软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(6):1565−1583 [doi: 10.13328/j.cnki.jos.005249]

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

面向社会网络融合的关联用户挖掘方法综述*

周小平[1,2], 梁 循[1], 赵吉超[1], 李志宇[1], 马跃峰[1]

[1](中国人民大学 信息学院,北京 100872)

[2](北京建筑大学 电气与信息工程学院,北京 100044)

通讯作者: 梁循, E-mail: xliang@ruc.edu.cn

摘 要: 现阶段大多数社会网络的研究都集中于单一的社会网络内部.社会网络融合为社会计算等各项研究提供更充分的用户行为数据和更完整的网络结构,从而更有利于人们通过社会网络理解和挖掘人类社会,具有重要的理论价值和实践意义.准确、全面、快速地关联用户挖掘,是大型社会网络融合的根本问题.社会网络中的关联用户挖掘旨在通过挖掘不同社会网络中同属于同一自然人的不同账号,从而实现社会网络的深度融合,近年来已引起人们的广泛关注.然而,社会网络的自身数据量大、用户属性相似、稀疏且存在虚假和不一致等特点,给关联用户挖掘带来了极大的挑战.分析了面向社会网络融合的关联用户挖掘所存在的困难,从用户属性、用户关系及其综合这 3 个方面梳理了当前关联用户挖掘的研究现状.最后,总结并展望了关联用户挖掘的研究方向.

关键词: 社会网络;社会网络融合;关联用户;用户属性;用户关系

中图法分类号: TP181

中文引用格式: 周小平,梁循,赵吉超,李志宇,马跃峰.面向社会网络融合的关联用户挖掘方法综述.软件学报,2017,28(6):1565−1583. http://www.jos.org.cn/1000-9825/5249.htm

英文引用格式: Zhou XP, Liang X, Zhao JC, Li ZY, Ma YF. Correlating user mining methods for social network integration: A survey. Ruan Jian Xue Bao/Journal of Software, 2017,28(6):1565−1583 (in Chinese). http://www.jos.org.cn/1000-9825/5249.htm

Correlating User Mining Methods for Social Network Integration: A Survey

ZHOU Xiao-Ping[1,2], LIANG Xun[1], ZHAO Ji-Chao[1], LI Zhi-Yu[1], MA Yue-Feng[1]

[1](School of Information, Renmin University of China, Beijing 100872, China)

[2](School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China)

Abstract: Social network (SN) has become a popular research field in both academia and industry. However, most of the current studies in this field mainly focuses on a single SN. Obviously, the integration of SNs, termed as social network integration (SNI), provides more sufficient user behavior data and more complete network structure for the studies on SN such as social computing. Additionally, SNI is more effective in excavating and understanding human society through SNs. Thus, it has significant theoretical and practical value to explore problems in SNI. Correlating users refer to the user accounts belonging to the same individual in different SNs. Since users naturally bridge the SNs, correlating user mining problem is the fundamental task of SNI, hence having attracted extensive attention. Due to the unfavorable characteristics of SN, correlating user mining problem is still a hard nut to crack. In this paper, the difficulties in the correlating user mining task are analyzed, and the methods addressing this issue are summarized. Finally, some potential future research work is suggested.

Key words: social network; social network integration; correlating user; user property; user relationship

* 基金项目: 国家自然科学基金(71271211, 71531012, 71601013); 北京市自然科学基金(4132067, 4174087); 北京市教委科技计划项目(SQKM201710016002)

Foundation item: National Natural Science Foundation of China (71271211, 71531012, 71601013); Beijing Natural Science Foundation (4132067, 4174087); Scientific Research Project of Beijing Educational Committee (SQKM201710016002)

收稿时间: 2016-09-28; 修改时间: 2016-12-07; 采用时间: 2017-01-04; jos 在线出版时间: 2017-01-20

CNKI 网络优先出版: 2017-01-20 16:06:38, http://www.cnki.net/kcms/detail/11.2560.TP.20170120.1606.013.html

社会网络(social network)是指人们用于创建、分享、交流信息和观点的虚拟社区[1].近年来,随着 Facebook, Twitter 的影响力不断提高,新浪微博、微信在人们日常生活中的深入渗透,社会网络已使得当前社会经济文化问题日益呈现出了动态性、快速性、开放性、交互性和数据海量化等特点[2].得益于社会网络所产生的海量用户行为数据,研究人员使用社会网络进行社区发现[3]、影响力分析[4]、链接分析[5]、情感分析[6]、观点挖掘[6]、商务智能[7]、企业决策支持[8]等应用[2,9,10].然而,大多数的社会网络研究都仅局限于单一的社会网络内部.高效、健壮地融合多个社会网络,为社会网络各项研究提供更为完善的用户行为数据(如图 1 所示),将使社会网络的研究更全面、更准确,也更有利于人们认识社会网络,进而通过社会网络认识人类社会.

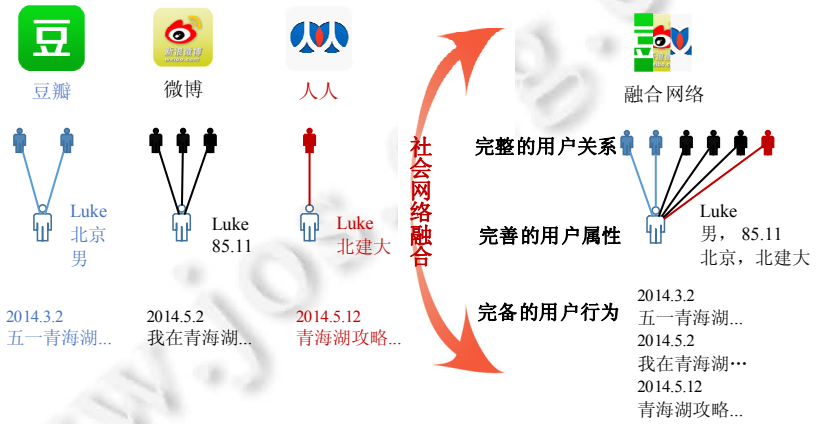


Fig.1 Social network integration

图 1 社会网络融合示意图

此外,集成不同社会网络上的用户信息,将为个性化服务及跨领域推荐[11]提供更全面的用户数据,也是解决“冷启动”问题[12]的重要途径之一.社会网络融合(social network integration)已成为现阶段社会网络研究的一个重要和热点问题.

用户是社会网络的主体.由于不同的使用需求,人们在不同的社会网络上注册用户,并在社会网络中构建用户关系,创造用户内容(user generated content,简称 UGC).因此,社会网络通常可表示为

$$SN=\{U,F,C\}.$$

其中,U 为社会网络中的用户集合,F 为社会网络的所有用户关系集合,C 为 UGC 集合.

不失一般性,本文使用 SN_A 表示社会网络 A,U_A,F_A 和 C_A 分别表示 SN_A 中的用户集合、用户关系集合和 UGC 集合.更进一步地,U_{Ai} 和 C_{Ai} 分别为表示 SN_A 中的用户 i 及其所创造的 UGC,$F_{Ai\sim j}$ 表示 SN_A 中 U_{Ai} 和 U_{Aj} 之间的好友关系.此外,我们有如下定义.

定义 1(社会网络融合). 通过匹配不同社会网络中的相同节点,将多个社会网络融合形成一个规模更大、信息更完备的社会网络,叫社会网络融合.

定义 2(关联用户). 假定 U_{Ai} 和 U_{Bj} 分别是大型社会网络 SN_A 和 SN_B 中的用户.若 U_{Ai} 和 U_{Bj} 都是现实世界中同一自然人 I 在 SN_A 和 SN_B 中的帐户(用户),则 U_{Ai} 和 U_{Bj} 是关联用户(correlating user,user linkage),记为 $U_{Ai}=U_{Bj}$ 或 $UL_{A\sim B}(i,j)$.关联用户也称关连用户.

定义 3(关联用户挖掘). 假定待融合的社会网络为 SN_A 和 SN_B,关联用户挖掘是指根据已知信息 $\mathcal{P}$,获取 SN_A 和 SN_B 中所有关联用户的方法.不同的融合环境,将可能使得已知信息 $\mathcal{P}$ 不尽相同.例如:基于用户属性的关联用户挖掘方法,$\mathcal{P}$ 主要为用户属性信息;而基于用户关系的关联用户挖掘方法,$\mathcal{P}$ 主要为社会网络的网络结构信息.通常,关联用户挖掘将转化为判定两个来自 SN_A 和 SN_B 的用户 U_{Ai} 和 U_{Bj} 在已知信息 $\mathcal{P}$ 下是否同属于一个自然人 I,即:

$$f(U_{Ai},U_{Bj}\mid\mathcal{P})=\begin{cases}1, & U_{Ai}=U_{Bj}\\ 0, & \text{其他}\end{cases} \tag{1}$$

社会网络间的关联用户挖掘旨在发现准确、全面的关联用户,以实现社会网络的深度融合(如图 2 所示).显然,用户是社会网络融合的天然桥梁,关联用户挖掘将直接从社会网络节点上融合社会网络[13].因此,如何构建准确、全面、快速的关联用户挖掘模型和方法,是社会网络融合的核心问题.

Fig.2 Correlating user mining

图 2 关联用户挖掘示意图

社会网络的研究涉及社会科学、管理科学和信息科学等众多学科,其研究和应用涉及人类生活的各个方面.随着人们对社会网络研究的深入,人们越来越渴望能够获取更全面的社会网络用户行为数据,以挖掘更有价值的用户行为规律和开展商务智能研究.社会网络融合和关联用户挖掘已引起了学术界和企业界的广泛关注,成为了社会网络研究的重要趋势.近两三年来,在国际数据库和数据挖掘的权威期刊和会议(如 IEEE Trans. on Knowledge and Data Engineering[13,14],ACM Conf. on Management of Data[15],ACM Knowledge Discovery and Data Mining[16,17],Int'l Conf. on Very Large Data Bases[18],ACM Int'l Conf. on Information and Knowledge Management[19,20],Int'l World Wide Web Conf.[21]等)中都刊载了关联用户挖掘的相关研究.本文将介绍社会网络关联用户挖掘所面临的问题,总结关联用户挖掘方法和研究进展.

1 关联用户挖掘面临的问题

早期,研究人员通过 E-mail 构建 Find Friend 机制构建关联用户挖掘方法[22].绝大多数的社会网络都通过 E-mail 注册账号(近年来兴起的移动社会网络中,有部分使用手机号注册账号).由于 E-mail 的唯一性,Find Friend 使用社会网络所提供的"E-mail 查找用户"功能挖掘不同社会网络间的关联用户.近年来,随着人们对自身网络隐私的重视以及社会网络平台对用户数据的保护,E-mail 等用户隐私数据已被各社会网络平台屏蔽,第三方应用可获取的用户属性信息越来越少.据统计,用户平均在一个社会网络中公开 4 项属性信息[23].这给关联用户挖掘及社会网络融合带来了极大的挑战.

