol{R}$ 的行号并寻找 $\boldsymbol{R}$ 第 k 行中非零元素;第 6 步为矩阵乘法结果计算.算法第 5 步的判断条件 $\boldsymbol{R}_{ij}=0$ 使得 $\boldsymbol{NS}^{\boldsymbol{h}}\cdot\boldsymbol{R}$ 可确定用户的候选社区(命题 1),并且使得当前用户已加入社区对应的矩阵 $\boldsymbol{AR}$ 元素值必为 0(定义 7);第 6 步累加各阶邻域的 $\boldsymbol{NS}^{\boldsymbol{h}}(i,k)\cdot\boldsymbol{R}(k,j)\big/2^{(h-1)}$ 得到 $\boldsymbol{AR}_{ij}$.算法第 7 步~第 8 步仍在稀疏矩阵乘法过程中.第 7 步计算 u_i 的 h 阶邻域用户在候选社区 c_j 中的参与度.第 8 步通过最多$|\boldsymbol{R}_i|$次比较运算得到两者之间的 UCT 距离(详见第 3.4 节复杂度分析).第 9 步~第 12 步基于邻域交互计算的结果,计算候选社区的新颖度和推荐度.算法第 9 步按行扫描矩阵 $\boldsymbol{NH}$ 和 $\boldsymbol{D}$;第 10 步定位 $\boldsymbol{NH}$ 和 $\boldsymbol{D}$ 的第 i 行中的非零元素,如果 $\boldsymbol{NH}_{ik}\neq 0$, 则 c_k 为候选社区;第 11 步计算社区新颖度,其中包括归一化运算;第 12 步计算社区的推荐度.

算法 2. NovelRec 在线推荐算法.

输入: 邻域用户相似度矩阵 $\boldsymbol{NS}^{\boldsymbol{h}}$, 用户社区交互矩阵 $\boldsymbol{R}$.

输出: 用户-社区推荐度矩阵 $\boldsymbol{UCTR}$.

```
1.  for h=1 to h_max
2.      矩阵 AR,NR,NH,D 所有元素初始化为 0
3.      for i=1 to m do
4.          for k where  NS^h_ik ≠ 0
5.              for j where  R_kj ≠ 0 ∧ R_ij = 0
6.                  AR(i,j) = AR(i,j) + NS^h(i,k)·R(k,j)/2^(h-1)
7.                  NH(i,j) = NH(i,j) + 1
8.                  UCTDistance (u_i,c_j,h)
9.  for i=1 to m
10.     for k where  NH_ik ≠ 0
11.         NR_ik = -log2(NH̄_ik · |c_k|/D̄_ik)
12.         UCTR_ik = NR̄_ik / AR̄_ik
```

研究表明,实际系统中用户和推荐项的数据密度一般小于 1%[21];本文数据集中 $\boldsymbol{NS}^1$ 密度为 0.013%,$\boldsymbol{NS}^2$ 密度为 0.6%,$\boldsymbol{NS}^3$ 为 9%,$\boldsymbol{R}$ 为 0.79%.算法 2 将 $\boldsymbol{NS}^{\boldsymbol{h}}$ 和 $\boldsymbol{R}$ 处理为稀疏矩阵,使用稀疏矩阵乘法技术 SpGEMM[25,26] 计算 $\boldsymbol{NS}^{\boldsymbol{h}}\cdot\boldsymbol{R}$, 并采用行压缩技术 CRS 进行矩阵存储.

3.4 NovelRec复杂度分析

NovelRec 离线部分的用户邻域建模使用稀疏矩阵乘法技术 SpGEMM 处理邻接矩阵 $\boldsymbol{A}$(本文数据集中 $\boldsymbol{A}$ 的密度 0.013%); $\boldsymbol{A}^{(h)}$ 的计算决定邻域建模的时间复杂度为 $O(flops+nnz(\boldsymbol{A})+m)$[25],其中,*flops* 指矩阵运算中非零算术运算次数,$nnz(\boldsymbol{A})$为矩阵 $\boldsymbol{A}$ 非零元素数量,m 为 $\boldsymbol{A}$ 的列数量.邻域用户相似度矩阵 $\boldsymbol{NS}^{\boldsymbol{h}}$ 建模的时间复杂度为 $O(\max\{\max\{nnz(\boldsymbol{N}^{\boldsymbol{h}})\},nnz(\boldsymbol{R})\})$, 其中, $nnz(\boldsymbol{N}^{\boldsymbol{h}})$ 和 $nnz(\boldsymbol{R})$ 表示 $\boldsymbol{N}^{\boldsymbol{h}}$ 和 $\boldsymbol{R}$ 的非零元素数量;该部分复杂度由 $\boldsymbol{N}^{\boldsymbol{h}}$ 和 $\boldsymbol{R}$ 的矩阵遍历决定.社区-社区主题距离矩阵 $\boldsymbol{CD}$ 建模的时间复杂度为 $O(n^2)$,n 为社区数量.

在线用户-社区UCT距离计算(算法1)最多需要进行$|\boldsymbol{R}_i|$次比较运算,$\boldsymbol{R}_i$表示用户-社区交互矩阵$\boldsymbol{R}$的行向量,$|\boldsymbol{R}_i|$表示用户u_i已加入社区的数量;用户交互社区的数量为常量,因此该算法时间复杂度为$O(1)$.算法1不需要存储中间运算结果,空间复杂度为$O(1)$.在线推荐计算(算法2)的时间复杂度为$O(flops2+\max\{nnz(\boldsymbol{NS}^h)\}+n)$[25],其中,$flops$ 2指算法2中非零算术运算次数,包括该算法第6步~第8步中的基本算术运算,$nnz(\boldsymbol{NS}^h)$为h阶邻域相似度矩阵非零元素数量,n为社区数量.算法2需要存储的中间计算结果包括矩阵$\boldsymbol{AR}$,$\boldsymbol{NH}$,$\boldsymbol{NR}$和$\boldsymbol{D}$.其中,$\boldsymbol{AR}$,$\boldsymbol{NH}$和$\boldsymbol{NR}$密度相同,非零元素位置严格对应,用$nnz(\boldsymbol{AR})$表示$\boldsymbol{AR}$的非零元素数量;矩阵$\boldsymbol{D}$的非零元素数量为$(1+n)n/2$,n为社区数量.则算法2的空间复杂度为$O(nnz(\boldsymbol{AR})+n^2)$.

