软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(3):811−823 [doi: 10.13328/j.cnki.jos.005448]

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

全视角特征结合众包的跨社交网络用户识别*

汪 潜, 申德荣, 冯 朔, 寇 月, 聂铁铮, 于 戈

(东北大学 计算机科学与工程学院,辽宁 沈阳 110169)

通讯作者: 汪潜, E-mail: workwith_young@163.com

摘 要: 随着互联网的普及和不断发展,用户通过多个社交网络进行社交活动,使用社交网络带来的丰富内容和服务.通过识别出不同社交网络上的同一用户,可以有助于进行用户推荐、行为分析、影响力最大化.已有方法主要基于用户的结构特征和属性特征来识别匹配用户,大多仅考虑局部结构,且受已知匹配用户数量的限制,提出一种基于全视角特征结合众包的跨社交网络用户识别方法(overall and crowdsourced user identification algorithm,简称OCSA).首先,利用众包提高已知匹配用户的数量;然后,应用全视角特征评价用户的相似度,以提升用户匹配的准确性;最后,利用两阶段的迭代式匹配方法完成用户识别工作.实验结果表明:该算法可显著提高用户识别的召回率和准确率,并解决了已知匹配用户数量不足时的识别问题.

关键词: 多社交网络;用户识别;众包

中图法分类号: TP311

中文引用格式: 汪潜,申德荣,冯朔,寇月,聂铁铮,于戈.全视角特征结合众包的跨社交网络用户识别.软件学报,2018,29(3):811−823. http://www.jos.org.cn/1000-9825/5448.htm

英文引用格式: Wang Q, Shen DR, Feng S, Kou Y, Nie TZ, Yu G. Identifying users across social networks based on global view features with crowdsourcing. Ruan Jian Xue Bao/Journal of Software, 2018,29(3):811−823 (in Chinese). http://www.jos.org.cn/1000-9825/5448.htm

Identifying Users Across Social Networks Based on Global View Features with Crowdsourcing

WANG Qian, SHEN De-Rong, FENG Shuo, KOU Yue, NIE Tie-Zheng, YU Ge

(College of Computer Science and Engineering, Northeastern University, Shenyang 110169, China)

Abstract: With the popularity and development of Internet, people like to take part in multiple social networks to enjoy different kinds of services. Consequently, an important task is to identify users in the networks, which is helpful for user recommendation, behavior analysis and impact maximization. Most state-of-the-art works on this issue are mainly based on the user's structure features and attribute features. They prefer to exploit user's local features and are limited by the number of the known matching users. In this paper, a method based on global view features is proposed to align users with crowdsourcing (OCSA). First, crowdsourcing is used to increase the number of known matching users on networks. Then, global view features are used to evaluate the similarity between users to improve the accuracy of user identification. Finally, an iterative two-stage matching method is put forward to answer the user identification. The results of experiments show that the presented method has better performance on precision and recall, especially when the number of known matching users is insufficient.

Key words: multiple-social network; user identification; crowdsourcing

* 基金项目: 国家自然科学基金(61472070, 61672142); 国家重点基础研究发展计划(973)(2012CB316201)

Foundation item: National Natural Science Foundation of China (61472070, 61672142); National Basic Research Program of China (973) (2012CB316201)

本文由基于图结构的大数据分析与管理技术专刊特约编辑林学民教授、杜小勇教授、李翠平教授推荐.

收稿时间: 2017-07-31; 修改时间: 2017-09-05; 采用时间: 2017-11-07; jos 在线出版时间: 2017-12-05

CNKI 网络优先出版: 2017-12-06 15:23:29, http://kns.cnki.net/kcms/detail/11.2560.TP.20171206.1523.012.html

社交网络作为 Web 2.0 时代的产物,为人们提供了丰富的社交服务,据统计,42%的社交网络用户同时使用多种社交网络[1],因为不同的社交网络拥有各自不同的社交方式,带给用户不同的社交服务.将这些社交网络融合成一个单一的环境,将有助于完善用户画像、提升好友推荐和提高商业广告投放精确度等.已有相关研究[1−3]的核心思想是:先基于社交网络中用户的结构特征和属性特征,使用机器学习模型化用户匹配模型,再从已匹配用户对出发,选取已匹配用户的邻居(局部特征)进行迭代扩散,实现匹配过程.

分析已有方法[1−3],典型具有如下几点不足.

- 第一,受限于用户匹配模型的识别准确率.已有的匹配模型严格依赖于提取的特征和基于机器学习的建立过程,然而在实际社交网络中,由于用户自身的安全考虑或网站的隐私保护等,存在所获取的用户属性并不完整、数据格式异构等问题,导致机器学习特征不完备,影响匹配准确性;
- 第二,模型仅基于用户的局部特征,易受到伪造用户的干扰.例如,社交网络上存在大量的营销或恶意账号,可成功欺骗该类匹配方法.如图 1 所示,下方的伪造用户从头像、个人信息、个人简介等方面都与上方真实用户完全一致,甚至用户名几乎相同;
- 第三,存在已知匹配用户对数量少的冷启动问题:当已知用户匹配对所占比例少时,在迭代式匹配、特征提取和参数训练过程中,都会因为传递信息过少,造成匹配结果正确率、召回率过低的局限性.

Fig.1 A fake user on social networks

图 1 微博上的伪造用户

为此,针对已有工作的不足,本文研究结合众包的用户识别问题.鉴于众包[4]是一种利用人工智慧来提高匹配准确性的有效方法,本文结合众包技术来弥补机器学习无法完全描述用户特征的局限,解决已知匹配用户对数量少的冷启动问题,且能有效甄别伪造用户,进而提高跨社网的用户匹配准确率;同时,模型化用户的全局特征(即多视角特征(见定义 3)),改善基于局部特征的识别用户的局限性,进一步提高跨社网匹配的准确性.