目前,社会网络关联用户挖掘所面临的挑战包括:

1) 相似性.随着用户数量的增加,社会网络出现了大量的具有相似或相同属性信息但不关联的用户.如图 2 所示,新浪微博和人人网都有上千用户名包含 luke 的用户;

2) 稀疏性.因许多用户未填写某项(些)属性而导致该项(些)属性信息较为稀疏.例如,头像是社会网络中

的一项重要属性,而只有 66%的用户会上传头像[24];

3) 虚假性.社会网络用户属性的虚假性主要源于:(I) 用户因不愿公开某项(些)属性而填写虚假的属性值;(II) 恶意用户因其需要设定用户属性与某(些)其他用户相同;(III) 用户填写属性信息时的随意性也容易造成虚假信息;

4) 不一致性.同一用户在不同的社会网络中对同一属性填写不同的值.例如,单位属性里,可能在社会网络 SN_A 中填“中国人民大学”,而在社会网络 SN_B 中填“人民大学”,或“人大”,或“RUC”.

5) 大数据.区别于传统的复杂网络,绝大多数的社会网络都拥有百万级以上的用户.其大数据也给现有的研究带来了巨大的挑战:

   - 一方面,用户属性是挖掘关联用户的最直接方法.现阶段,大多数的关联用户发现方法都基于用户属性(如昵称、头像)相似度的计算.然而社会网络中用户属性的相似性、稀疏性、虚假性和不一致性,使得单纯使用用户属性挖掘关联用户的方法易受恶意用户的攻击,健壮性较差;
   - 另一方面,用户关系,尤其是好友关系,是社会网络中较稳定、不易受攻击且可获取的信息.目前,基于用户关系挖掘关联用户的研究大都针对匿名化的社会网络在线发布数据的还原,又称去匿名化(de-anonymization)[25].使用用户关系的相似性构建社会网络间的关联用户挖掘模型,易于挖掘网络结构相似的关联用户.社会网络存在大量网络结构高度相似的用户节点[26],单独使用网络结构较难准确、全面地挖掘关联用户.

无疑,合理融合用户属性和用户关系,是构建准确、全面的关联用户挖掘模型的更优方法.

- 一方面,在基于用户属性的关联用户挖掘方法中融入用户关系,采用用户关系甄别恶意用户,将提升原有算法的健壮性;
- 另一方面,在基于用户关系的关联用户挖掘方法中融入用户属性,采用用户属性对所挖掘的关联用户进行二次校验,将提升原有算法的准确率.

目前,部分学者[14,15]已经着手研究利用已有数据,融合用户属性和用户关系构建关联用户挖掘模型.

此外,从图的角度上看,社会网络融合可以归结为图的同构问题[27].理论上,图同构问题的求解所需时间为

$$\sum_{i=1}^{\min(n_1,n_2)} \frac{n_1!n_2!}{n!(n_1-i)!n!(n_2-i)!} \tag{2}$$

其中,n_1 和 n_2 分别为两个图的节点数,$n!$为 n 的阶乘.显然,图的同构问题是一个 NP 问题[28].面对百万级以上的网络节点,图的同构求解方法根本无法应用于社会网络的关联用户挖掘.

随着社会网络在人们日常生活中的不断渗透,其用户数量越来越大.如何提升关联用户挖掘方法的效率,减少其运行时间,满足大数据处理的需求,是关联用户挖掘模型和方法可实践性的关键.

2 关联用户挖掘的方法

社会网络由用户 U、用户关系 F 和 UGC C 组成.相应地,关联用户挖掘可以从用户属性、用户关系和 UGC 着手.通常,用户属性指用户个人资料中所公开的信息,包括用户名、头像、年龄、教育背景、工作信息等.人们注册不同的社会网络以满足不同的使用需求,这也使得用户在不同的社会网络中发布不同的 UGC.因此,UGC 通常并不适用于关联用户挖掘.然而,UGC 所包含的用户行为信息,包括 UGC 发布地点、发布时间以及书写风格等,却是发现关联用户的有效方法之一.本文将用户行为信息也视为用户属性的一种.为此,关联用户挖掘主要从用户属性和用户关系着手进行.其中,大多数的关联用户挖掘研究仅考虑某一或几项用户属性,并较准确地挖掘部分关联用户;而用户关系多用于去匿名化研究.因此,从社会网络的组成要素出发,现有的关联用户挖掘方法可以分为 3 类:基于用户属性、基于用户关系和综合属性和关系(如图 3 所示).

多个社会网络间的关联用户挖掘通常转化为两两社会网络间的关联用户挖掘.因此,当前社会网络关联用户挖掘主要在两个社会网络中展开.本文所调研的文献也都在两个社会网络间开展关联用户挖掘研究.

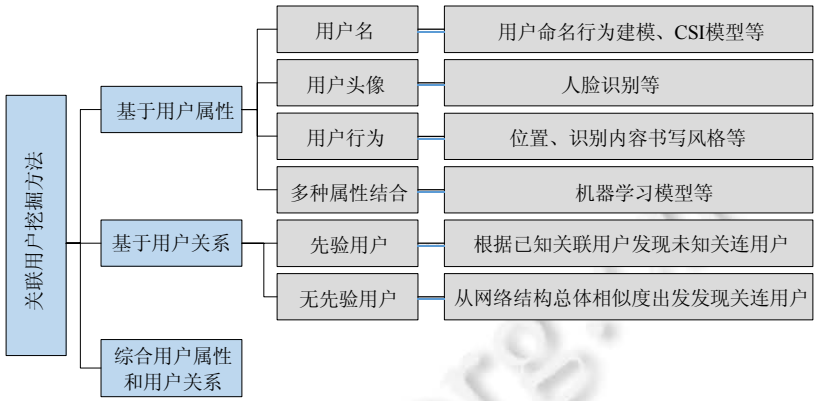


Fig.3 Correlating user mining methods

图 3 关联用户挖掘方法示意图

2.1 基于用户属性的关联用户挖掘

基于用户属性的关联用户挖掘是目前研究最广的社会网络融合方法,此类方法认为:如果两个用户属性相同或相似,则这两个用户为关联用户.当前,研究人员从用户名、用户头像、其他用户属性、UGC 等各方面开展了广泛的研究,表 1 对该类方法的已知信息 $\mathcal{P}$ 进行了统计分析.

Table 1 $\mathcal{P}$ in the methods using users' profiles and UGC

表 1 基于用户属性的关联用户挖掘方法中的 $\mathcal{P}$

类别	方法	$\mathcal{P}$
基于用户名的关联用户挖掘方法	Ref.[31]	海量用户名
	Ref.[16]	用户名,部分已知关联用户
	Ref.[19]	用户名,用户购买记录
基于头像的关联用户挖掘方法	Ref.[32]	用户头像
综合用户属性的关联用户挖掘方法	Ref.[33]	用户标签
	Ref.[34–36]	所有用户文本属性(用户名、教育背景、年龄等)
	Ref.[38]	5 个内部特征(用户名、使用语言、URL、描述、好友数)和 2 个外部特征(位置、头像)
基于 UGC 的关联用户挖掘方法	Ref.[20,21]	UGC 的空间位置、时间戳和文本
	Ref.[39,40]	UGC 文本
	Ref.[41]	UGC 的空间位置和时间戳

2.1.1 基于用户名的关联用户挖掘方法

用户名是目前关联用户挖掘建模使用最多的方法.Perito 等人通过实证分析了 Google,eBay,LDAP 和 MySpace 的用户名数据,探索了用户名在不同社会网络的唯一性,并验证了其应用于关联用户挖掘的可行性[29].Zafarani 等人[30]对 12 个社交网络中的上千个用户名进行了相似的实证验证.在此基础上,研究人员采用无监督[31]和有监督分类[16]两种方法进行关联用户识别.

Liu 等人[31]认为,不同社会网络中的稀有且相似的用户名极有可能是关联用户.例如 SN_A 和 SN_B 中,用户名同是 pennystar88 的用户极有可能是关联用户,而同是 tank 的用户极有可能属于不同自然人.为此,他们根据用户名采用别名区分(alias-disambiguation)构建了基于无监督分类的关联用户发现方法.他们首先对用户名进行分词,例如将 pennystar88 分解为 penny,star,88 或者 pen,ny,star,88;而后,采用 *n*-gram 概率[59]计算所分解的词组的稀有性.若用户名 *username* 可分解为词语 $w_1,w_2,\ldots,w_m$,则 *username* 出现的概率为

$$p(username)=p(w_1,w_2,...,w_m)=\prod_{i=1}^{m}p(w_i\mid w_{i-n-1}...w_{i-1}) \tag{3}$$

显然,$p(username)$越小,*username* 越稀有,越有可能用于关联用户挖掘.给定海量用户名,则有:

$$p(w_i \mid w_{i-n-1}...w_{i-1}) = \frac{C(w_{i-n-1},...,w_i)}{C(w_{i-n-1},...,w_{i-1})} \tag{4}$$

其中,$C(w_{i-n-1},\ldots,w_i)$为所有用户名分解后,词 $w_{i-n-1},\ldots,w_i$ 出现的频率.最后,根据稀有用户名的相似度来获取不同社会网络的关联用户.显然,该方法无需先验关联用户即可完成关联用户挖掘.由于 $p(username)$的衡量是该方法准确性的关键,为保证 $p(username)$的客观性和有效性,该方法需要较大数据集作为数据支撑.

Zafarani 等人[16]认为,用户名背后隐藏着用户名命名的行为特征.为此,他们从人类群体局限、个体外在因素和个体内在因素这 3 个方面建立用户名命名行为特征模型(如图 4 所示):人类群体局限是指受时间、记忆和知识的影响,同一自然人在用户名命名时,其长度相近,内容相近且都限制于已知词汇量和字母表的范围;个体外在因素是指用户名命名受键盘和语言习惯等的影响,例如某些用户的用户名可能是有序的 qwer1234;个体内在因素是指用户名命名受用户属性和用户习惯等的影响,例如用户的出生日期等.在此基础上,他们采用 SVM 等方法对已知关联用户进行特征学习,进而识别未知关联用户.区别于 Liu 等人的方法[31],该方法需要已知先验关联用户作为支撑.