算法2中任意两个用户的在线推荐计算过程相互独立:根据该算法,用户u_i候选社区的确定,及其准确度、新颖度以及推荐度在$\boldsymbol{NS}^h_i\cdot\boldsymbol{R}$基础上得到,$\boldsymbol{NS}^h_i$表示$\boldsymbol{NS}^h$中$u_i$对应的行向量;换言之,$\forall 1\leqslant i,j\leqslant m$,$u_i$和$u_j$的推荐计算过程相互独立.该独立性质由矩阵乘法性质决定,也因为邻域用户主题相似度矩阵$\boldsymbol{NS}^h$建模了用户之间在邻域、社区以及社区主题上的关联.在线推荐计算的用户间独立性,使得单个用户的推荐计算量与用户数量线性相关.如果实际应用中只需对部分用户(比如k个用户)进行计算,则在线部分只需要全局计算量的k/m.此外,根据矩阵乘法性质,单个用户的推荐计算量与社区数量也线性相关.上述线性相关性质,表明对邻域用户相似度的离线建模能够保证在线推荐计算的低复杂度.

3.5 用户冷启动分析

推荐系统向新增加用户推荐社区时会遇到冷启动问题:由于新用户缺乏数据,对其进行推荐难以达到较好的效果.冷启动用户可大致分为两类:第1类为无数据的冷启动用户,即新用户与社区和其他用户均无交互;第2类为数据稀疏的冷启动用户,该类新用户与社区和其他用户的交互数据少.针对第1类无数据用户,根据社区流行度(加入社区的成员数量)向其推荐热门社区,因为新用户更有可能对热门流行社区感兴趣.针对第2类冷启动用户,如果NovelRec推荐给该类用户的社区不足k个(k为推荐列表长度,见定义1),提出如下策略补足推荐列表:不妨设u_q为第2类用户,且由于缺乏数据NovelRec仅向其推荐了$k'(k'<k)$个社区.该策略定位u_q已加入社区所属的社区分类集合$\{T_1,T_2,\ldots,T_a\}$,将属于该分类集合的社区按流行度排序,取出前$(k-k')$个社区补充到u_q的推荐列表.第4.4节将分析冷启动用户的推荐效果.

4 实 验

本节由实验数据分析、推荐准确度分析和推荐新颖度分析这3部分组成.第4.1节对本文使用的豆瓣社区数据进行了展示和分析,包括用户交互网络、用户-社区交互矩阵以及社区成员数等;同时确定用户邻域阶数的上界$h_{\max}$(见定义2).第4.2节和第4.3节根据相应度量标准,比较了NovelRec与3种其他推荐方法[4,13,15]在推荐准确性和推荐新颖性上的表现.本文实验环境如下:Intel(R) Core(TM) i5-2320 3.00GHz处理器,4GB内存,Windows 7旗舰版64位操作系统,使用Java 1.6语言编写程序,数据库为Mysql 5.0.

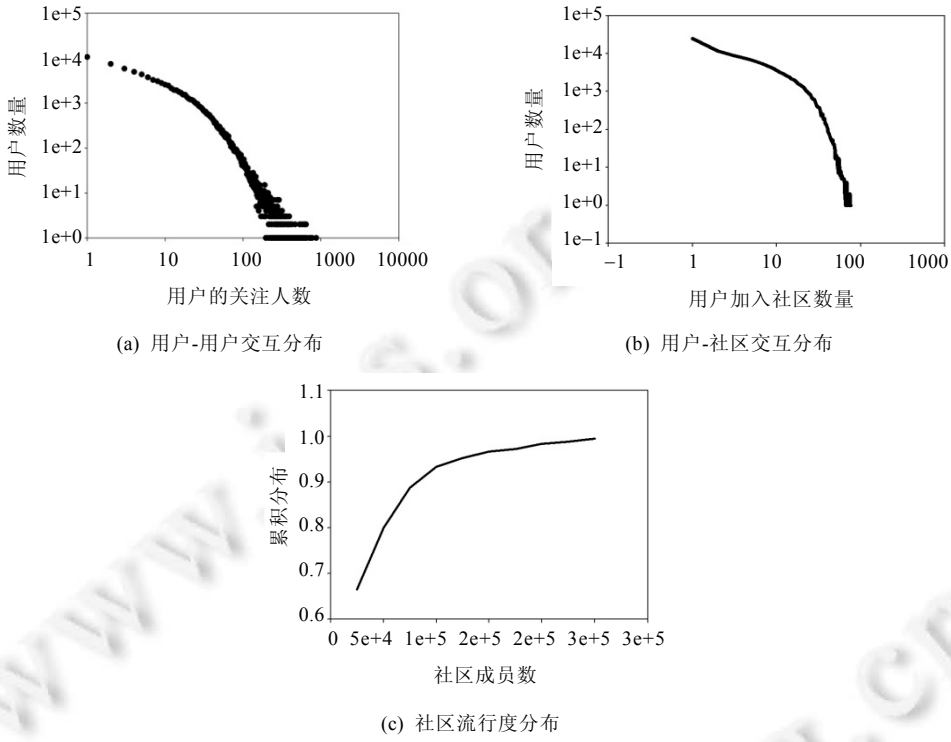
4.1 实验数据分析

4.1.1 数据集分析

豆瓣社区是典型的社会网络应用,目前拥有超过70 000 000用户和320 000社区.本文数据集包括:115 962个用户,1 041个社区,949 226条用户-社区交互记录,1 841 964条用户-用户交互记录(实验部分使用的用户交互关系为“关注”关系),包含根节点在内共76个分类节点,社区分类树T中处于$h(T)-1$层(见定义3),即第3层的分类节点数为67.用户-社区交互记录形式为(*Uid*,*Cid*,*Rating*),*Uid*和*Cid*分别为匿名化后用户和社区的标识符,*Rating*为用户与社区交互的频度.用户-用户交互记录的形式为(*Uid*,*Uid*),即连接用户节点的边.社区记录的形式为(*Cid*,*Pop*,*Tid*),其中,*Pop*为社区流行度,*Tid*为该社区直属的分类节点.社区分类记录的形式为(*Tid*,*P_Tid*),其中,*Tid*,*P_Tid*分别为该分类节点及其父亲分类节点的标识符.