本文的主要贡献如下:提出了一种基于全视角特征结合众包的跨社交网络用户识别方法(overall and crowdsourced user identification algorithm,简称 OCSA).

- 首先,利用众包平台匹配部分用户,解决机器学习算法中存在的已知用户匹配度少的冷启动问题;
- 其次,依据社交网络中用户按社区分布的特征,提出用户全视角特征的定义和计算方法,提高跨网的用户识别准确率;
- 第三,基于前面的工作,提出了两阶段的用户识别框架;
- 最后,通过实验证明,本文提出的方法能够有效地识别社交网络用户,减少了对已知匹配用户对的依赖,提高了用户识别的准确率.

1 相关工作

已有多社交网络上的用户识别研究,按照使用特征信息的不同,大致分为 3 种:基于用户属性信息的匹配方法[5−7]、基于用户拓扑结构的匹配方法[8,9]和两种特征混合的匹配方法[1−3].其中,早期大部分都是围绕用户的属

性信息展开研究,近年来侧重后两种方法研究.

基于属性的用户识别算法主要考虑用户的个人信息和发表内容,通过建立分类模型,使用相应的匹配策略来完成识别工作.例如:Zafarani 和 Liu[5]提出了基于用户名的精确匹配方法;Vosecky 等人[6]提出了结合属性权重的用户匹配思想,以及精确匹配、模糊匹配、部分匹配等不同的匹配策略;Raad[7]则提出一种基于 FOAF 的属性匹配框架.以上这些基于属性的识别方法都面临着领域知识不足、属性信息不充分等问题,而且无法识别恶意用户的干扰.

针对以上不足,出现了大量结合图匹配、随机图模型等的用户识别算法的研究工作.典型的是将社交网络看做是用户好友关系构成的图结构,从而将用户识别问题转化成图匹配问题.例如:Liu[8]基于用户行为和核心网络的思想提出了 HYDRA 算法,该算法将用户的行为模型和用户的核心结构模型结合在一起,利用用户属性和用户生成内容(user generated content,简称 UGC)建模用户行为;Korula[9]将社交网络用户识别问题抽象为严格的数学定义,认为不同的社交网络其实都是由一个潜在的用户图结构通过概率生成的,并且图的边的选择过程是近似概率的,且存在一种级联效应.基于这样的思想,采用迭代的方法来识别同一实体用户.

近几年提出的混合式用户识别的研究,这种匹配方法可以充分利用用户属性相似度、结构相似度及其传递性关系.例如:Kong 等人[2]提出的 MNA 算法考虑用户的局部结构,通过稳定婚姻匹配算法来满足用户映射的一对一约束;Zhang 等人[1]从能量模型的角度出发,把不同的属性和结构匹配分成不同的能量级,使得当匹配结果达到最优情况时能量最低,从而将匹配问题转化成能量模型的求参过程,并且采用对偶问题分解的思路来提高问题解决的效率;Zhang 等人[3]提出了元路径的概念,即,通过不同的语义路径描述两点之间的关系,并且转化成特征向量的形式,同时,通过预剪枝、自我匹配等方式提高算法效率和准确率.以上这些算法主要是集中于提取特征和构造模型,对于机器学习难以表达的复杂细节很难识别.

众包是一种利用人工进行批量处理的方法,通过划分任务、设定相应的问题,并选择交给人工去判断,根据返回结果进行相应的计算,可以解决问题模型建立的问题[4].较为成熟的众包平台主要有 Amazon Mechanical Turk 和 CrowedFlower,这些平台以社区的形式来提供众包服务.

已有一些基于众包的实体识别研究[10−13],主要侧重众包任务生成和判断过程的优化.例如:Whang 等人[12]通过实体识别中存在的传递闭包关系来减少众包任务数量,最大程度地降低需要人工的部分;Gokhale 等人[13]摆脱了传统众包方法中仅参与结果判断阶段的局限性,使之参与实体识别各个阶段中.与单纯的机器学习算法相比,都具有很好的准确率和召回率.

本文拟采用众包来解决用户识别问题,并通过启发式的算法来减少需要识别的众包任务,进而提高效率.与之前这些算法相比,本文的几点创新工作是:

(1) 将众包技术融入跨社交网络的用户匹配中,改善冷启动问题和匹配精度,当前还没有将众包技术结合到跨社交网络的用户匹配的相关工作;

(2) 利用用户的全视角特征(用户参与的多个社群的综合特征)提高匹配精度,而已有工作主要利用用户的局部特征实现用户匹配.

2 问题定义

本节给出一些必要的定义及研究问题描述.

定义 1(社交网络). 给定 $G=(U,E,A)$来表示社交网络,其中,U 代表用户集合,E 代表用户之间的关系集合,A 代表用户的属性集合.特别的,本文中用 $u\in U$ 代表单独用户.

定义 2(社区). 社区是社交网络上用户的集合,每个社区由具有相同兴趣爱好或者强合作关系的用户构成.这些社区内部联系紧密,社区之间联系稀疏.

给定社交网络 $G=(U,E,A)$,$S_i\subseteq U$ 代表一个社区,$S\{S_1,S_2,S_3,\ldots,S_n\}$表示单个社交网络上拥有的 n 个社区的社区划分.

定义 3(全视角特征). 将用户和各社区之间的关联关系描述为用户的全视角特征.对于社区集合 $S\{S_1,S_2,$

$S_3,\ldots,S_n\}$,用户 u 的全视角特征为 $\boldsymbol{V_u}=(v_1,v_2,\ldots,v_n)$,其中,$v_i$ 代表用户 u 和社区 S_i 之间的关联关系.