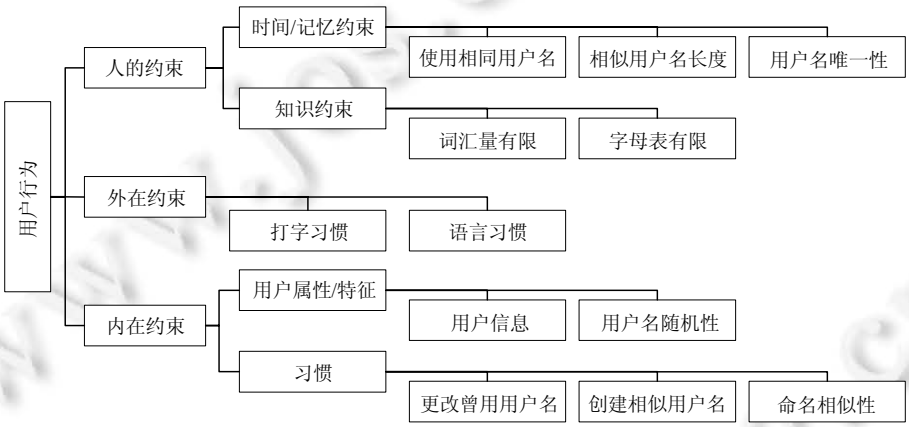


Fig.4 User behavior patterns on naming in social network[16]

图 4 用户命名行为模式[16]

其他相似的研究还包括:Lu 等人通过用户名和购买记录构建了 CSI(custom-social identification)模型,帮助企业获取客户在社会网络上的信息[19].虽然用户名内潜藏着有机可循的关联用户挖掘方法,然而在社会网络中,大量用户名相似的用户(如图 2 所示)使得此类方法健壮性较差.

2.1.2 基于用户头像的关联用户挖掘方法

用户头像也是社会网络中较为重要的用户属性之一.虽然只有 66%左右的用户会上传用户头像[24],然而用户的真实头像是区分关联用户和非关联用户的重要依据之一.Acquisti 等人采用人脸识别算法计算用户头像的相似度来实现社会网络融合[32].然而在大型社会网络中,存在着大量使用非自身照片的用户.如图 2 所示,7 名新浪微博用户中,有 5 名为非自身人脸头像(含 1 名小孩头像).因此,该类方法的召回率较低.

2.1.3 综合用户属性的关联用户挖掘方法

为构建更为健壮的关联用户挖掘模型,研究人员综合使用多项用户文本属性,以挖掘更准确、更全面的结果.Iofciu 等人根据用户属性中的标签建立标签向量,而后,通过计算标签向量之间的相似度来挖掘关联用户[33].Motoyama 等人对用户文本属性(教育背景、职业等)进行分词,而后采用词袋模型(bags of words)计算用户相似度,完成关联用户挖掘[34].即,用户的相似度为

$$s(U_{Ai},U_{Bj}) = \frac{|W_{Ai} \cap W_{Bj}|}{|W_{Ai} \cup W_{Bj}|} \tag{5}$$

其中,$|\cdot|$表示集合$\cdot$的数量,W_{Ai} 为用户 U_{Ai} 属性进行分词后词袋中的词集合.叶娜等人针对识别社交网络用户时存

在的模式不一致问题,将所有文本属性融合为一个字符串计算相似度,进而基于分块和二部图进行用户识别[35].此类方法都是在假定“所有用户属性对关联用户挖掘具有一样的效用”的基础上挖掘关联用户.用户属性之间是异质的.虽然该类方法解决了异质属性的融合问题,然而不同用户属性对关联用户挖掘具有不一致的效用,例如虚假性强的属性,其将噪音引入关联用户挖掘模型,并可能对挖掘效果起反作用.

为解决不同用户属性对关联用户挖掘效用不一致问题,研究人员引入了机器学习.Zhang 等人认为,单纯使用用户名进行关联用户挖掘建模不可靠,提出了针对用户属性的关联框架 OPL(online profile linkage).OPL 分别计算了用户的 5 个内部特征(用户名、使用语言、URL、描述、好友数)和 2 个外部特征(位置、头像)的相似度,而后采用朴素贝叶斯进行用户区分[36].Cortis 等人采用 NCO(nepomuk contact ontology)[37]构建用户属性本体,建立关联用户挖掘的方法[38].对于任意用户 U_{Ai},其属性可以描述为

$$P_{Ai}=\{P_{Ai}^1,P_{Ai}^2,...,P_{Ai}^m\} \tag{6}$$

其中,P_{Ai}^i 为用户 U_{Ai} 的第 j 个属性值,m 为属性总数.而后有:

$$s(U_{Ai},U_{Bj})=\frac{\sum_{n=1}^{m}\left(\frac{sim(P_{Ai}^n,P_{Bj}^n)+type_w(P^n)}{1+type_w(P^n)}\right)}{m} \tag{7}$$

其中,$sim(P_{Ai}^n,P_{Bj}^n)$ 和 $type_w(P^n)$分别为第 n 个属性的相似度和权重.

虽然这些方法都在实验数据集下都取得了更好的结果,然而在大型社会网络中,用户属性的相似性、虚假性使得该类方法较为脆弱,恶意用户极易伪造虚假用户,从而影响模型的健壮性.

2.1.4 基于 UGC 的关联用户挖掘方法

为解决基于用户资料属性挖掘关联用户易受攻击的问题,少数研究引入了用户行为属性.Kong 等人综合了 UGC 的空间位置、时间戳和文本的相似度,采用 SVM 构建了多网络锚点(multi-network anchoring,简称 MNA)的关联用户挖掘方法[20].在空间位置上,MNA 采用共同位置、cos 相似度和平均距离等 3 个方面计算两个用户的相似度;在时间上,MNA 采用相同发 UGC 时间、cos 相似度计算两个用户的相似度;在文本内容上,MNA 对内容进行分词,建立词袋模型,而后采用向量内积和 cos 相似度计算用户的文本相似度.最后,MNA 采用 SVM 等分类算法进行关联用户挖掘.Zheng 等人提出了一种基于对内容书写风格识别的关联用户挖掘方法[39].Almishari 也验证了采用书写风格挖掘关联用户的可行性[40].由于书写风格识别技术在短文本中的适用性还较差,Goga 等人综合了 UGC 空间位置、发布时间和用户书写风格,使用分类器挖掘关联用户[21].Nie 等人在 240 人的数据集上验证了使用用户习惯挖掘关联用户的可行性[41].毋庸置疑,用户行为是提升关联用户挖掘模型的有力方法,例如:空间位置虽然能较为精确地挖掘关联用户,然而社会网络中空间位置信息极为稀疏,且大多数用户不愿意公开其空间位置信息.这些都使得现阶段该类方法在大型社会网络中的召回率较低.

2.2 基于用户关系的关联用户挖掘

用户关系是社会网络中稠密、可靠且可获取的用户属性.目前,使用用户关系融合社会网络的研究还较少.大多数研究旨在解决社会网络隐私保护中的去匿名化问题.去匿名化主要针对节点去匿名化,它通过识别节点以获取该节点的真实数据信息.去匿名化通过识别匿名信息网络和真实信息网络中的相同节点,获取节点真实数据信息.因此,一定程度上说,去匿名化问题和关联用户挖掘问题具有一定的相似性.然而,二者也有本质的不同:在去匿名化问题研究中,通常所涉及的两个网络在某个子网上具有高度的重叠性;而在关联用户挖掘研究中,所涉及的两个网络其重叠度大致在 60%[13,35].因此,去匿名化的方法通常并不适用于关联用户挖掘.

当前,根据所形成的算法是否需要事先给定一定量已知关联节点,可将该类研究方法分为基于先验节点的关联用户挖掘方法和无先验节点的关联用户挖掘方法两类.表 2 分析了当前基于用户关系的关联用户挖掘算法的时间复杂度.

Table 2 Time complexities of correlating user mining algorithms using user relationship

表 2 基于用户关系的关联用户挖掘算法时间复杂度分析

类别	方法	算法复杂度
基于先验节点的关联用户挖掘方法	Ref.[46,47]	$O(en^2)$
	Ref.[13,18,44]	$O(n^3)$
无先验节点的关联用户挖掘方法	Ref.[49]	$O(n^4)$
	Ref.[50,51]	$O(n^5)$

注:e,n 分别表示两个待融合社会网络中的最大用户关系总数和最大用户总数.

2.2.1 基于先验节点的关联用户挖掘方法

根据已知关联用户(也称种子用户、种子节点、先验用户等)定义未关联用户间的相似度,相似度较高的用户对视为关联用户,并通过迭代方法关联越来越多的用户(如图 5 所示).

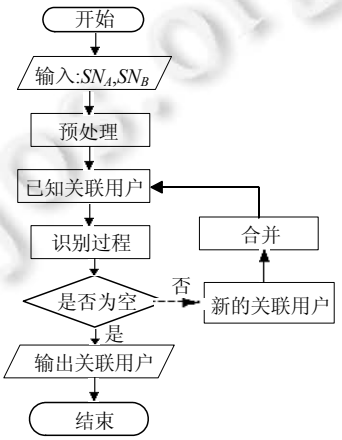


Fig.5 Correlating user mining based on user relationship[13]

图 5 基于用户关系的关联用户挖掘[13]

国内外学者已将 Tonimoto 系数[42]、Jaccard 系数[43]、好友共献[13,18,44]等用于去匿名化建模.徐钦根据网络结构以及先验关联用户信息计算节点相似度矩阵,再由遗传算法求得网络间相似度之和最大的节点匹配方案[45].Narayanan 等人针对去匿名化问题建立了 NS 算法[46],NS 算法是去匿名化问题中较具有影响力的算法,并赢得了 IJCNN 2011 社会网络挑战赛的冠军[47].NS 算法综合考虑了节点的出、入度构建相似度计算公式:

$$s(U_{Ai},U_{Bj})=\frac{c_{in}(U_{Ai},U_{Bj})}{\sqrt{d_{in\text{-}Bj}}}+\frac{c_{out}(U_{Ai},U_{Bj})}{\sqrt{d_{out\text{-}Bj}}} \tag{8}$$

其中,$c_{in}(U_{Ai},U_{Bj})$是 U_{Ai} 和 U_{Bj} 同时连入已知关联用户的数量,$c_{out}(U_{Ai},U_{Bj})$是同时连入 U_{Ai} 和 U_{Bj} 的已知关联用户的数量,$d_{in\text{-}Bj}$ 和 $d_{out\text{-}Bj}$ 分别是 U_{Bj} 的出、入度.在每一次迭代中,NS 只寻找满足如下条件的 U_{Bj}:

$$\frac{\max(U_{Ai})-\max_2(U_{Ai})}{\sigma(U_{Ai})}\geqslant\rho \tag{9}$$

其中,$\max(U_{Ai})$和 $\max_2(U_{Ai})$分别为$\{s(U_{Ai},\cdot)\}$的最大值和次大值,ρ为给定的阈值参数.由于在该公式中 NS 只考虑了 U_{Bj} 的出度、入度,为提高算法的准确率,NS 采用反向传播相似度验证,以修正因种子用户不足造成的错误关联.经真实社会网络数据实验验证,NS 算法大约能挖掘 30%左右的关联用户[46].为减少种子用户对算法效果的影响,Nilizadeh 等人在 Narayanan 等人的基础上对种子节点进行社区划分,并得到了更为准确的结果[48].