实验部分沿用表1中的符号描述.用户交互网络的节点数为115 962,边数为1 841 964,其中用户最多关注870个用户,最少关注1个用户.本节对用户关注的分布进行了统计分析,如图4(a)所示,横轴为用户关注人数,纵

轴为拥有相应关注人数的用户数量;在 log-log 标度下其符合幂律(power law)分布,表明用户交互网络具有典型的社会网络特征.图 4(b)为用户-社区交互分布,横轴所示的用户加入社区数量即用户交互的社区数量,纵轴为加入相应数量社区的用户数量,用户-社区交互矩阵 ***R*** 的密度为 0.79%,其中有 24 746 个用户仅与 1 个社区交互,而单个用户最多与 87 个社区交互;在 log-log 标度下,用户与社区的交互也符合幂律(power law)分布.图 4(c)为社区流行度分布,坐标轴为线性标度,横坐标为社区成员数即社区流行度,纵轴为累积分布;社区成员数最大为 349 700,最小为 6.如图所示,80% 的社区其成员数不超过 50 000;而成员数超过 100 000 的社区仅占全体社区的 7%,可见本数据集中的社区流行度分布符合帕累托法则,即 80/20 法则.



(a) 用户-用户交互分布 (b) 用户-社区交互分布

(c) 社区流行度分布

Fig.4 Analysis of the Douban community dataset

图 4 豆瓣社区数据集分析

4.1.2 邻域参数确定

用户邻域阶数上界 $h_{\max}$(见定义 2)的选择对 NovelRec 方法有重要意义:$h_{\max}$ 影响用户候选社区集合的大小和在线推荐算法(见算法 2)的计算量,也对候选社区推荐度计算结果产生影响.本节根据实验分析选择 $h_{\max}$=3,同时验证了进行高质量新颖性社区推荐需要考虑“三度影响理论”[17].

定义 6 依据社会网络用户行为理论[17,18],将目标用户 u_q 的 h 阶邻域用户加入且 u_q 其未加入的社区 c_i 作为其候选社区,即 c_i 成为候选社区由于 u_q 受其邻域用户的影响.本节通过随机抽样和全局统计,分析用户的候选社区集合随邻域阶数上界 $h_{\max}$ 的变化,见表 2.表中第 1 列对用户真实 ID 做了匿名处理;第 2 列~第 5 列分别为不同 $h_{\max}$ 取值下用户候选社区集合 C_q 的大小.

表 2 中前 3 行为 3 个随机抽取用户的候选社区数量与 $h_{\max}$ 的对应关系,表中数据说明 1 跳~3 跳关系($h_{\max}$=3)对用户的候选社区影响明显,同时影响程度随跳数增加而减小.为不失一般性,表 2 的第 4 行显示了不同 $h_{\max}$ 下矩阵 ***AR*** 密度的变化;***AR*** 的密度变化能够反映用户候选社区的变化,因为 ***AR*** 非零元素即为用户与其候选社区的准确度(见定义 7).数据反映 ***AR*** 的密度受 1 跳~3 跳关系影响较大,而 4 跳关系的影响很小;也反映抽样数据(表 2 前 3 行)与统计数据基本保持一致.以上分析表明新颖性社区推荐需要考虑“三度影响理论”[17]:3 跳以

内的邻域关系(h_{max}=3)对推荐影响明显;邻域关系的影响随跳数增加而减小.

Table 2 $|C_q|$ vs. h_{max}
表 2 $|C_q|$ vs. h_{max}

目标用户	$\|C_q\|(h_{max}=1)$	$\|C_q\|(h_{max}=2)$	$\|C_q\|(h_{max}=3)$	$\|C_q\|(h_{max}=4)$
u_1	210	747	1 003	1 019
u_2	340	885	992	1 000
u_3	196	788	988	1 008
全体用户	9.3%	48.2%	83.5%	84.9%

4.2 推荐准确性分析

本节比较了 NovelRec 与 3 种对比方法[4,13,15]在推荐准确性上的表现.为便于叙述,实验部分将文献[4]所提方法称为 Orkut,文献[13]称为 Auralist,文献[15]称为 TRelation(total relation).Orkut 为准确性 Web 社区推荐方法,利用 LDA 模型从矩阵 $\boldsymbol{R}$ 中计算出用户与社区基于潜在主题的关联,并向目标用户推荐关联度高的社区.Auralist 为新颖性推荐方法,构建以项为节点、项相似度为边的图,则用户已评分的项对应该图的子图.该方法将特定项节点加入目标用户的子图,并以该项的聚集因子作为项对目标用户的新颖度.TRelation 为新颖性推荐方法,通过挖掘用户网络中存在的潜在社团,确定用户的候选新颖性社区,并根据潜在社团中用户的行为计算社区的新颖度.以上 3 种对比方法没有充分利用用户与社区的交互、用户交互网络和社区主题进行推荐.实验结果表明,NovelRec 方法能够保证推荐的准确性.

本节选择通用的“leave-one-out”[16]方法衡量推荐的准确性.对用户-社区交互矩阵 $\boldsymbol{R}$ 中的每个用户,随机从其加入的社区中抽出 1 个(leave one out)作为测试集,$\boldsymbol{R}_{leave}$ 为训练集,$\boldsymbol{R}_{leave}$ 表示矩阵 $\boldsymbol{R}$ 被抽掉 115 962($\boldsymbol{R}$ 中的用户数)个元素后的矩阵. $\boldsymbol{R}_{leave}$ 的推荐结果中,被抽出的测试社区在对应用户的推荐列表中的排名越高,说明推荐方法越准确.

因为 $\boldsymbol{R}$ 中有 24 746 个用户仅交互 1 个社区(见第 4.1.1 节),所以 $\boldsymbol{R}_{leave}$ 中用户数降为 91 216,社区数降为 1 034,则 91 216 个用户的推荐列表中最多出现 1 034 个不同的独立社区.Orkut 和 Auralist 均用到 LDA 模型,为其统一设置 LDA 参数如下:迭代次数为 1 000;主题数 67,与豆瓣数据集保持一致;α=2,β=0.5.推荐准确性对比结果如图 5 和图 6 所示.