例如,用户会因为共同的兴趣爱好、相似的教育背景及工作经历等,参与到多个社区(如摄影爱好群、数据挖掘群)中,形成一种稳定的关联关系.

定义 4(激活用户). 社交网络中,活跃且有影响力的用户称为激活用户.他们个人信息完善,能够吸引更多的用户与他们交互和分享信息,是构成社交网络中社区的核心部分.

定义 5(激活用户锚点对). 跨社交网络中,代表同一真实用户的激活用户对称为激活用户锚点对.激活用户锚点对相对于社交网络的对齐和信息的传递具有重要的作用.

本文主要研究社交网络间的用户识别问题,不失一般性,以两个社交网络为例,且用户在同社交网络上最多一个账号,即,满足一对一约束.基于社交网络中用户的结构特征和属性特征,结合众包技术,采用相应的匹配策略,匹配出属于同一真实用户的账号.

3 跨社交网络用户识别模型

本文主要从两方面改善用户识别的准确性:(1) 识别激活用户,利用众包识别激活用户锚点对,改善完全依赖于机器学习模型的不足,同时提高已匹配用户对数量;(2) 利用用户在多社交网络中的全视角特征,提高用户识别的准确性.

正如前面提到的,由于各社交网络平台用户信息的异构性、匿名性和不完备性等,很难直接获取到用于匹配用户的足够信息,导致匹配用户识别率低,传递信息不充分,识别结果不理想.为此,本文提出采用众包技术,利用人工智慧去分析和判断部分用户的匹配关系,既可以增加已知的用户匹配数量,也可以有效甄别伪造用户,进而提高用户识别的准确率.受限于众包的时间效率和金钱代价,本文提出优先对激活用户进行众包识别,避免高代价地对所有用户进行众包识别.

社交网络中,用户通常会因共同的爱好,或者相同的工作、籍贯、生活经历等,形成大大小小的社区.这些社区构成了社交网络的主体,使得社交网络呈现出星网状的结构.从社交网络结构上来看,激活用户位于社区核心,非激活用户位于社区之间,联系着不同的社区.从单一用户来看,激活用户在某一方面具有自身的兴趣或身份特征,可以代表社区;而非激活用户和不同激活用户之间的关系,反映了他们自身的兴趣爱好和身份特征,并且在一段时间内较为稳定.单一用户参与多个社区所表现出来的特征,即为全视角特征(见定义 3).不同于传统的局部结构特征,全视角特征不再局限于用户的邻居,即使他的邻居重合度或者相似度发生变化,只要用户本身没有发生兴趣或者身份的变化,从全局上来看都是稳定的.伪造用户很难构造出一模一样的兴趣特征,因此可以更好地分辨出伪造用户.

综上,本文提出了一个结合众包的两阶段的迭代式用户匹配模型.如图 2 所示.第 1 阶段,结合网络表示和聚类算法划分社区并选取激活用户,利用众包来构建激活用户锚点对,以此来提高已知匹配用户对数量;第 2 阶段,利用激活用户锚点对计算用户全视角特征,并结合用户的属性和局部结构特征来识别未匹配的用户对;为了提高算法召回率,在一次匹配完成后,激活用户锚点对集合动态更新,重新计算全视角特征,实现迭代的用户识别.

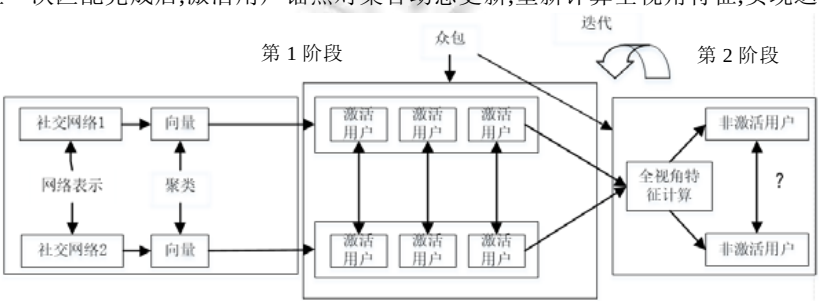

Fig.2 Two-Phase iterative user identification model based on crowdsourcing

图 2 结合众包的两阶段的迭代式社交网络用户识别模型

4 应用众包的激活用户锚点对构建

激活用户锚点对是对齐不同社交网络以及获取全视角特征的重要途径.一方面,激活用户锚点对相比于普通的用户锚点对,能够传递更多的属性和结构信息,有助于应用机器学习抽取特征;另一方面,通过激活用户,可以计算用户的全视角特征.并且利用不同社交网络间激活用户的映射关系将不同社交网络上用户的全视角特征对应起来,因而采用众包技术识别激活用户锚点对.首先,需要选取激活用户并按社区分类,然后在跨社网间使用众包来构建激活用户锚点对.

4.1 激活用户选择和分类

Barabasi 和 Albert[14]认为,在复杂网络上存在一种优先连接模型.该模型网络中的节点优先连接一些度比较高的节点,这就使得网络由一部分度数较高的核心节点和大部分的边缘节点构成,且节点度数满足幂等律.社交网络也满足这种网络特征,这些得到优先连接的用户即为激活用户,他们被其他用户包围着,构成局部社区,社区内部维系着共同话题或者兴趣爱好等.激活用户典型具有以下特征.

1) 用户位于社交网络的中心,与社区内部用户联系紧密,兴趣较为单一;

2) 处于激活状态,局部密度满足要求.