虽然 NS 算法[46]及其衍生算法[48]在去匿名化领域取得了较好的效果,并适用于重叠度较大情况下的社会网络关联用户挖掘,然而在社会网络关联用户挖掘任务中,其还存在一定的不足,主要体现在:① 微博类社会网络关注行为的无限制性使得关注关系稳定性较弱,因此,NS 算法采用有向图构建社会网络模型,分别考量节点的出度、入度,这有可能为关联用户挖掘带来较大的噪声;② 区别于去匿名化,社会网络关联用户挖掘的对象往

往是稀疏重叠的网络,因此,NS 算法中对度的引入,有可能降低关联用户挖掘的召回率和准确率等.实验部分也验证了:NS 算法在关联用户挖掘任务中的效果差于针对社会网络关联用户挖掘的算法——FRUI[13].

通常,用户在不同的社会网络维护相似的好友圈.Xuan 和 Wu 等人通过 QQ 和手机通信调查发现,60%的电话通信对象是 QQ 好友[42].调查发现,用户与 67.5%的新浪微博好友建立人人网的好友关系[13].在此基础上提出了基于好友关系的社会网络关联用户挖掘方法 FRUI[13].FRUI 认为,社会网络是真实社会的抽样.假定真实世界是一个随机网络,且任意两个人是好友的概率为 $p(0<p<1)$.若社会网络 SN_A 和 SN_B 分别以概率 s_a 和 $s_b(0<s_a,s_b<1)$ 抽样真实世界的好友关系,则 SN_A 和 SN_B 中任意两个用户是好友的概率为 ps_a 和 ps_b.因此,给定已知信息 $\mathcal{P}$ 为已关联用户 F,则分别来自 SN_A 和 SN_B 的待关联用户 U_{Ai} 和 U_{Bj} 的已知好友为

$$c(U_{Ai},U_{Bj}\mid \mathcal{P}=F)=\begin{cases}|F|ps_as_b, & U_{Ai}=U_{Bj}\\ |F|p^2s_as_b, & U_{Ai}\neq U_{Bj}\end{cases} \tag{10}$$

其中,$|F|$为已关联用户 F 的大小.显然,关联用户和非关联用户的已知好友之间有 $1/p$ 倍的差距.不难发现:s_a 和 s_b 描述了 SN_A 和 SN_B 的重叠度,p 表征真实社会网络的密度.当 p 一定时,较大的 s_a 和 s_b 表示 SN_A 和 SN_B 的重叠度较大,此时,较小的$|F|$即可得出较好的关联用户挖掘效果;当 s_a 和 s_b 一定时,较小的 p 将能扩大 $U_{Ai}=U_{Bj}$ 和 $U_{Ai}\neq U_{Bj}$ 下已知好友之间的差距倍数.然而,为确保 $c(U_{Ai},U_{Bj}|F,U_{Ai}=U_{Bj})-c(U_{Ai},U_{Bj}|F,U_{Ai}\neq U_{Bj})$足够大,将需要较多的先验用户 F.反过来,随着 p 的增大,真实社会网络越来越趋近于簇(cluster),从而使得网络中任何两个节点的网络相似度增大.此时,需要更多的先验用户以确保能够正确区分关联用户和非关联用户.

在此基础上,FRUI 给出了好友相似度计算方法:

$$s(U_{Ai},U_{Bj})=|F_{Ai}\cap F_{Bj}|+\frac{|F_{Ai}\cap F_{Bj}|}{\min(|F_{Ai}|,|F_{Bj}|)} \tag{11}$$

鉴于 FRUI 也只考虑一对一匹配,在每一轮迭代中,FRUI 给出了如下关联用户判别公式:

$$f(U_{Ai},U_{Bj})=(s(U_{Ai},U_{Bj})==\max_u(s)) \tag{12}$$

其中,$\max_u(s)$为所有未关联用户对中无冲突的一对用户对的相似度值.FRUI 理论上较为简单,但大量实验上验证了,其效果要优于 NS 算法.

FRUI 和 NS 等都是一对一匹配,即,假定一个自然人在一个社会网络中只存在一个账户.对于一个用户的多个账户,FRUI 和 NS 等算法视之为不同用户.在多对多的匹配中,往往认为相似度大于某个设定阈值的用户对即为关联用户[18].

基于先验节点的关联用户挖掘算法根据已知关联用户建立未关联用户之间的相似度模型,该类方法的挖掘质量和数量一定程度上依赖于种子用户的数量和质量.现阶段,由于社会网络平台和平台用户对隐私保护的日益重视,社会网络中的种子用户的获取将变得越来越困难.

2.2.2 无先验节点的关联用户挖掘方法

在无先验的情况下,关联用户挖掘旨在找出一种最优节点匹配方法,使得两个网络中的边的重叠度最高.假定两个网络的邻接矩阵分别为 A 和 B,则该类方法旨在求解下述问题:

$$\arg\min_P \| A-PBP^T \| \tag{13}$$

其中,P 为排列矩阵,P^T 是 P 的转置矩阵.虽然图同构的一些研究成果可用于该问题的求解,然而图同构是 NP 问题,且其往往适用于两个基本相同的网络.社会网络关联用户挖掘所研究的对象通常是两个有一定重叠度的网络,例如 QQ 和通讯录的重叠度在 60%左右、新浪微博和人人网的重叠度在 67.5%左右,因此,图同构的方法在社会网络的关联效果较差.

当前,无先验节点的关联用户挖掘方法大多将问题转化为优化问题进行求解,即:以总体相似度最大为目标建立关联用户挖掘模型及其求解方法,挖掘两个图模型中所有节点最可能的关联情况.

在单一社会网络中,通常认为一个节点重要是因为其邻居节点重要引起的.在此基础上,人们使用 $x=A\square x$,通过不断迭代来挖掘重要节点(PageRank 算法).相似地,在基于网络的关联用户挖掘中,通常认为其邻居都为关联用户的一对用户也为关联用户.因此,对于任意一对用户(U_{Ai},U_{Bj}),其相似度可以看做是其邻居相似度的综合.为

此,Signh 等人提出了 Global Network Alignment(GNA)算法[49].若 S 为两个社交网络的用户相似度矩阵,$S_{ij}=s(U_{Ai},U_{Bj})$表示用户 U_{Ai} 和 U_{Bj} 的相似度,则有:

$$S_{ij}=\sum_{a\in N_{Ai},b\in N_{Bj}}\frac{1}{|N_a||N_b|}s(a,b) \tag{14}$$

其中,N_{Ai} 为 U_{Ai} 的邻居节点.通过不断迭代,直至相似度矩阵 S 收敛.此时,S 中值大于设定阈值的项所对应的用户对即为所待挖掘的关联用户.针对用户(U_{Ai},U_{Bj}),其邻居 N_{Ai} 和 N_{Bj} 可形成一个完全二分图.由于关联用户未知,该算法根据完全匹配实现相似度的迭代.该方法在两个基本一致的图中效果较好,并被应用于蛋白质交互网络的检索[49]等.

针对在线加密社会网络数据的还原,Fu 等人也建立了类似的关联节点识别方法 Neighbor Matching(NM). NM 算法认为,同一用户在不同社会网络具有相似的好友关系挖掘关联用户,并通过迭代的方法进行计算[50].也即:两个图中的节点相似度是由邻居节点相似度决定的,而邻居节点的相似度是由一对一最优匹配衡量.NM 算法的基本流程如下:

1) 初始化两个图中节点之间的相似度,通常可都为 1;

2) 假设 m 和 n 分别是两个图中的两个节点,则第 k+1 轮迭代中节点的相似度 $s(m,n)$可以递归为

$$s^{k+1}(m,n)=\sum_{l\in N(m)}s^k(l,\theta(l)) \tag{15}$$

其中,$N(m)$表示节点 m 的邻居节点,θ是节点 m 和 n 的邻居之间的一一映射;

3) 根据二分图最优匹配法找出两个图中一一对应的节点,挖掘关联用户.

NM 算法和 GNA 算法方法具有一定的相似性,两者都旨在通过邻居节点迭代计算两个不同网络节点的相似度.然而在相似度迭代上,NM 算法采用二分图的一对一最优匹配算法,而 GNA 算法采用二分图全匹配权值加和.显然,NM 算法具有较高的匹配挖掘效果,然而其具有更高的时间复杂度(见表 2).NM 算法是当前该类方法中效果最好的关联用户挖掘方法,并获得了“WSDM2013 去匿名化挑战赛”的冠军[51].

此外,Pedarsani 等人在无种子节点的情况下,采用贝叶斯方法进行关联用户挖掘,在两个较为相近的网络中取得了较好的结果[52].为挖掘更准确的关联用户,Zhang 等人[17]通过充分考虑局部相似性和全局相似性构建了异质社会网络关联用户挖掘模型 COSNET.受种子用户获取困难的影响,目前无先验节点的关联用户挖掘方法已经引起了较为广泛的关注.然而,现阶段该类方法主要应用于社会网络的去匿名化研究中.去匿名化研究多应用于两个网络结构相近甚至相同的社会网络,且其对节点度较高的节点识别准确率较高.而在社会网络中,用户及用户关系的重叠度都较低,也就是说,社会网络在网络结构上的差异较大,因此,去匿名化的理论和方法直接应用于关联用户挖掘的可行性较差.此外,社会网络是无尺度网络,其内存在着大量度较低的节点也较难通过用户关系进行挖掘.

2.3 综合用户属性和用户关系的关联用户挖掘

用户关系是社会网络中较为稳定的要素.在用户属性中融入用户关系,构建关联用户挖掘模型,可以避免模型受恶意用户的攻击,提升模型的准确率;在用户关系中融入用户属性,可以更准确地识别度数较低的用户,提升关联用户挖掘模型的准确率和召回率.因此,通过融合用户属性和用户关系,一方面将有利于构建不易受攻击的关联用户挖掘模型;另一方面,有利于提升模型的准确率和召回率.现阶段,少数研究尝试将用户关系同用户属性相结合.

Jain 等人在 Facebook 和 Twitter 间挖掘关联用户,他们首先通过种子用户判断种子用户的好友中是否有用户名一致的用户.该方法并未从本质上融合用户属性和用户关系,且实验效果也证实,其所构建的模型中用户关系对关联用户的挖掘结果作用不大[53,54].