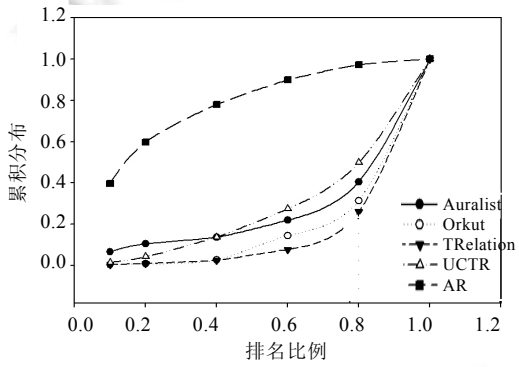


Fig.5 Macro view of recommendation accuracy
图 5 推荐准确性的宏观视图

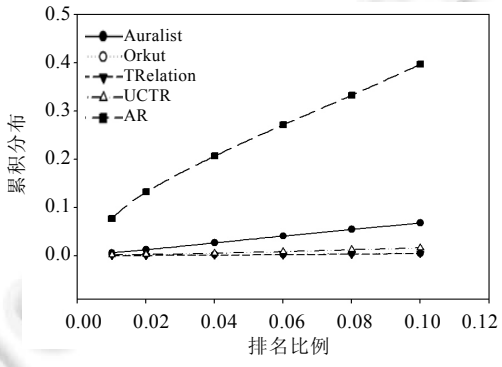


Fig.6 Micro view of recommendation accuracy
图 6 推荐准确性的微观视图

图 5 为测试社区排名累积分布的宏观视图,其中,$\boldsymbol{AR}$ 为第 3.3.1 节计算的推荐准确度,横坐标 0.1 代表测试社区在对应用户的推荐列表中排名前 10%;100%表示该社区在列表中排名最后.$\boldsymbol{AR}$ 的准确性远高于对比方法,其 40%的测试社区排名比例在前 10%,而 NovelRec(即 UCTR)有 1.6%的测试社区排名比例在前 10%,Auralist 有 6.7%的测试社区排名比例在前 10%,Orkut 为 0.46%,TRelation 为 0.42%.宏观视图下排名比例在前 40%时,NovelRec 在准确性上优于 Orkut 和 TRelation,但稍逊于 Auralist;而当排名比例超过 40%时,NovelRec 在准确性

上优于所有对比方法.图 5 中代表 5 种方法的曲线汇聚在点(100%,100%),因为测试社区在所有用户的推荐列表中的排名一定在前 100% (最差情况下排名最后).图 6 为测试社区排名累积分布的微观视图,显示测试社区排名比例在前 1%~10%的累积分布,其结果与图 5 所展示的结果保持一致:排名比例在前 1%~10%时,NovelRec 在准确性上优于 Orkut 和 TRelation,但稍逊于 Auralist,***AR*** 的准确性远高于所有对比方法.

图 5 和图 6 的结果表明,虽然 NovelRec 在推荐准确性上部分情况(测试社区排名比例在前 40%)下稍逊于 Auralist,但其他情况下均优于 Auralist;且 NovelRec 全局优于 Orkut 和 TRelation.这说明 NovelRec 能够保证推荐的准确性.同时观察到 ***AR*** 在准确性上优势明显,说明第 3.3.1 节由邻域交互计算确定的社区准确度效果良好.NovelRec 的准确性在部分情况下的劣势,由用户推荐度 UCTR 的计算公式(见定义 12)造成,即社区的准确性排名被其新颖性排名拉低.虽然 NovelRec 牺牲了一定的推荐准确性,但极大地提高了推荐新颖性(详见第 4.3 节).

4.3 推荐新颖性分析

本节比较了上述方法在推荐新颖性上的表现,使用的度量标准为流行度和覆盖率[16].流行度标准指用户推荐列表中的前 k 个社区的流行度分布,该分布与社区流行度分布(如图 4(c)所示)和 k 的取值有关;社区流行度由其成员数衡量.在流行度标准下,高流行度的社区被推荐得越多,推荐方法的新颖性就越低.覆盖率是指进行 top-k 推荐时,用户被推荐的所有社区中独立社区的数量与独立社区总数的比值,该比值越大说明被推荐的独立社区越多.在覆盖率标准下,该比值越大,则推荐方法越新颖.实验结果表明,在新颖性度量标准,包括覆盖率和流行度上,NovelRec 均明显优于已有方法.

在“leave-one-out”比较方法下,推荐用户数为 91 216,社区总数为 1 034;社区流行度分布如图 4(c)所示,80%的社区其成员数不超过 50 000;而成员数超过 100 000 的社区仅占全体社区的 7%.图 7~图 9 显示了在 top-1、top-5 和 top-10 推荐下,各种方法的推荐结果在社区流行度上的分布,其中,***NR*** 表示第 3.3.2 节所计算的社区新颖度.

图 7 显示,在 top-1 推荐情景下,NovelRec,TRelation 和 ***NR*** 在流行度上的表现接近:这 3 种方法推荐的 91 216(被推荐用户数)个社区中 97.7%的流行度不超过 25 000;而全体社区中有 66.6%其流行度不超过 25 000.相比之下,Auralist 所推荐社区中仅 5%其流行度在 25 000 以下;Orkut 相应比例为 0.超过 80%的社区其流行度不超过 50 000,***NR*** 推荐的社区中有 99.4%其流行度在 50 000 以下,NovelRec 有 98.8%在 50 000 以下,TRelation 相应比例为 99.3%;相比之下,Auralist 为 21.3%,而 Orkut 仍为 0.可见,在 top-1 推荐情景下,NovelRec,TRelation 和 ***NR*** 在流行度上的表现基本接近,且远远好于其他对比方法;Auralist 在流行度上优于 Orkut.图 7 同时显示,在进行 top-1 推荐时,Orkut 推荐的社区中有约 15%的社区,其流行度超过 300 000,而这样的社区在全体社区中所占比例小于 0.5%.