基于以上特征,本文采用近似的社区划分方法,思想是:在满足激活用户特征的条件下,选择社区的密度中心作为激活用户,并通过密度中心之间的合并过程来完成激活用户的聚类,忽略社区边缘一些度数较低的用户,以此提高激活用户选择和分类效率.具体步骤分两步实现:首先,采用 DeepWalk 图表示算法[15]将社交网络表示成向量,以此降低问题的维度;接下来,考虑到社交网络中社区的分布呈现出星网状的结构,社区内部联系紧密,而社区之间联系相对稀疏,也就是说,密度高的部分被众多密度低的部分包围着,此处采用 CFSFDP 聚类算法[16]初始计算用户的局部密度,并通过不断地合并和收缩来寻找聚类中心,通过对聚类决策图的观察,根据经验选取同时具有高δ和ρ值的点来划分社交网络,将用户聚类.将聚类结果中局部密度大于阈值β的作为激活用户,并按聚类结果划分到不同社区.通过对不同的密度阈值下实验结果进行评估来设定密度阈值β,通过实验部分证明,这种方法选取的关键点能够保证算法具有较好的冷启动效果.

4.2 结合众包的激活用户锚点对构建

通过观察,社交网络间社区分布相似,某个社交网络上的激活用户在其他网络上也能大概率地成为激活用户.若直接将这些不同网络间的激活用户交叉构成匹配对并生成众包任务,会带来巨大的金钱和时间花费.为此,本文提出一种启发式的候选激活用户锚点对构建策略,并结合众包识别锚点对.

在激活用户选取过程中,已经将多个社交网络上的激活用户按照不同社区划分,且通常同一社区的用户具有相同的兴趣.由于人的精力所限,兴趣往往较稳定.例如,若在社交网络 G_A 上,u_a 和 u_b 在同一社区中,在社交网络 G_B 中,对应的用户 u'_a 和 u'_b 也很大可能位于同一社区中.基于此,激活用户锚点对构建过程分为如下两个部分:首先,通过众包匹配不同社交网络间的社区;然后,在跨社网的匹配社区间,利用众包完成激活用户锚点对的构建.

由于用户是通过随机游走来向量表示的,密度最大的点说明其与社区外耦合较少、社区内联系紧密,兴趣相对稳定,且更能够代表社区的兴趣属性,因此,可基于桶思想优化匹配实现过程.例如,如图 3 所示,社交网络 G_A 和社交网络 G_B 分别具有 3 个社区(S_1,S_2,S_3),将激活用户按社区分别放入 3 个桶中,在桶中按照用户的局部密度进行排序,首先对跨网络间桶进行匹配,然后在桶间识别激活用户.在对桶进行匹配时,每次从尚未匹配的桶中取出密度最大的用户与其他网络的桶中的用户进行匹配,然而,该激活用户可能在另一网络中并没有注册账号或者对应账号并没有成为密度最大的用户,需要考虑从桶中取出密度次之的用户继续匹配.在实际匹配过程中,将待匹配的激活用户对发布到众包平台上,包括用户属性信息和用户的主页链接地址等,交由众包用户判断是否为同一用户.经过一轮后,如果有桶没有与其他桶匹配成功,选取密度次之的用户与其他桶匹配,进行以上操作,直到所有桶中均已匹配过为止.当不同社交网络上社区数量不等时,对其中最匹配的几对社区完成匹配,而忽略掉其他未能匹配的社区.这也是大多数基于子图分割的匹配算法的策略.从实现上,这种做法对于后续相似度的计算和匹配过程的实现并没有任何影响.从对算法的表现上看,确实会丢失一些细节信息,但在实际匹配过

程中,只是将社区的匹配结果作为不同网络间向量的对应关系.在每轮匹配完成后,根据新匹配用户对和已有社区匹配对比较,以更新社区匹配结果,从而尽可能地减少对信息的损失.

在众包平台中,常用的提交任务分为配对任务和基于簇的众包任务.如图 3 所示,为了降低众包代价以及提高效率,选择横向的生成簇的众包任务.以 S_1 匹配过程为例,众包工人需要判断 G_B 未匹配的各桶中当前密度最大的未匹配用户中是否有用户和 G_A 中 S_1 中该用户相似,将这样一次判断过程生成一条众包任务提交给平台.如果匹配不成功,取出每个桶中下一条用户记录,直到匹配成功或桶中所有记录均已提交.对于匹配成功的激活用户称为候选激活用户.

对于桶中用户匹配,纵向地生成簇任务,例如:若 G_A 中桶 1 和 G_B 中桶 2 匹配,G_A 中桶 1 中的每位用户与 G_B 桶 2 中所有用户均构成一条簇任务提交给平台.

通过以上步骤,我们利用众包完成了激活用户锚点对的构建.

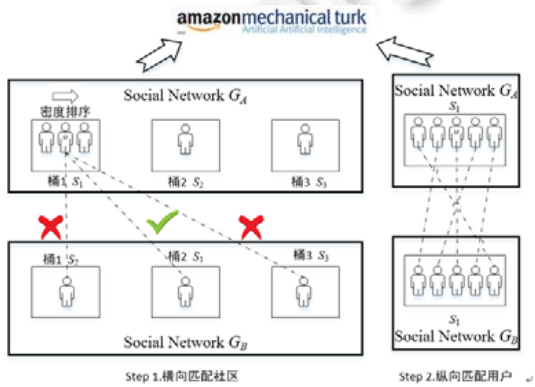

Fig.3 Active user's pairing process with crowdsourcing

图 3 结合众包的激活用户配对过程

5 基于全视角特征的非激活用户识别

通过构建用户的全视角特征向量,计算跨社交网络的用户相似度,并迭代匹配用户.