Yu 通过在单一社会网络内部计算节点之间的相似度,将社会网络构建为加权图,从而将关联用户挖掘转化为加权图的匹配问题,最终将 1 000 个用户的 DBLP 数据集抽样为两个社会网络进行实验验证[55].用户因不同的需要使用不同的社会网络,同一用户在不同社会网络中的 UGC 存在着较大的差异性,因此,Yu 的方法具有一定

的局限性.

为挖掘大规模社会网络中的关联用户,Liu 等人[14,15]构建了统一的挖掘框架 HYDRA.HYDRA 通过动态信息匹配和行为分布分析构建混杂属性信息模型,通过网络结构相似性和一致性构建结构一致性模型;最终,将问题转换为多目标优化问题进行解决.在混杂属性信息建模上,HYDRA 分别考虑了用户属性、UGC 和用户行为等信息的相似度,形成 d 维的属性相似度矩阵 S_D.若给定已知关联用户集合 $F=\{(x(U_{Ai},U_{Bi}),y(U_{Ai},U_{Bi}))\}$,其中,$x(U_{Ai},U_{Bi})$为$(U_{Ai},U_{Bi})$的相似度向量,$y(U_{Ai},U_{Bi})\in\{-1,1\}$表征$(U_{Ai},U_{Bi})$是否为关联用户.在此基础上,HYDRA 在属性上形成决策模型 f:

$$f(x)=\boldsymbol{w}^Tx+\boldsymbol{b} \tag{16}$$

其中,$\boldsymbol{w}$ 和 $\boldsymbol{b}$ 为模型参数,可通过下述最优化模型获得:

$$\min_{\boldsymbol{w}} F_D(\boldsymbol{w})=\frac{\gamma_L}{2}\|\boldsymbol{w}\|^2+\sum\xi(U_{Ai},U_{Bi})\ \text{s.t.}\ y(U_{Ai},U_{Bi})(\boldsymbol{w}^Tx(U_{Ai},U_{Bi})+\boldsymbol{b})\geqslant 1-\xi(U_{Ai},U_{Bi}) \tag{17}$$

其中,ξ为误差参数.在用户关系上,HYDRA 认为,现实世界的好友之间在社会网络上会有频繁的交互和相似的用户兴趣.为此,从社会网络结构一致性上,HYDRA 建立模型决策模型 $y(U_{Ai},U_{Bi})=\boldsymbol{w}^{\mathrm{T}}x(U_{Ai},U_{Bi})$,该模型可通过下述最优化进行求解:

$$\min_{\boldsymbol{w}} F_S(\boldsymbol{w})=\boldsymbol{w}^TX^T(D-M)X\boldsymbol{w}\ \text{s.t.}\ \|\boldsymbol{w}\|^2\leqslant s, D(a,a)=\sum\nolimits_b M(a,b) \tag{18}$$

其中,s 为预定义正整数值,a 和 b 为待匹配用户对.若 $a=(U_{Ai},U_{Bi})$,$b=(U_{Am},U_{Bn})$,则有:

$$M(a,b)=\exp\left(\frac{-(\|x_{Ai}-x_{Bj}\|^2+\|x_{Am}-x_{Bn}\|^2)}{2\sigma_1^2}\right)\left(1-\frac{(d(U_{Ai},U_{Am})-d(U_{Bj},U_{Bn}))^2}{\sigma_2^2}\right) \tag{19}$$

其中,σ_1 和σ_2 分别为用户交互行为和用户结构的权重调节参数,$d(U_{Ai},U_{Am})$为用户 U_{Ai} 和 U_{Am} 之间的最短路径距离.由于 HYDRA 需要对公式(13)和公式(14)进行最优化求解,为此,转化为多目标优化进行求解.也即:

$$\min_{\boldsymbol{w}} F(\boldsymbol{w})=[F_D(\boldsymbol{w}),F_S(\boldsymbol{w})]\ \text{s.t.}\ y(U_{Ai},U_{Bi})(\boldsymbol{w}^Tx(U_{Ai},U_{Bi})+\boldsymbol{b}\geqslant 1-\xi(U_{Ai},U_{Bi}),\|\boldsymbol{w}\|^2\leqslant s \tag{20}$$

由于 HYDRA 充分利用了社会网络的所有可用资源,其取得了较好的实验效果.然而,HYDRA 是一个半监督学习方法,需要一定的先验用户,因此具有一定的局限性.

用户属性和用户关系是社会网络的不同要素,用户在属性上的相似性易于用相似度表达.用户关系是一种网络结构,现有的理论和方法较难给出一种适用于关联用户挖掘的网络结构相似度计算模型.也即,现有的理论和方法无法将用户属性和用户关系统一于不同维度上的相似度融合.因此,在关联用户挖掘中,用户属性和用户关系的融合存在着不一致性.综合用户属性和用户关系的关联用户挖掘建模研究还处于初步探索阶段.

3 实验分析

3.1 实验方法和数据分析

当前,社会网络隐私问题已引起了学术界和企业界的广泛关注.一方面,它催生了社会网络隐私保护这一研究领域;另一方面,它也使得当前关联用户挖掘几乎没有公开的数据集.为此,当前较难对各类方法进行统一有效地实验对比和分析.

通常,我们认为社会网络是真实世界的映射,是真实世界的一种抽样.鉴于较难对用户属性进行有效抽样,形成实验数据,本文仅对基于用户关系的关联用户挖掘方法进行比对分析.本文实验分析了所调研文献中效果较好的 3 种方法,包括:

(1) NS 算法[46].NS 算法设计的初衷在于解决社会网络去匿名化问题,是当前最好的基于先验知识的去匿名化方法,并赢得了 IJCNN2011 社会网络挑战赛的冠军[47];NS 算法的健壮性也使得其能在一定程度上满足关联用户挖掘的需求;

(2) NM 算法[50].类似于 NS 算法,NM 算法主要应用于社会网络去匿名化问题,然而其是当前该领域无需先验知识的最好算法,赢得了 WSDM2013 社会网络去匿名化挑战赛的冠军[51].为此,本文也将此进行

实验分析;

(3) FRUI 算法[13].FRUI 算法是专门针对社会网络关联用户挖掘而设计的,其需要先验知识进行驱动.

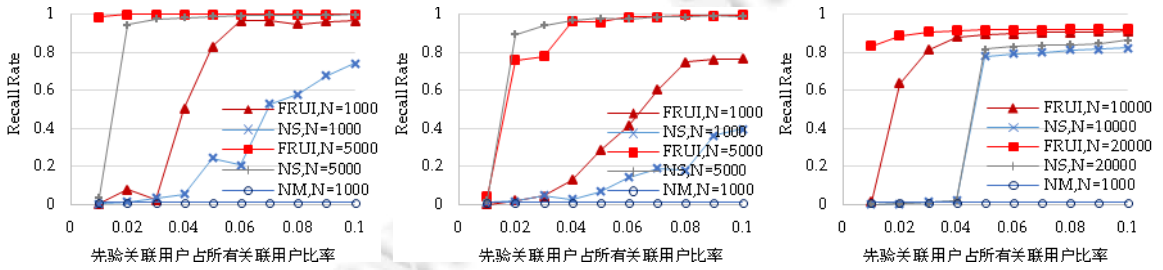
在数据集方面,虽然在社会网络隐私保护领域有部分数据可参考,例如 KDD12 和 WSDM13 去匿名化挑战赛数据集,然而去匿名化任务中的数据与本文所讨论的社会网络间关联用户挖掘有极大的不同:去匿名化任务中,其两个网络中存在某个高度重合的子网.为此,本文拟通过人工模拟网络建立实验数据集进行实验分析.

3.2 人工数据集实验分析

为进行充分的实验分析,本文首先针对 3 种典型复杂网络进行抽样实验.3 种人工复杂网络包括 Erdős-Rényi(ER)网络[56]、Watts-Strogatz(WS)网络[57]和 Barabási-Albert(BA)网络[58].ER 随机网络是早期研究比较多的一类复杂网络,其基本思想是:网络中的 N 个节点,两两之间以概率 p 存在边.WS 小世界网络核心思想是:对于一个有 N 个节点、每个节点有 k 个邻居,以概率 p_r 随机化重连其每条边.显然,当 p_r=1 时,WS 小世界网络转变为 ER 网络.经大量实验,pr 取[0,1]间的任意值时,本文所分析实验方法在召回率、准确率和 F1-Measure 等方面的对比结果基本一致.不失一般性,在本文实验中,p_r=0.5.BA 网络是近年来得到广泛研究的一类网络,其思想是:对于任意加入网络中的节点,其以正比于已有节点度的概率与已有节点建立 m 个连接.在度分布上,ER 网络和 WS 网络都服从钟型分布,其区别在于:ER 网络的度分布为正态分布,BA 网络的度分布服从幂律分布.

在人工数据集上,我们分别采用 3 种不同的网络模型产生人工网络 SN;而后,对所产生的每个人工网络的每条边,分别以概率 s_a 和 s_b 进行抽样保留,形成一对实验网络 SN_a 和 SN_b.在各项实验中,鉴于召回率、准确率、F1-Measure 等指标具有相似的趋势,因此,本文仅对召回率进行分析.

图 6 是在 s_a=s_b=0.5 的情况下,3 种算法随先验关联用户数变化的对比图.图 6(a)和图 6(b)是 p=0.05 的情况下,ER 网络和 WS 网络的实验结果对比.实验分别在包含 1 000 个和 5 000 个节点的两组网络中开展.不难看出:随着先验关联用户数的增加,FRUI 和 NS 算法的召回率也增加.在 n=1000 的 ER 网络中,FRUI 算法需要 60 个先验关联用户可挖掘出几乎所有的关联用户;而给定 100 个先验关联用户,NS 算法仅找出不到 80%的关联用户.在 n=5000 的 ER 网络中,FRUI 算法和 NS 算法分别需要 50 和 100 个先验关联用户即可找出几乎所有的关联用户.在 n=1000 的 WS 网络中,给定 100 的先验关联用户,FRUI 算法可识别 80%的关联用户,而 NS 算法可识别 40%的关联用户.在 n=5000 的 WS 网络中,给定 2%的先验关联用户,FRUI 和 NS 算法都可识别出大约 80%的关联用户.图 6(c)是 3 种算法在 m=20 的情况下,n=10000 和 n=20000 的 BA 网络中的实验对比.不难看出,FRUI 的算法要明显优于 NM 和 NS 算法.此外,由于两个网络的重叠度为 0.5×0.5=0.25,重叠度较低,因此,NM 算法的效率较差,本文仅列出 n=1000 的实验结果.FRUI 算法在该系列实验中都要优于 NS 算法和 NM 算法.