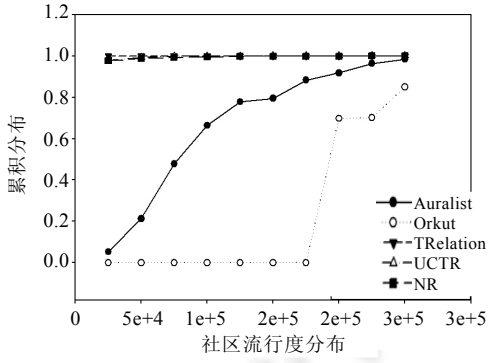


Fig.7 Popularity distribution of top-1 recommendation

图 7 Top-1 推荐的流行度分布

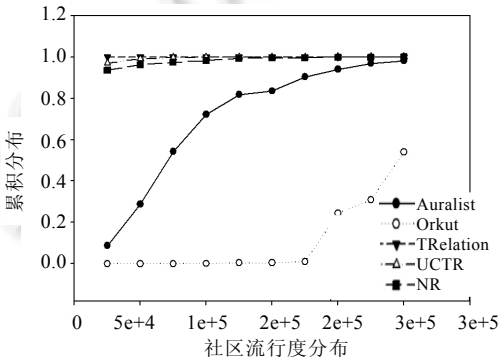


Fig.8 Popularity distribution of top-5 recommendation

图 8 Top-5 推荐的流行度分布

图 8 显示在 top-5 推荐下,各方法在流行度上的表现,此时方法推荐的社区数为 91216×5=456080.图 8 与图 7 显示的情况大体一致,但在 top-5 推荐下,NovelRec 在流行度上稍好于 ***NR***,与 TRelation 接近.NovelRec 推荐的社区中有 97%其流行度在 25 000 以下,***NR*** 为 93.7%;NovelRec 推荐的社区中有 99.2%其流行度在 50 000 以下,***NR*** 为 96.3%.NovelRec,TRelation 和 ***NR*** 在流行度上明显优于其他对比方法,Auralist 优于 Orkut.图 9 为 top-10 推荐下各方法在流行度上的比较,此时推荐的社区数为 912 160.图 9 显示的情况与图 8 基本一致,进一步验证了 NovelRec 和 ***NR*** 在流行度上优于 Orkut 和 Auralist.

图 10 显示各方法在覆盖率上的对比结果,横轴的 K 为推荐列表的长度,纵轴为覆盖率的值.NovelRec 在覆盖率上略优于 ***NR***,这两种方法在覆盖率上明显优于其他对比方法;Auralist 优于 Orkut 和 TRelation.

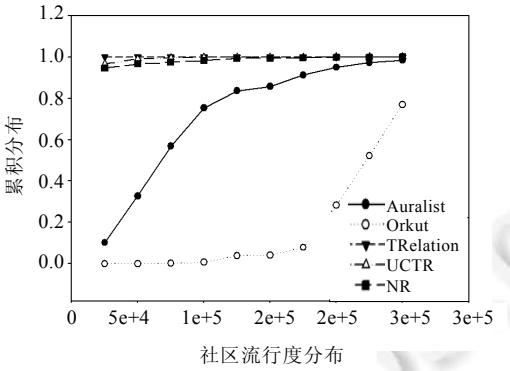


Fig.9 Popularity distribution of top-10 recommendation

图 9 Top-10 推荐的流行度分布

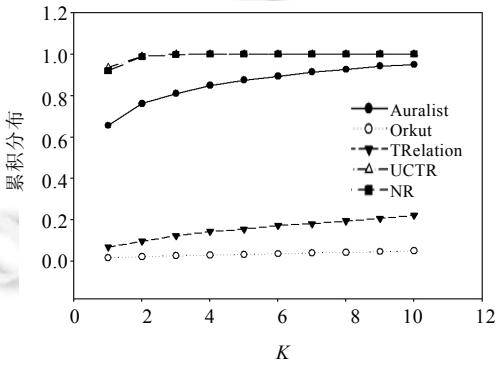


Fig.10 Coverage distribution of top-K recommendation

图 10 Top-K 推荐的覆盖率分布

当进行 top-1 推荐时,91 216 个社区被推荐,NovelRec 推荐了 972 个独立社区,覆盖率为 0.936;***NR*** 推荐独立社区 952 个,覆盖率为 0.921;Auralist 覆盖率为 0.656;TRelation 为 0.069 3;Orkut 推荐独立社区 18 个,覆盖率 0.017 4.进行 top-3 推荐时,91216×3=273648 个社区被推荐,NovelRec 推荐的独立社区数为 1 034,即覆盖率达到 100%;***NR*** 推荐独立社区 1 031 个,覆盖率为 0.997;Auralist 覆盖率 0.81,而 Orkut 覆盖率仍仅为 0.027,推荐独立社区 28 个.进行 top-4 推荐时,364 864 个社区被推荐,NoverRec 覆盖率维持在 100%;***NR*** 的覆盖率达到 100%;Auralist 覆盖率为 0.849.进行 top-10 推荐时,912 160 个社区被推荐,Auralist 覆盖率为 0.95,仍只推荐到 982 个独立社区;而 Orkut 推荐的独立社区仅为 52 个.TRelation 方法虽然有良好的流行度分布,但覆盖率低.

基于以上流行度和覆盖率的对比分析,NovelRec 方法在推荐新颖性上明显优于其他对比方法.根据第 3.3.2 节的新颖度定义,该方法推荐在邻域和主题上距离目标用户较远的社区,从而能够推荐更多在邻域和主题上较为分散的社区,并得到较高的覆盖率;该方法推荐邻域用户较少参与同时成员数较少的社区,从而能够推荐更多相对小众、流行度低的社区.NovelRec 相对于 3 种对比方法,能够更充分地利用 Web 社区特性,从而达到更优的新颖性.

在以上对比分析中,部分情况下,NovelRec 在新颖性上甚至优于 ***NR***,如进行 top-5 推荐时,NovelRec 在流行度上稍好于 ***NR***:NovelRec 推荐的社区中有 97%其成员数在 25 000 以下,***NR*** 只有 93.7%的推荐社区其成员数小于 25 000.该情况侧面反映了定义 12 会拉低社区的准确度并提高了其新颖度,部分解释了 NovelRec 的准确性在部分情况下差于 Auralist 的现象,.

4.4 NovelRec性能分析

本节分析了 NovelRec 方法性能随数据量大小以及数据维度的变化.首先分析了 NovelRec 性能随用户数量的变化,如图 11 所示,横轴为用户数量;左纵轴表示 NovelRec 为相应数量用户进行推荐的运行时间,单位为 ms;右纵轴为相应的社区推荐数量.该部分实验随机取样了 12 组用户,统计了不同用户数量下,NovelRec 的运行时

间和推荐社区数量的变化;其中用户数量从 108 逐渐增加到 30 464,对应构成曲线的 12 个数据点.比较两条曲线的变化趋势,社区推荐数量的增幅超过运行时间,说明 NovelRec 可扩展性强,该可扩展性由方法性质决定,即第 3.4 节所述,单个用户的推荐计算量与用户数量和社区数量均线性相关.