5.1 全视角特征提取和相似度计算

显然,只有那些构成激活用户锚点对的激活用户才能将全视角特征映射到不同社交网络上,因而这里只考虑这部分激活用户,称为激活用户锚点.直观上,用户和不同激活用户锚点之间的距离反映着自身的特征信息,将用户之间的关联关系看做路径长度,以经过路径的数量(由六度空间理论可知,不能达到或较长的用户路径长度为 6)作为全视角特征.然而,这种距离信息往往具有一些问题.图 4 中,11 号用户和两边激活用户锚点 4 号和 17 号的距离都一样,均为 1,但是从图 4 中可以很显然看出,11 号点离左边激活用户锚点更近一些.所以,简单把用户到激活用户锚点的距离作为位置信息的参照并不准确.

实际上,这种情况在社交网络中很普遍,用户可能会由于微博推荐或者热点新闻去关注一些热点用户.相比于其真实的兴趣,这些关注随着时间推移会演化.比如在奥运期间,大量用户关注奥运新闻的社区核心用户.需要区分这种短暂的兴趣和真正的爱好.下面以一个简单的情况为例来说明:对于真正的奥运迷,不会满足于仅通过奥运新闻来关注,而是关注一切和奥运领域有关系的核心用户.或者,作为一个明星的粉丝,好友列表就会有大量的粉丝团成员以及和明星相关的公众主页.因而在计算位置信息时,需要综合考虑邻居的位置信息.如公式(1),将用户和其直接邻居对各锚点的距离加权作为位置信息,即为全视角特征:

$$\bar{d}(A,u)=\frac{d(A,u)+\theta\times\sum_{j\in N(u)}d(A,j)}{1+\theta\times|N(u)|} \tag{1}$$

其中,u 为用户,A 代表激活用户锚点,θ表示权重,$N(u)$代表被用户 u 关注或者关注 u 的用户.

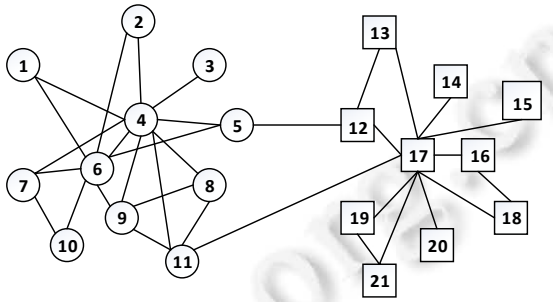

Fig.4 A social network structure diagram

图 4 某网络结构图

对于给定的社交网络上的用户 u,将与来自同一社区的位置信息进行算术平均,构成全视角特征向量,表示成 $\boldsymbol{d}(\boldsymbol{v}_u)$.例如,$a_1,a_2,a_3$ 为激活用户锚点,其中,a_1,a_2 来自同一社区,计算 u_1 特征向量时,将 $\bar{\boldsymbol{d}}(\boldsymbol{a}_1,\boldsymbol{u}_1)$ 和 $\bar{\boldsymbol{d}}(\boldsymbol{a}_2,\boldsymbol{u}_1)$ 进行算术平均.以 $((\bar{\boldsymbol{d}}(\boldsymbol{a}_1,\boldsymbol{u}_1)+\bar{\boldsymbol{d}}(\boldsymbol{a}_2,\boldsymbol{u}_1))/2,\bar{\boldsymbol{d}}(\boldsymbol{a}_3,\boldsymbol{u}_1))$ 作为用户的全视角特征向量值.在社交网络图上,采用弗洛伊德算法或者随机游走来计算距离.

对于不同网络间用户全视角特征向量的相似度(Sim_{oaf}),首先计算对应网络上的全视角特征向量,然后计算其余弦相似度,由公式(2)给出:

$$Sim_{oaf}(u_1,u_2)=\frac{d(v_{u_1})\times d(v_{u_2})}{|d(v_{u_1})||d(v_{u_2})|} \tag{2}$$

其中,$\boldsymbol{d}(\boldsymbol{v}_{u_1}),\boldsymbol{d}(\boldsymbol{v}_{u_2})$ 分别代表来自不同社交网络的用户 u_1,u_2 的全视角特征向量.

5.2 迭代匹配过程

依据本文提出的跨社网用户识别模型,迭代匹配过程具体如下.

步骤 1. 利用第 1 阶段中结合众包匹配的锚点对为已知用户匹配对,完成跨社交网络间的用户之间的匹配.本文通过综合用户属性特征、局部结构特征和用户全视角特征加权求和计算用户匹配相似度.这里采用元路径[7]提取局部结构特征.用户之间的相似度(Sim_{all})计算如下:

$$Sim_{all}(u_1,u_2)=\alpha\times Sim_{oaf}(u_1,u_2)+(1-\alpha)\times Sim_{attr\&structure}(u_1,u_2) \tag{3}$$

其中,u_1,u_2 为来自不同社交网络的用户,而 $Sim_{attr\&structure}(u_1,u_2)$表示用户的局部结构相似度和属性相似度,$Sim_{oaf}(u_1,u_2)$表示用户的全视角特征,$\alpha$表示不同相似度的权重.

步骤 2. 在社交网络间采用改进的稳定婚姻匹配方法[7]来识别匹配用户对,并保证用户对的一对一约束关系:对于在稳定婚姻匹配中相似度较低的用户对,需要通过众包进行人工判断.

步骤 3. 用户对识别完成后,若某一用户的密度大于激活用户锚点对集合中任意激活用户的密度,则加入集合中,并重新计算用户的全视角特征,进行下一次迭代,直到迭代中没有新的用户匹配对产生.迭代匹配计算过程见算法 1.

算法 1. 迭代匹配过程.