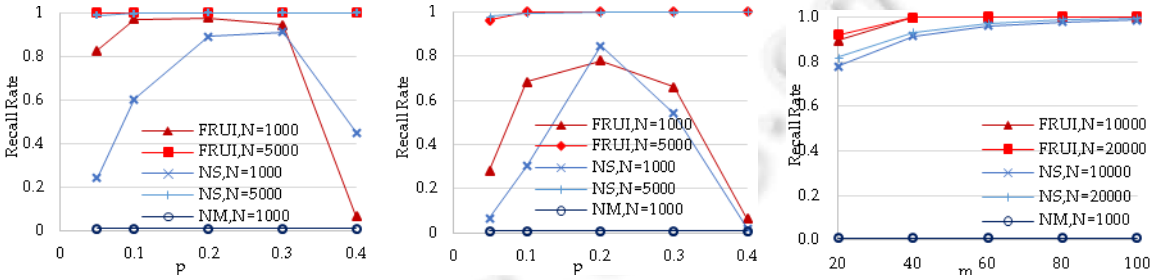
(a) ER 网络,p=0.05 (b) WS 网络,p=0.05 (c) BA 网络,m=20

Fig.6 Comparisons of FRUI, NS and NM in three synthetic networks along with the number of prior knowledge

图 6 FRUI,NS 和 NM 算法在 3 种不同人工网络中随先验用户数变化的实验对比

图 7 是给定 5%先验关联用户的情况下,3 种算法在不同密度网络中的执行效率.图 7(a)和图 7(b)为 3 种算法分别在 p=0.05,0.1,0.2,0.3,0.4 中的召回率对比.在 n=1000 的 ER 和 WS 网络中,FRUI 和 NS 算法随着网络密度的增加,关联用户挖掘效果增强,而后降低;当 p=0.2 时,两种算法的效率最好.在 n=5000 的 ER 和 WS 网络中,FRUI

和 NS 算法都能够挖掘几乎所有的关联用户.整体上,FRUI 的召回率都优于 NS 算法.图 7(c)为 3 种算法在 BA 网络中,m 值以 20 的梯度从 20 增加到 100 的召回率对比.显然,随着网络密度的增加,FRUI 和 NS 算法的召回率也增加.在 3 种人工网络中,FRUI 算法的召回率都略优于 NS 算法;而 NM 算法几乎很难挖掘出关联.不难发现,图 7(a)和图 7(b)还显示:在 ER 和 WS 网络中,当 p 值较大或较小时,FRUI 和 NS 算法都需要更多的先验用户才能得到更好的关联用户挖掘效果.这也佐证了第 2.2.1 节对 FRUI 算法的分析.

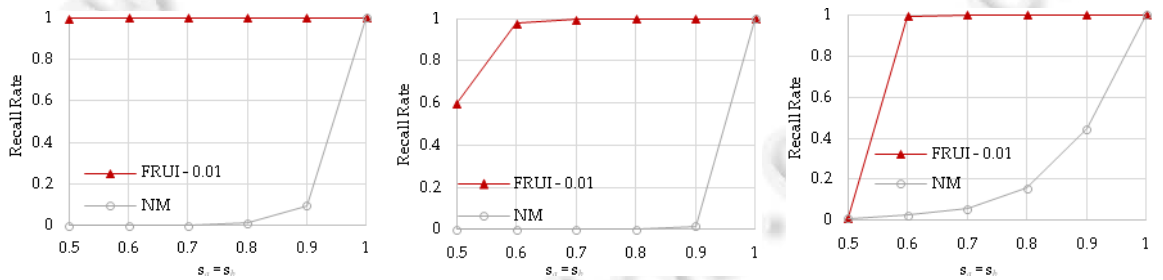


(a) ER 网络 (b) WS 网络 (c) BA 网络

Fig.7 Comparisons of FRUI, NS and NM in three synthetic networks along with the density of the network

图 7 FRUI,NS 和 NM 算法在 3 种不同人工网络中随网络密度变化的实验对比

图 8 为 FRUI 算法和 NM 算法随网络重叠度增加的实验对比.由于 FRUI 算法整体上优于 NS 算法,因此并未对 NS 算法作进一步对比.在 FRUI 算法中,随机给定 1%的先验关联用户.在 ER 和 WS 网络中,p=0.05;在 BA 网络中,m=20.显然,随着网络重叠度增加,两种算法的召回率也跟着增加.然而可以明显地看到:NM 算法只有在两个网络的重叠度很大的情况下才能取得较好的关联用户挖掘效果;而 FRUI 算法在较稀疏的网络重叠度下,也能取得较好的召回率.这也说明了当前去匿名化的相关方法在关联用户挖掘中的适应性较差.



(a) ER 网络,p=0.05 (b) WS 网络,p=0.05 (c) BA 网络,m=20

Fig.8 Comparisons of FRUI, NS and NM in three synthetic networks along with the overlap of the networks

图 8 FRUI 和 NM 算法在 3 种不同人工网络中随网络重叠度变化的实验对比

3.3 真实数据集实验分析

为进一步验证算法在关联用户挖掘的适用性,本文在新浪微博和人人网好友关系网络数据集上作了进一步实验.两个数据集的数据统计见表 3:新浪微博有 117 万个用户和 190 万条好友关系,人人网有 550 万个用户和 1 460 万条好友关系.两个网络的度分布如图 9 所示.

在此基础上,本文分别从两个网络中抽取 50 000 用户及其用户关系,而后对所抽取的网络进行随机抽样,形成具有一定重叠度的一堆重叠网络.本文使用 Jaccard 系统来定义两个网络的节点/关系重叠度,即:

$$overlap(X,Y)=\frac{|X\cap Y|}{|X\cup Y|} \tag{21}$$

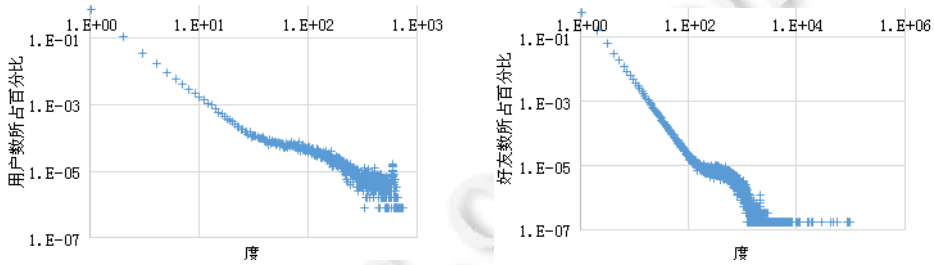
其中,$overlap()$,X 和 Y 分别表示两个网络的重叠度和节点/关系的集合.因此,当两个网络有 2/3 的节点相同时,其

节点重叠度为 0.5.

Table 3 Networks of the ground truth dataset

表 3 真实数据集网络

数据集	节点数	边数	平均度数
新浪微博	1.17M	1.9M	3.2
人人网	5.5M	14.6M	5.3



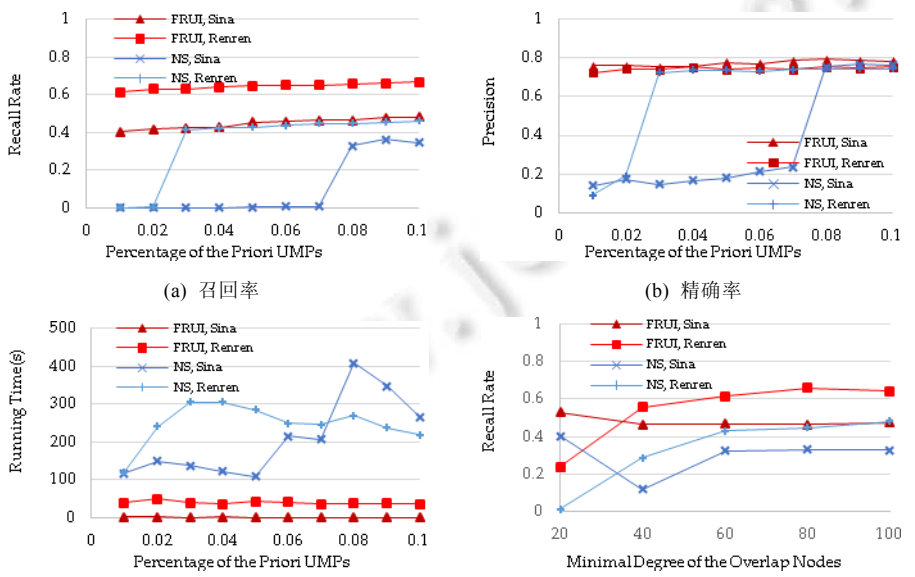
(a) 新浪微博数据集 (b) 人人网数据集

Fig.9 Degree distributions of networks in ground truth dataset

图 9 真实数据集数据集度分布

实验中,我们随机的选取一部分关联节点作为先验节点.鉴于 NM 算法在关联用户挖掘中的效用相对较差,本节只对比 FRUI 和 NS 算法.具体地,我们将先验节点以 0.01 的梯度从 0.01 逐步增加到 0.1 以观察实验结果.由于新浪微博和人人网都相对较稀疏,存在大量度很低的节点.因此,本文只选用度数值不低于ε的节点作为种子节点.为比较不同种子节点对关联用户挖掘的影响,本文以 20 的梯度逐步增加ε值来进行实验对比.

图 10(a)对比了ε=80 的情况下,FRUI 和 NS 算法的召回率(图 10 中,节点重叠度为 33%,边重叠度 33%.在图 10(a)~图 10(c)中,节点最小度值为 80.在图 10(d)中,先验关联用户占比 8%.).



(a) 召回率 (b) 精确率

(c) 运行时间 (d) 召回率

Fig.10 Comparison between FRUI and NS in Sina and Renren datasets

图 10 FRUI 和 NS 算法在新浪微博和人人网数据集中的实验对比

在新浪微博数据集中,给定 5%的先验用户,FRUI 能够挖掘大约 50%的关联用户;而 NS 在 10%的先验用户下只获得 40%的召回率.同样地,FRUI 在人人网数据集上也有更好的表现.图 10(b)对比了两种方法的在ε=80 情况下的准确率.显然,FRUI 的准确率要高于 NS 算法.图 10(c)显示,FRUI 算法的运行效率也要明显优于 NS.图 10(d)描述了 FRUI 和 NS 两种算法随ε值变化召回率的变化曲线.虽然ε值对两种算法都有一定的影响,但 FRUI 的算法基本上都要优于 NS 算法.因此可以说,FRUI 算法是当前基于网络结构算法中效果相对较好的算法.这也佐证了第 2.2.1 节中对 FRUI 算法和 NS 算法的分析.