本节接下来分析 NovelRec 方法性能随不同数据维度的变化.如前文所述,NovelRec 方法利用了 3 个维度的数据,分别为用户-社区交互、用户-用户交互和社区-社区交互维度.应对不同的维度组合,基于 NovelRec 有 4 个基准方法,见表 3. NovelRec-CC 方法仅考虑社区-社区交互维度,该方法的推荐度计算公式去除了与用户相关的维度数据.NovelRec-UC 仅考虑用户-社区交互维度,其推荐度计算公式去除了用户-用户交互维度的数据.NovelRec-UUC 不考虑社区-社区交互维度,则计算 $\boldsymbol{AR}_{qi}$ 时去除了用户相似度,计算 $\boldsymbol{NR}_{qi}$ 时去除了主题距离和社区流行度.NovelRec-UCC 不考虑用户-用户交互维度,其推荐度计算公式中去除了邻域和邻域距离.

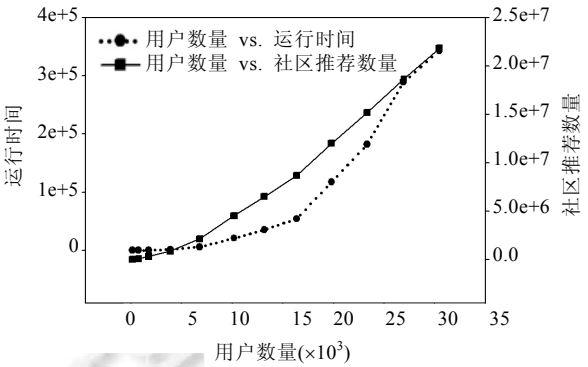


Fig.11 NovelRec efficiency vs. number of users

图 11 NovelRec 效率 vs. 用户数量

Table 3 Baseline methods based on NovelRec

表 3 基于 NovelRec 的基准方法

方法名称	用户-社区交互	用户-用户交互	社区-社区交互	方法描述	推荐度计算公式
NovelRec-CC	–	–	✓	仅考虑社区-社区交互维度	$\boldsymbol{UCTR}_{qi}=\overline{\lvert c_i\rvert}$
NovelRec-UC	✓	–	–	仅考虑用户-社区交互维度	$\boldsymbol{UCTR}_{qi}=\boldsymbol{AR}_{qi}=\sum \boldsymbol{S}_{qk}\boldsymbol{R}_{ki}$
NovelRec-UUC	✓	✓	–	考虑两种维度的组合	$\boldsymbol{AR}_{qi}=\sum_{h=1}^{h_{\max}}\sum_{u_k\in N^h(u_q)}\boldsymbol{N}^{\boldsymbol{h}}{}_{qk}\boldsymbol{R}_{ki}\big/2^{(h-1)}$, $\boldsymbol{NR}_{qi}=-\log_2\left(\overline{\boldsymbol{NH}}_{qi}\big/\overline{\boldsymbol{ND}}_{qi}\right)$, $\boldsymbol{UCTR}_{qi}=\overline{\boldsymbol{NR}}_{qi}\big/\overline{\boldsymbol{AR}}_{qi}$
NovelRec-UCC	✓	–	✓	考虑两种维度的组合	$\boldsymbol{AR}_{qi}=\sum \boldsymbol{S}_{qk}\boldsymbol{R}_{ki}$, $\boldsymbol{NR}_{qi}=\sum -\log_2\left(\lvert\overline{c_i}\rvert\big/\boldsymbol{TD}_{qi}\right)$, $\boldsymbol{UCTR}_{qi}=\overline{\boldsymbol{NR}}_{qi}\big/\overline{\boldsymbol{AR}}_{qi}$

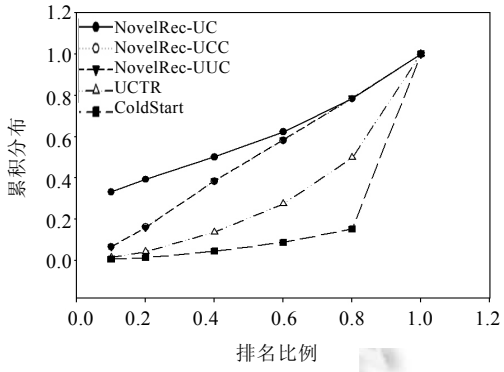


Fig.12 Macro view of recommendation accuracy: NovelRec vs. baselines

图 12 推荐准确性宏观视图:NovelRec vs. 基准方法

根据第 4.2 节和第 4.3 节所述度量标准,图 12~图 14 比较了 NovelRec 及其基准方法的推荐准确性和新颖性.图 12 为推荐准确性的宏观视图.因为 NovelRec-UC 退化为准确性推荐方法,所以其准确性最高;NovelRec- UCC 和 NovelRec-UUC 的准确性接近;前述 3 种基准方法在准确性上略优于 NovelRec.NovelRec-CC 缺少用户信息,不能构建测试集和训练集,无法衡量其准确性,因此没有出现在图 12 中.第 4.2 节已分析 NovelRec 准确性稍差的原因:社区的准确性排名被其新颖性排名拉低.在流行度和覆盖率标准下,NovelRec 均明显优于其基准方法.NovelRec 能够达到更优的推荐新颖性,因其考虑了社区推荐场景下更丰富维度的

数据,充分利用了 Web 社区特性.NovelRec-CC 在图 13、图 14 中对应与横轴“平行”的线条,因为该方法对所有用户均推荐流行度最高的 k 个社区.流行度 top-10 社区的成员数量均超过 300 000,因此图 13 中 NovelRec-CC 对应线条的纵坐标均为 0;图 14 中当 K=1 时其纵坐标为 1/1034,当 K=10 时其纵坐标为 10/1034=0.97%.