输入:要匹配的社交网络 $G_1=(U_1,G_1,A_1),G_2=(U_2,G_2,A_2)$,标志位 *flag*;

输出:识别出的用户匹配对集合 A.

1 分别找出 G_1,G_2 上的候选激活用户;

2 进行第 1 阶段众包用户识别,计算出激活用户锚点对,完成部分激活用户对的识别;

3 第 2 阶段建立机器学习模型,完成匹配;
4 WHILE *flag*=true /*初始值 *flag* 为 true*/
5 FOR EACH $u_1 \in U_1, u_2 \in U_2$ DO
6 计算全视角特征和属性、局部结构特征相似度;
7 END FOR
8 基于众包的稳定婚姻匹配策略匹配用户
9 将匹配结果放入集合 *A* 中;
10 IF 匹配工作并未完成 THEN
11 比较新匹配结果用户的密度和距离
对锚点对集合更新;
12 更新激活用户锚点对;
13 ELSE
14 *Flag*=false; /*更新标志位*/
15 END IF
16 END WHILE
17 RETURN *A*;

在第 1 行中,选择激活用户.在第 2 行,对激活用户进行众包用户识别.第 5 行~第 7 行,计算出用户的全视角特征,并利用众包得到的用户识别结果,构建机器学习训练模型,并计算用户相似度.第 8 行匹配用户对.第 10 行~第 14 行中,利用新匹配结果来对激活用户锚点对集合进行更新,以提高识别算法的召回率.可以看出,第 3 行~第 16 行构成了一个迭代的计算过程.

6 实验与结果

本节在真实的数据集上对本文算法进行了实验评估.

6.1 数据集

这里使用的是 Twitter-Flickr 数据集,Twitter 是一种常用的在线分享微博网络,而 Flickr 是一种以照片分享为主的社交网站.利用爬虫从两个社交网站中爬取用户,并且利用 Google Profiles service 提供的数据来构建事实集.

表 1 为数据集的基本信息:

Table 1 Statistics of Twitter-Flickr dataset

表 1 Twitter-Flickr 数据集统计信息

社交网络	用户数	用户关系数
Twitter	15 302	527 381
Flickr	12 749	407 824

6.2 对比方法和评估

本文提出的方法与 SVM、MNA[7]、COSNET[1]算法进行了对比实验.

- SVM:在匹配过程中仅考虑用户姓名、URL、出生地等属性的相似度.本文以该方法为基准方法;
- MNA:从社交网络中的用户关联关系、用户生成内容、时空等信息中抽取特征,并基于稳定婚姻匹配约束用户的映射关系;
- COSNET:基于局部结构和属性相似度构建候选匹配子图,通过建立最优能量模型来解决用户识别问题,把问题分割转化成对偶问题,提高了算法的效率;
- OCSA:本文提出的算法,结合众包基于全视角特征的跨社网迭代识别用户;

- OCSA-:结合众包不考虑全视角特征的跨社网迭代用户识别算法;
- OCSA_no:仅考虑全视角特征的跨社网迭代用户识别算法.

采用传统的准确率和召回率对实验结果评估.

6.3 实验和结果

在本节中,我们设置多个实验来验证本文方法的正确性和可靠性.

(1) 用户分布规律及聚类参数选择

在实验之前,首先统计了数据集中各社区网络上用户节点度数分布,如图 5 所示.可以看出:不论是 Twitter 还是 Flickr,用户分布都遵从幂律分布,激活用户仅占很少一部分,符合优先连接模型.其次,在图 6 中展示了利用 CFSFDP[15]算法对 Twitter 社交网络选取社区中心和进行社区划分的依据决策.其中,横轴代表ρ值,纵轴代表δ值.根据经验选取决策图中δ和ρ值较大的节点作为聚类中心,以便于对激活用户进行划分.