4 总结与展望

关联用户挖掘建模在形式上与其他领域的许多研究相似或相关,如自然语言处理中的共献问题[59]、实体匹配[60]、数据库记录链接[61−63]和信息检索中的命名辨别问题[64−66]等.虽然这些方法为社会网络关联用户挖掘提供借鉴,然而,由于社会网络中数据量大以及用户属性的相似性、稀疏性、虚假性和不一致性,使得面向社会网络融合的关联用户挖掘方法将面临更多的挑战,也将更为复杂.

综上所述,目前面向社会网络融合的关联用户挖掘研究现状可以总结为:

1) 从用户属性(含用户行为)中挖掘关联用户是目前研究最多、效果最好的方法,但在社会网络中,由于用户属性的相似性、稀疏性、虚假性和不一致性,使得单纯使用用户属性挖掘关联用户的方法健壮性不足,易于受恶意用户的攻击,且召回率可进一步提升;
2) 基于用户关系的关联用户挖掘研究多在去匿名化研究领域.在较为相似或相同的图中,该方法能够较准确挖掘节点度较高的节点;
3) 目前,大多数关联用户挖掘方法需要种子用户:一方面,其挖掘质量较依赖于种子用户的质量;另一方面,在许多社会网络中,种子用户的获取越来越困难.目前,针对无种子用户的关联用户挖掘建模研究还很少;
4) 融合用户属性和用户关系以挖掘准确、全面的关联用户的研究较少;针对社会网络,建立快速的关联用户挖掘模型也有待进一步研究.

围绕社会网络关联用户挖掘当前的研究现状,未来可开展的研究方向包括:

1) 用户属性对关联用户挖掘的效用评价体系.社会网络包含用户名、头像、用户行为等多种属性信息,然而不同的用户属性对关联用户挖掘的效用不一样,甚至某些属性可能会对关联用户挖掘起反作用.为此,如何从社会网络用户属性的相似性、稀疏性、虚假性和不一致性出发构建面向关联用户挖掘的用户属性效用评价体系,遴选适合关联用户挖掘的用户属性,将为构建准确、全面的关联用户挖掘模型和方法建立基础;
2) 在无先验用户的情况下,基于用户关系的关联用户挖掘方法.随着先验用户获取越来越困难,面向无先验用户的关联用户挖掘方法已引起了广泛关注.鉴于用户关系(尤其是好友关系)的可靠性和稳定性,如何在无先验或者极少先验的情况下精确地挖掘关联用户,将是当前关联用户挖掘的重要研究内容.其研究成果将可为基于先验关联用户的方法提供先验知识,也将可为去匿名化提供借鉴思路;
3) 面向大数据的关联用户挖掘模型及其求解方法.现阶段,许多关联挖掘模型都局限于小规模的数据量,例如 NM 算法[50],如何采用低秩矩阵分解[67]、深度学习[68]、并行计算[69]等前沿理论和方法高效解决海量数据下的社会网络关联用户挖掘,将是一个重要的研究方向;
4) 综合用户属性和用户关系的关联用户挖掘混杂模型和方法.在用户属性中融入用户关系,构建关联用户挖掘模型,可以避免模型受恶意用户的攻击,提升模型的准确率;在用户关系中融入用户属性,可以更准确地识别度数较低的用户,提升关联用户挖掘模型的准确率和召回率.综合用户属性和用户关系,是社会网络关联用户挖掘的必然趋势.然而,用户属性和用户关系是社会网络的不同要素,用户在属性上的相似性易于用相似度表达.不同的社会网络具有不同的用户关系,在无先验用户的情况下,现有的理论和方法较难给出一种适用于关联用户挖掘的用户关系相似度计算模型,从而无法将用户

属性和用户关系统一于不同维度上的相似度融合.统一用户属性和用户关系,构建关联用户挖掘混杂模型和方法,将是一种必然.

5) 跨社会网络研究.社会网络融合将为社会网络各项研究提供更充分的数据基础.如何利用社会网络融合的研究成果开展跨社会网络研究,将是未来的一个重要趋势.例如协同推荐的冷启动问题[70],对于 SN_A 中的用户 U_{Ai},其用户关系为 F_{Ai}.通过关联用户挖掘,F_{Ai} 中部分用户在 SN_B 的关联用户集合为 F'_{Bi}.当 U_{Ai} 新注册 SN_B 时,可合理为其推荐其潜在好友 F'_{Bi},从而为冷启动提供新的解决思路.

面向社会网络融合的关联用户挖掘方法已逐渐引起了学术和产业界的关注:一方面,其方法研究将能为社会网络的去匿名化问题提供借鉴,为协同推荐的冷启动问题提供新的解决思路;另一方面,其将直接从网络节点上建立社会网络间的关联,为社会计算等社会网络挖掘提供更充分地用户行为数据和网络结构,有利于人们更好地通过社会网络认识人类社会,其研究具有重要的理论和实践意义.

References:

[1] Ellison NB. Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 2007,13(1): 210−230. [doi: 10.1111/j.1083-6101.2007.00393.x]

[2] Wang FY, Li XC, Mao WJ, Wang T. Social Computing: Methods and Applications. Hangzhou: Zhejiang University Press, 2014 (in Chinese).

[3] Zhou XP, Liang X, Zhang HY. User community detection on micro-blog using R-C model. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2808−2823 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4720.htm [doi: 10.13328/j.cnki.jos.004720]

[4] Morone F, Makse HA. Influence maximization in complex networks through optimal percolation. Nature, 2015,527(7579):544. [doi: 10.1038/nature14604]

[5] Aakas Z, Xun L, Zhou X. Learning structural features of nodes in large-scale networks for link prediction. In: Proc. of the AAAI. 2016.

[6] Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2008,2(1-2):1−135.

[7] Chen H, Chiang RHL, Storey VC. Business intelligence and analytics: From big data to big impact. MIS Quarterly, 2012,36(4): 1165−1188.

[8] Xiang Z, Gretzel U. Role of social media in online travel information search. Tourism Management, 2010,31(2):179−188. [doi: 10.1016/j.tourman.2009.02.016]

[9] Tang L, Liu H, Wrote; Wen YM, Bi YZ, Trans. Community Detection and Mining in Social Media. Beijing: China Machine Press, 2012 (in Chinese).

[10] Liang X, Yang XP, Zhou XP, Zhang HY. Social Computing on the Big Data of Social Media. Beijing: Tsinghua University Press, 2014 (in Chinese).

[11] Li CY, Lin SD. Matching users and items across domains to improve the recommendation quality. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2014. 801−810. [doi: 10.1145/2623330.2623657]

[12] Bodhit A, Amin K. Possible solutions of new user or item cold-start problem. Int'l Journal of Mathematics, 2013,1(3).

[13] Zhou XP, Liang X, Zhang HY, Ma YF. Cross-Platform identification of anonymous identical users in multiple social media networks. IEEE Trans. on Knowledge and Data Engineering, 2016,28(2):411−424. [doi: 10.1109/TKDE.2015.2485222]

[14] Liu S, Wang S, Zhu F. Structured learning from heterogeneous behavior for social identity linkage. IEEE Trans. on Knowledge and Data Engineering, 2015,27(7):2005−2019. [doi: 10.1109/TKDE.2015.2397434]

[15] Liu S, Wang S, Zhu F, Zhang J, Krishnan R. Hydra: Large-scale social identity linkage via heterogeneous behavior modeling. In: Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2014. 51−62. [doi: 10.1145/2588555.2588559]

[16] Zafarani R, Liu H. Connecting users across social media sites: A behavioral-modeling approach. In: Proc. of the 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2013. 41−49. [doi: 10.1145/2487575.2487648]

[17] Zhang Y, Tang J, Yang Z, Pei J, Yu PS. COSNET: Connecting heterogeneous social networks with local and global consistency. In: Proc. of the 21th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2015. 1485−1494. [doi: 10.1145/2783258.2783268]

[18] Korula N, Lattanzi S. An efficient reconciliation algorithm for social networks. Proc. of the VLDB Endowment, 2014,7(5). [doi: 10.14778/2732269.2732274]

[19] Lu CT, Shuai HH, Yu PS. Identifying your customers in social networks. In: Proc. of the 23rd ACM Int'l Conf. on Information and Knowledge Management. ACM Press, 2014. 391−400. [doi: 10.1145/2661829.2662057]

[20] Kong X, Zhang J, Yu PS. Inferring anchor links across multiple heterogeneous social networks. In: Proc. of the 22nd ACM Int'l Conf. on Information and Knowledge Management. ACM Press, 2013. 179−188. [doi: 10.1145/2505515.2505531]

[21] Goga O, Lei H, Parthasarathi SHK, Friedland G, Sommer R, Teixeira R. Exploiting innocuous activity for correlating users across sites. In: Proc. of the 22nd Int'l Conf. on World Wide Web. New York: ACM, 2013. 447−458. [doi: 10.1145/2488388.2488428]

[22] Balduzzi M, Platzer C, Holz T, Kirda E, Balzarotti D, Kruegel C. Abusing social networks for automated user profiling. In: Proc. of the Recent Advances in Intrusion Detection. Berlin, Heidelberg: Springer-Verlag, 2010. 422−441. [doi: 10.1007/978-3-642-15512-3_22]

[23] Irani D, Webb S, Li K, Pu C. Large online social footprints—An emerging threat. In: Proc. of the Int'l Conf. on Computational Science and Engineering (CSE 2009). IEEE, 2009. 271−276. [doi: 10.1109/CSE.2009.459]

[24] Chen Y, Zhuang C, Cao Q, Hui P. Understanding cross-site linking in online social networks. In: Proc. of the 8th Workshop on Social Network Mining and Analysis. ACM Press, 2014. 6. [doi: 10.1145/2659480.2659498]

[25] Ma XJ, Sun YQ, Liu FP. Privacy protection in social media. China Computer Federal Communication, 2011,7(1):52−56 (in Chinese).

[26] Xiao Y, Xiong M, Wang W, Wang H. Emergence of symmetry in complex networks. Physical Review E, 2008,77(6):066108. [doi: 10.1103/PhysRevE.77.066108]

[27] Lyzinski V, Fishkind DE, Fiori M, Vogelstein JT, Priebe CE, Sapiro G. Graph matching: Relax at your own risk. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2016,38(1):60−73. [doi: 10.1109/TPAMI.2015.2424894]

[28] Ullmann JR. An algorithm for subgraph isomorphism. Journal of the ACM (JACM), 1976,23(1):31−42. [doi: 10.1145/321921.321925]

[29] Perito D, Castelluccia C, Kaafar MA, Manils P. How unique and traceable are usernames? In: Proc. of the Privacy Enhancing Technologies. Berlin, Heidelberg: Springer-Verlag, 2011. 1−17. [doi: 10.1007/978-3-642-22263-4_1]

[30] Zafarani R, Liu H. Connecting corresponding identities across communities. In: Proc. of the 3rd Int'l Conf. on Weblogs & Social Media (ICWSM), 2009. 354−357.