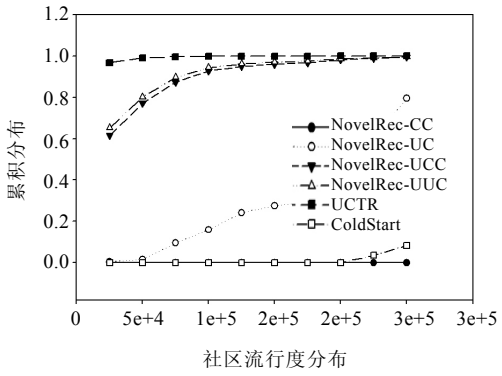


Fig.13 Popularity distribution of top-10 recommendation: NovelRec vs. baselines

图 13 Top-10 推荐的流行度分布: NovelRec vs. 基准方法

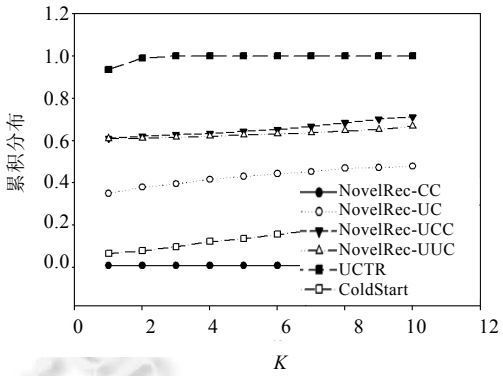


Fig.14 Coverage distribution of top-K recommendation: NovelRec vs. baselines

图 14 Top-K 推荐的覆盖率分布: NovelRec vs. 基准方法

本节同时对冷启动用户的推荐效果进行了实验,构建了由 100 个(实验数据集中用户数量的 0.1%)冷启动用户组成的集合:其中,10 个用户为第 3.5 节所述第 1 类无数据的冷启动用户;剩余 90 个为第 2 类数据稀疏用户.为保证数据稀疏,从本文数据集中随机挑选 90 个用户:保证这些用户仅与 1 个用户存在交互,同时,其中 10 个用户仅加入 2 个社区,另外 80 个仅加入 1 个社区.利用第 3.5 节所述策略对该用户集合进行推荐.

图 12~图 14 统计了冷启动用户的推荐结果(对应 ColdStart 曲线).第 4.2 节所述“leave-one-out”方法需要构建测试社区集合,因此准确性衡量只对加入 2 个社区的 10 个用户有效.图 12 表明,冷启动用户的准确度结果比 NovelRec 及其基准方法要差,因为该 10 个用户的数据过于稀疏,导致测试社区的排名很低.图 13 表明,冷启动用户的推荐社区大多是流行度高的社区,该结果由第 3.5 节所述推荐策略造成.因为大多数推荐结果集中在流行度高的社区,所以图 14 显示新用户推荐结果的覆盖率较低.

社区推荐场景下从数据性质的角度,有以下几种特殊用户:(1) 冷启动用户;(2) 仅与社区交互的用户,该类用户仅产生与社区交互的数据,不与其他用户交互;(3) 仅与其他用户交互的用户,该类用户与其他用户进行互动,但不加入社区.针对冷启动用户的推荐策略第 3.5 节已述;基准方法 NovelRec-UC 仅考虑用户-社区交互维度,NovelRec-UCC 仅考虑用户-社区以及社区-社区交互维度,因此该两种方法可针对上述第 2 种用户进行推荐,用户-用户交互维度数据对该两种方法没有影响;同理 NovelRec-UUC 仅考虑用户-社区以及用户-用户维度数据,可针对上述第 3 种用户进行推荐,该方法不受社区-社区交互维度数据影响.NovelRec 方法集中了上述基准方法的优点,可针对不同的特殊用户进行推荐.同时 NovelRec 方法针对用户数量的变化可扩展性强,如图 11 所示.因此,NovelRec 方法拥有良好的鲁棒性.

5 结束语

NovelRec 的社区新颖度计算充分考虑了用户之间的邻域关系、用户与社区之间基于邻域和主题的关联,以及社区自身特性,以提高推荐新颖性;社区准确度计算根据协同过滤和社会化推荐思想,结合邻域用户相似度与邻域用户行为,以保证推荐准确性.NovelRec 通过将用户之间的邻域关系和主题关联,离线建模到邻域用户相似度矩阵中,使得单个用户的推荐计算量分别与用户数量、社区数量线性相关,从而保证方法的低复杂度.实验结果表明,NovelRec 方法在推荐新颖性上优于已有方法:针对不同的推荐列表长度,该方法能够提升 5.1%~29%

的覆盖率;针对不同的流行度,该方法能够提升 1.6%~90%.实验结果同时表明,该方法能够保证推荐结果的准确性.

References:

[1] Deshpande M, Karypis G. Item-Based top-*n* recommendation algorithms. ACM Trans. on Information Systems, 2004,22(1): 143−177. [doi: 10.1145/963770.963776]

[2] Castagnos S, Boyer A. A client/server user-based collaborative filtering algorithm: Model and implementation. In: Brewka G, Coradeschi S, Perini A, eds. Proc. of the European Conf. on Artificial Intelligence. Riva del Garda: IOS Press, 2006. 617−621.

[3] Wang J, De Vries AP, Reinders MJT. Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In: Efthimiadis NE, Dumais ST, Hawking D, Jarvelin K, eds. Proc. of the Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM, 2006. 501−508. [doi: 10.1145/1148170.1148257]

[4] Chen WY, Chu JC, Luan J, Bai H, Wang Y, Chang EY. Collaborative filtering for orkut communities: Discovery of user latent behavior. In: Quemada J, Leon G, Maarek Y, Nejdl W, eds. Proc. of the Int'l Conf. on World Wide Web. New York: ACM, 2009. 681−690. [doi: 10.1145/1526709.1526801]

[5] Konstas I, Stathopoulos V, Jose JM. On social networks and collaborative recommendation. In: Allan J, Aslam JA, Sanderson M, Zhai CX, Zobel J, eds. Proc. of the Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM, 2009. 195−202. [doi: 10.1145/1571941.1571977]

[6] Jiang M, Cui P, Liu R,Yang Q, Wang F, Zhu W, Yang SQ. Social contextual recommendation. In: Chen XW, Lebanon G, Wang HX, Zaki MJ, eds. Proc. of the ACM Int'l Conf. on Information and Knowledge Management. New York: ACM, 2012. 45−54. [doi: 10.1145/2396761.2396771]

[7] McNee SM, Riedl J, Konstan JA. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In: Olson GM, Jeffries R, eds. Proc. of the CHI 2006 Extended Abstracts on Human Factors in Computing Systems. New York: ACM, 2006. 1097−1101. [doi: 10.1145/1125451.1125659]

[8] Oh J, Park S, Yu H, Song M, Park ST. Novel recommendation based on personal popularity tendency. In: Cook DJ, Pei J, Wang W, Zaiane OR, Wu XD, eds. Proc. of the Int'l Conf. on Data Mining. Vancouver: IEEE Computer Society, 2011. 507−516. [doi: 10.1109/ICDM.2011.110]