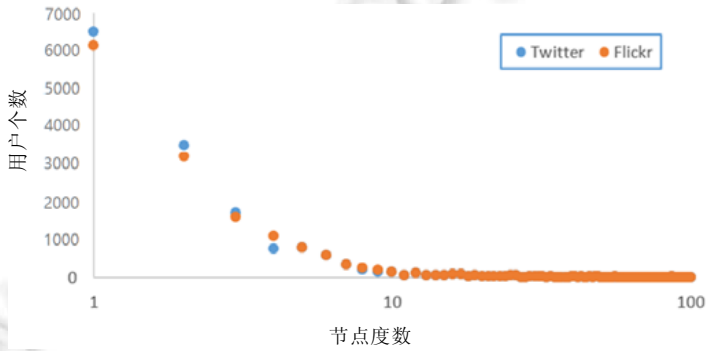

Fig.5 Degrees distribution of users

图 5 用户节点度数分布

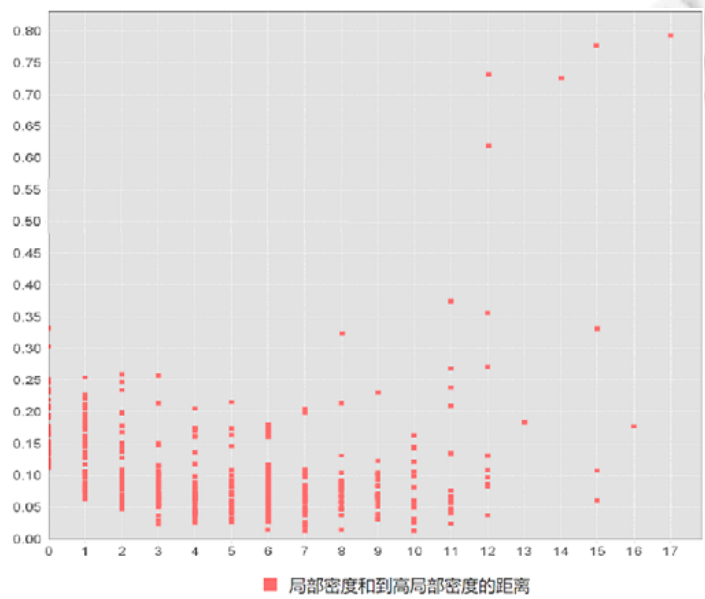

Fig.6 Clustering decision diagram

图 6 CFSFDP 聚类决策图

(2) 不同用户识别算法的准确率和召回率对比

图 7 显示了不同算法的识别结果,实验时,从事实集中抽取 1 000 条记录作为已知用户对,θ设为 1.3(见后文图 9),α设为 0.4(见后文图 10),密度阈值β设为 28(见后文图 11).从图 7 可以看出,本文提出的 OCSA 算法相比其他算法具有较高的准确率和召回率.通过 OCSA 算法和 OCSA-的对比可以看出:全视角特征可以识别出同名或属性较为相似的用户,提升了用户识别准确率,并在一定程度上避免了误识别伪造用户.

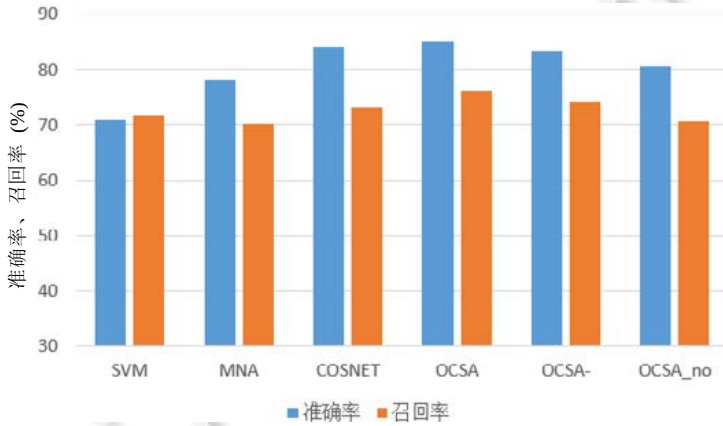

Fig.7 Perfomance of different methods

图 7 不同算法的准确率、召回率

(3) 已知匹配用户对数量对各算法的影响

为了验证已知匹配用户比例对于准确率和召回率的影响,从事实集中选取了 1 000 条记录,按照不同比例抽取作为已知匹配用户,采用上面相同的参数下,将 SVM、MNA 和本文的 OCSA 算法在不同比例下对比实验.如图 8 所示,在已知匹配用户比例不足的情况下,OCSA 算法能够明显地通过选取激活用户和全视角特征来提高准确率和召回率,并受已知匹配用户数量影响较小.由于 MNA 相比 SVM 能够提取更多的特征,并且能够通过一对一约束减少伪造用户对用户匹配的干扰,因而在已知匹配用户较少的情况下,也能获得更多的特征信息,准确率也较高,但也能看到召回率受已匹配用户数量影响较大.还观察到:在低匹配用户数量情况下,由于 SVM 不受一对一约束的限制,生成用户对较多,MNA 表现不如 SVM.

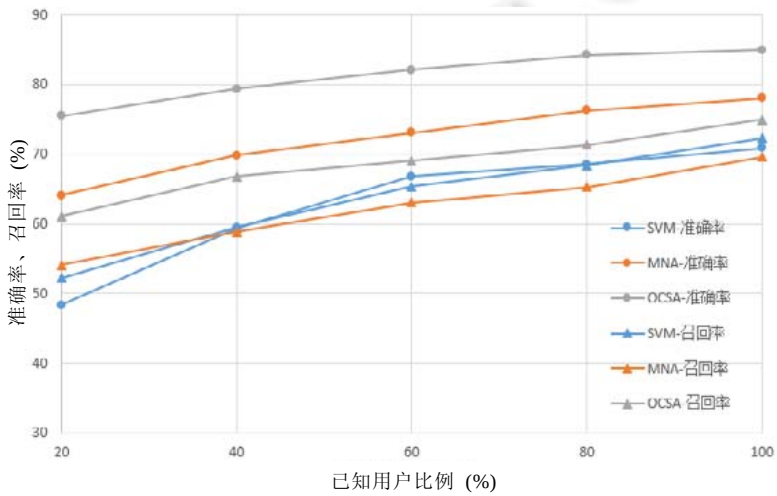

Fig.8 Impact of number of aligning users

图 8 匹配用户数量对准确率、召回率的影响

(4) θ,α,β对准确率、召回率的影响

图 9 表示用户和邻居位置关系权重θ对 OCSA 算法准确率和召回率的影响:一开始,随着θ的增大,准确率和召回率逐渐增加,但增加的幅度逐渐减小,一直达到稳定,最后略有下降.可以看出:综合考虑邻居的位置信息能够避免短期兴趣的干扰,但也需综合考虑用户自身的兴趣爱好选择.

图 10 显示了全视角特征和属性、局部结构特征的权重α对准确率和召回率的影响.随着α值的增加,准确率和召回率先增加后急速下降.可见:引入全视角特征可以分辨社交网络上存在的伪造用户,有助于识别用户,但其不能完全替代属性特征和局部结构特征的作用.