[31] Liu J, Zhang F, Song X, Song Y, Lin C, Hon H. What's in a name? An unsupervised approach to link users across communities. In: Proc. of the 6th ACM Int'l Conf. on Web Search and Data Mining. ACM Press, 2013. 495−504. [doi: 10.1145/2433396.2433457]

[32] Acquisti A, Gross R, Stutzman F. Faces of facebook: Privacy in the age of augmented reality. 2011.

[33] Iofciu T, Fankhauser P, Abel F, Bischoff K. Identifying users across social tagging systems. In: Proc. of the ICWSM. 2011.

[34] Motoyama M, Varghese G. I seek you: Searching and matching individuals in social networks. In: Proc. of the 11th Int'l Workshop on Web Information and Data Management. ACM Press, 2009. 67−75. [doi: 10.1145/1651587.1651604]

[35] Ye N, Zhao YL, Bian GQ, Li J, He J. A schema-independent user identification algorithm in social networks. Journal of Xi'an Jiaotong University, 2013,47(12):19−25 (in Chinese with English abstract).

[36] Zhang H, Kan MY, Liu Y, Ma SP. Online social network profile linkage. In: Proc. of the Information Retrieval Technology. Springer Int'l Publishing, 2014. 197−208. [doi: 10.1007/978-3-319-12844-3_17]

[37] Mylka A, Sauermann L, Sintek M, van Elst L. Nepomuk contact ontology. Technical Report, 2007.

[38] Cortis K, Scerri S, Rivera I, Handschuh S. An ontology-based technique for online profile resolution. In: Proc. of the Social Informatics. Springer Int'l Publishing, 2013. 284−298. [doi: 10.1007/978-3-319-03260-3_25]

[39] Zheng R, Li J, Chen H, Huang Z. A framework for authorship identification of online messages: Writing-Style features and classification techniques. Journal of the American Society for Information Science and Technology, 2006,57(3):378−393. [doi: 10.1002/asi.20316]

[40] Almishari M, Tsudik G. Exploring linkability of user reviews. In: Proc. of the Computer Security (ESORICS 2012). Berlin, Heidelberg: Springer-Verlag, 2012. 307−324. [doi: 10.1007/978-3-642-33167-1_18]

[41] Nie Y, Huang J, Li A, Zhou B. Identifying users based on behavioral-modeling across social media sites. In: Proc. of the Web Technologies and Applications. Springer Int'l Publishing, 2014. 48−55. [doi: 10.1007/978-3-319-11116-2_5]

[42] Qi X, Wu TJ. Node matching between complex networks. Physical Review E, 2009,80(2):026103. [doi: 10.1103/PhysRevE.80.026103]

[43] Bartunov S, Korshunov A, Park ST, Ryu W, Lee H. Joint link-attribute user identity resolution in online social networks. In: Proc. of the 6th Int'l Conf. on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis. ACM Press, 2012.

[44] Li XF, Liang X, Zhou XP, Shi XJ, Shen H, Zhang HY. A user matching method for cross-platform microblogging services. Patent, No.201410000302.9, 2014 (in Chinese).

[45] Xun Q. Node matching between complex networks based on genetic algorithm. Journal of Heilongjiang Institute of Science & Technology, 2011,21(3):244−248 (in Chinese with English abstract).

[46] Narayanan A, Shmatikov V. De-Anonymizing social networks. In: Proc. of the 2009 30th IEEE Symp. on Security and Privacy. IEEE, 2009. 173−187. [doi: 10.1109/SP.2009.22]

[47] Narayanan A, Shi E, Rubinstein BIP. Link prediction by de-anonymization: How we won the kaggle social network challenge. In: Proc. of the 2011 Int'l Joint Conf. on Neural Networks (IJCNN). IEEE, 2011. 1825−1834. [doi: 10.1109/IJCNN.2011.6033446]

[48] Nilizadeh S, Kapadia A, Ahn YY. Community-Enhanced de-anonymization of online social networks. In: Proc. of the 2014 ACM SIGSAC Conf. on Computer and Communications Security. ACM Press, 2014. 537−548. [doi: 10.1145/2660267.2660324]

[49] Singh R, Xu J, Berger B. Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc. of the National Academy of Sciences, 2008,105(35):12763−12768. [doi: 10.1073/pnas.0806627105]

[50] Fu H, Zhang A, Xie X. Effective social graph deanonymization based on graph structure and descriptive information. ACM Trans. on Intelligent Systems and Technology (TIST), 2015,6(4):49. [doi: 10.1145/2700836]

[51] Fu H, Zhang A, Xie X. De-Anonymizing social graphs via node similarity. In: Proc. of the Companion Publication of the 23rd Int'l Conf. on World Wide Web Companion. 2014. 263−264. [doi: 10.1145/2567948.2577366]

[52] Pedarsani P, Figueiredo DR, Grossglauser M. A Bayesian method for matching two similar graphs without seeds. In: Proc. of the Allerton. 2013. 1598−1607. [doi: 10.1109/Allerton.2013.6736720]

[53] Jain P, Kumaraguru P, Joshi A. @ i seek'fb. me': Identifying users across multiple online social networks. In: Proc. of the 22nd Int'l Conf. on World Wide Web Companion. 2013. 1259−1268. [doi: 10.1145/2487788.2488160]

[54] Jain P, Kumaraguru P. Finding nemo: Searching and resolving identities of users across online social networks. arXiv preprint arXiv:1212.6147, 2012.

[55] Yu M. Entity linking on graph data. In: Proc. of the Companion Publication of the 23rd Int'l Conf. on World Wide Web Companion. 2014. 21−26. [doi: 10.1145/2567948.2567954]

[56] Erdős P, Rényi A. On random graphs I. Publicationes Mathematicae Debrecen, 1959,6:290−297.

[57] Watts DJ, Strogatz SH. Collective dynamics of 'small-world'networks. Nature, 1998,393(6684):440−442. [doi: 10.1038/30918]

[58] Barabási AL, Albert R. Emergence of scaling in random networks. Science, 1999,286(5439):509−512. [doi: 10.1126/science.286.5439.509]

[59] Cai J, Strube M. End-to-End coreference resolution via hypergraph partitioning. In: Proc. of the 23rd Int'l Conf. on Computational Linguistics. Association for Computational Linguistics, 2010. 143−151.

[60] Wang J, Li G, Yu JX, Feng JH. Entity matching: How similar is similar. Proc. of the VLDB Endowment, 2011,4(10):622−633. [doi: 10.14778/2021017.2021020]

[61] Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate record detection: A survey. IEEE Trans. on Knowledge and Data Engineering, 2007,19(1):1−16. [doi: 10.1109/TKDE.2007.250581]

[62] Hassanzadeh O, Pu KQ, Yeganeh SH, Miller RJ, Popa L, Hernández MA, Ho H. Discovering linkage points over Web data. Proc. of the VLDB Endowment, 2013,6(6): 445−456. [doi: 10.14778/2536336.2536345]

[63] Sadinle M, Fienberg SE. A generalized Fellegi–Sunter framework for multiple record linkage with application to homicide record systems. Journal of the American Statistical Association, 2013,108(502):385−397. [doi: 10.1080/01621459.2012.757231]

[64] Kalashnikov DV, Chen Z, Mehrotra S, Nuray-Turan R. Web people search via connection analysis. IEEE Trans. on Knowledge and Data Engineering, 2008,20(11):1550−1565. [doi: 10.1109/TKDE.2008.78]

[65] Qian Y, Hu Y, Cui J, Zheng QH, Nie ZQ. Combining machine learning and human judgment in author disambiguation. In: Proc. of the 20th ACM Int'l Conf. on Information and Knowledge Management. ACM Press, 2011. 1241−1246. [doi: 10.1145/2063576.2063756]

[66] Tang J, Fong ACM, Wang B, Zhang J. A unified probabilistic framework for name disambiguation in digital library. IEEE Trans. on Knowledge and Data Engineering, 2012,24(6):975−987. [doi: 10.1109/TKDE.2011.13]

[67] Lee J, Kim S, Lebanon G, Singer Y. Local low-rank matrix approximation. In: Proc. of the Proceedings of the 30th Int'l Conf. on Machine Learning. 2013,28:82−90. [doi: 10.1038/nature14539]

[68] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015,521(7553):436−444.

[69] Chin WS, Zhuang Y, Juan YC, Lin CJ. A fast parallel stochastic gradient method for matrix factorization in shared memory systems. ACM Trans. on Intelligent Systems and Technology (TIST), 2015,6(1):2. [doi: 10.1145/2668133]

[70] Huang S, Zhang J, Wang L, Hua XS. Social friend recommendation based on multiple network correlation. IEEE Trans. on Multimedia, 2016,18(2):287−299. [doi: 10.1109/TMM.2015.2510333]

附中文参考文献:

[2] 王飞跃,李晓晨,毛文吉,王涛.社会计算的基本方法和应用.杭州:浙江大学出版社,2014.

[3] 周小平,梁循,张海燕.基于 R-C 模型的微博用户社区发现.软件学报,2014,25(12):2808−2823. http://www.jos.org.cn/1000-9825/4720.htm [doi: 10.13328/j.cnki.jos.004720]

[9] Tang L,Liu H,著;文益民,闭应洲,译.社会计算:社区发现和社会媒体挖掘.北京:机械工业出版社,2012.

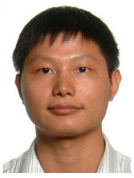
[10] 梁循,杨小平,周小平,张海燕.面向社会化媒体大数据的社会计算.北京:清华大学出版社,2014.

[25] 马晓君,孙宇清,刘发朋.社会网络中的隐私保护.中国计算机学会通讯,2011,7(1):52−56.

[35] 叶娜,赵银亮,边根庆,李健,何菁.模式无关的社交网络用户识别算法.西安交通大学学报,2013,47(12):19−25.

[44] 李晓菲,梁循,周小平,施晓菁,申华,张海燕.一种跨平台微博社区账户匹配方法.国家发明专利,申请号:201410000302.9.

[45] 徐钦.基于遗传算法的复杂网络节点匹配问题.黑龙江科技学院学报,2011,21(3):244−248.



周小平(1985－),男,福建寿宁人,讲师,主要研究领域为社会计算,数据挖掘.

梁 循(1965－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为数据挖掘,商务智能,社会计算.

赵吉超(1992－),女,硕士生,主要研究领域为社会计算,推荐系统.

李志宇(1991－),男,博士生,CCF 学生会员,主要研究领域为深度学习.

马跃峰(1976－),男,副教授,CCF 学生会员,主要研究领域为机器学习.