[9] Yin HZ, Cui B, Li J, Yao JJ, Chen C. Challenging the long tail recommendation. In: Alonso G, Cetintemel U, Dalvi N, Freire J, Korth H, Sacan A, Tatbul N, Tung A, eds. Proc. of the VLDB Endowment. Istanbul: VLDB Endowment. 2012. 896−907. [doi: 10.14778/2311906.2311916]

[10] Onuma K, Tong HH, Faloutsos C. TANGENT: A novel, "Surprise me", recommendation algorithm. In: Elder JF IV, Fogelman-Soulie F, Flach PA, Zaki MJ, eds. Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2009. 657−666. [doi: 10.1145/1557019.1557093]

[11] Nakatsuji M, Fujiwara Y, Tanaka A, Uchiyama T, Fujimura K, Ishida, T. Classical music for rock fans: Novel recommendations for expanding user interests. In: Huang J, Koudas N, Jones GJF, Wu XD, Collins-Thompson K, An AJ, eds. Proc. of the ACM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2010. 949−958. [doi: 10.1145/1871437.1871558]

[12] Kawamae N. Serendipitous recommendations via innovators. In: Crestani F, Marchand-Maillet S, Chen HH, Efthimiadis EN, Savoy J, eds. Proc. of the Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2010. 218−225. [doi: 10.1145/1835449.1835487]

[13] Zhang Y C, Seaghdha DO, Quercia D, Jambor T. Auralist: Introducing serendipity into music recommendation. In: Adar E, Teevan J, Agichtein E, Maarek Y, eds. Proc. of the Int'l Conf. on Web Search and Data Mining. New York: ACM Press, 2012. 13−22. [doi: 10.1145/2124295.2124300]

[14] Vargas S, Castells P. Rank and relevance in novelty and diversity metrics for recommender systems. In: Mobasher B, Burke RD, Jannach D, Adomavicius G, eds. Proc. of the ACM Conf. on Recommender Systems. New York: ACM Press, 2011. 109−116. [doi: 10.1145/2043932.2043955]

[15] Yu Q, Peng ZY, Hong L, Liu B, Peng HP. Novel community recommendation based on a user-community total relation. In: Bhowmick SS, Dyreson CE, Jensen CS, Lee ML, Muliantara A, Thalheim B, eds. Proc. of the Int'l Conf. on Database Systems for Advanced Applications. Bali: Springer-Verlag, 2014. 281−295. [doi: 10.1007/978-3-319-05813-9_19]

[16] Herlocker JL, Konstan JA, Terveen LG, Riedl JT. Evaluating collaborative filtering recommender systems. ACM Trans. on Information Systems, 2004,22(1):5−53. [doi: 10.1145/963770.963772]

[17] Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. The New England Journal of Medicine, 2007,357(4):370−379. [doi: 10.1056/NEJMsa066082]

[18] Singla P, Richardson M. Yes, there is a correlation:-from social networks to personal behavior on the Web. In: King I, Baeza-Yates, RA, eds. Proc. of the Int'l Conf. on World Wide Web. New York: ACM Press, 2008. 655−664. [doi: 10.1145/1367497.1367586]

[19] Ziegler CN, Lausen G, Schmidt-Thieme L. Taxonomy-Driven computation of product recommendations. In: Grossman DA, Gravano L, Zhai CX, Herzog O, Evans DA, eds. Proc. of the ACM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2004. 406−415. [doi: 10.1145/1031171.1031252]

[20] Liu JG, Zhou T, Wang BH. Progress of the personalized recommendation systems. Progress of Nature and Science, 2009,19(1):1−15 (in Chinese with English abstract). [doi: 10.3321/j.issn:1002-008X.2009.01.001]

[21] Ma H, Yang HX, Lyu MR, King I. Sorec: Social recommendation using probabilistic matrix factorization. In: Shanahan JG, Amer-Yahia S, Manolescu I, Zhang Y, Evans DA, Kolcz A, Choi KS, Chowdhury A, eds. Proc. of the ACM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2008. 931−940. [doi: 10.1145/1458082.1458205]

[22] Hu X, Meng XW, Zhang YJ, Shi YC. Recommendation algorithm combing item features and trust relationship of mobile users. Ruan Jian Xue Bao/Journal of Software, 2014,24(8):1817−1830 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4491.htm [doi: 10.13328/j.cnki.jos.004491]

[23] Ma H. An experimental study on implicit social recommendation. In: Jones GJF, Sheridan P, Kelly D, Rijke MD, Sakai T, eds. Proc. of the Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2013. 73−82. [doi: 10.1145/2484028.2484059]

[24] Leskovec J, Kleinberg J, Faloutsos C. Graphs over time: Densification laws, shrinking diameters and possible explanations. In: Grossman R, Bayardo RJ, Bennett KP, eds. Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2005. 177−187. [doi: 10.1145/1081870.1081893]

[25] Buluc A, John R. G. Parallel sparse matrix-matrix multiplication and indexing: Implementation and experiments. SIAM Journal on Scientific Computing, 2012,34(4):C170−C191. [doi: 10.1137/110848244]

[26] Gustavson, Fred G. Two fast algorithms for sparse matrices: Multiplication and permuted transposition. ACM Trans. on Mathematical Software, 1978,4(3):250−269. [doi: 10.1145/355791.355796]

附中文参考文献:

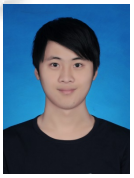
[20] 刘建国,周涛,汪秉宏.个性化推荐系统的研究进展.自然科学进展,2009,19(1):1−15. [doi: 10.3321/j.issn:1002-008X.2009.01.001]

[22] 胡勋,孟祥武,张玉洁,史艳翠.一种融合项目特征和移动用户信任关系的推荐算法.软件学报,2014,25(8):1817−1830. http://www.jos.org.cn/1000-9825/4491.htm [doi: 10.13328/j.cnki.jos.004491]

余骞(1985－),男,湖北孝感人,博士生,主要研究领域为 Web 社区管理,Web 社区推荐.

洪亮(1982－),男,博士,副教授,CCF 专业会员,主要研究领域为数据管理,社会网络.

彭智勇(1963－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为 Web 数据管理,复杂数据管理,可信数据管理.



万言历(1994－),男,学士,主要研究领域为数据库,数据挖掘.