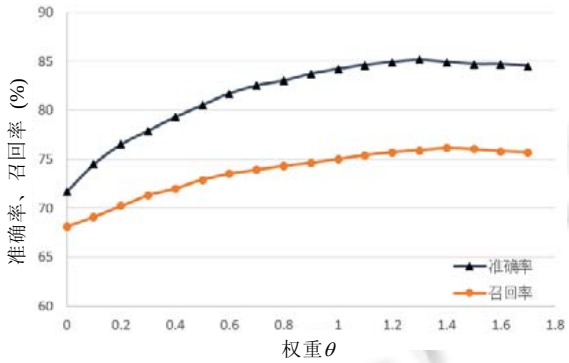

Fig.9 Impact of θ on precision and recall

图 9 θ对准确率、召回率的影响

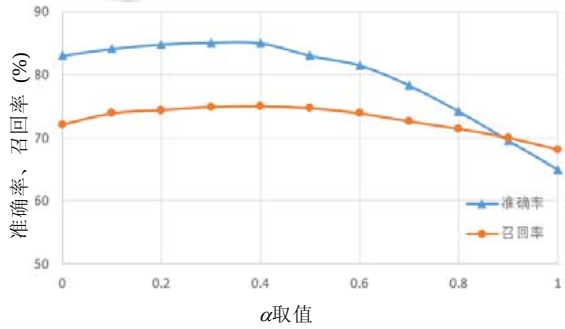

Fig.10 Impact of α on precision and recall

图 10 α对准确率、召回率的影响

图 11 显示了选取激活用户的密度阈值β对准确率和召回率的影响.如图 11 所示:当阈值越低时,选取的激活用户越多,通过第 1 阶段的众包识别可以显著提高识别的准确率,但激活用户的增多也影响了用户全视角特征的刻画,造成召回率的下降;反之,随着阈值的增加,准确率下降,而召回率上升.

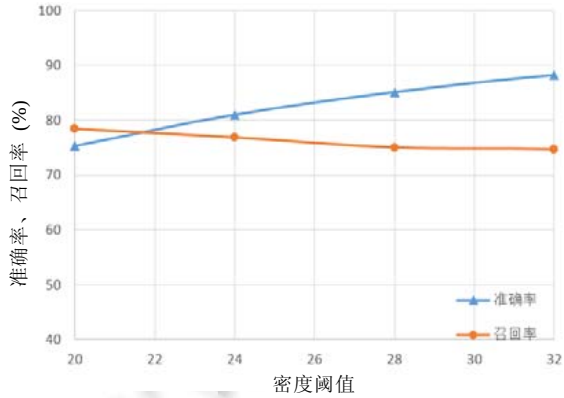

Fig.11 Impact of β on precision and recall

图 11 密度阈值对准确率、召回率的影响

(5) 激活用户锚点与匹配社区分布统计

表 2 统计了激活用户锚点是否位于匹配社区的数量分布,同时也统计了非激活用户锚点对的数量分布.可以看出:用户锚点往往同时存在匹配社区中,且大多数激活用户锚点对都能通过社区间的匹配得到.可见,本文提出结合众包的激活用户锚点对构建策略可以发现绝大多数激活用户锚点对.

Table 2 Statistics of anchor link

表 2 锚点对统计信息

锚点对种类	匹配社区	非匹配社区
激活用户锚点对	982	147
非激活用户锚点对	6 217	1 236

7 总 结

本文提出了结合众包的跨社交网络用户识别方法,通过匹配激活用户对提高已知匹配用户数量;提出了全视角特征的概念,精准描述用户画像;利用众包并进行迭代匹配,提高用户识别准确性.实验结果证明,该方法可以很好地解决已匹配用户过少以及误识别伪造用户的问题.同时,利用众包解决了传统机器学习算法中表达能力有限、冷启动等问题,从而提高识别的准确率和召回率.

同时也看到:众包需要等待人工的识别和处理过程,因而在处理效率上比不上传统的识别算法.本文在进行众包识别过程中采用了启发式的生成算法,优先对激活用户进行识别,但并没有考虑众包任务之间的传递依赖关系.在今后的工作中,希望能够对众包任务的生成过程进一步优化,利用已有的众包结果进行自动剪枝并批量生成众包任务,以减少迭代次数,满足对处理效率的要求.

References:

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

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

[3] Zhang J, Shao W, Wang S, Kong X, Yu PS. Partial network alignment with anchor meta path and truncated generic stable matching. Computer Science, 2015.

[4] Wang J, Kraska T, Franklin MJ, Feng J. CrowdER: Crowdsourcing entity resolution. Proc. of the VLDB Endowment, 2012,5(11): 1483−1494. [doi: 10.14778/2350229.2350263]

[5] Zafarani R, Liu H. Connecting corresponding identities across communities. In: Proc. of the Int'l Conf. on Weblogs and Social Media. Menlo Park: AAAI Press, 2009.

[6] Vosecky J, Hong D, Shen VY. User identification across multiple social networks. In: Proc. of the Int'l Conf. on Networked Digital Technologies. Piscataway: IEEE, 2009. 360−365. [doi: 10.1109/NDT.2009.5272173]

[7] Raad E, Chbeir R, Dipanda A. User profile matching in social networks. In: Proc. of the Int'l Conf. on Network-Based Information Systems. New York: IEEE Computer Society, 2010. 297−304. [doi: 10.1109/NBiS.2010.35]

[8] 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. New York: ACM Press, 2014. 51−62. [doi: 10.1145/2588555.2588559]

[9] Korula N, Lattanzi S. An efficient reconciliation algorithm for social networks. Proc. of the VLDB Endowment, 2014,7(5): 377−388. [doi: 10.14778/2732269.2732274]

[10] Vesdapunt N, Bellare K, Dalvi N. Crowdsourcing algorithms for entity resolution. Proc. of the VLDB Endowment, 2014,7(12): 1071−1082. [doi: 10.14778/2732977.2732982]

[11] Wang S, Xiao X, Lee CH. Crowd-Based deduplication: An adaptive approach. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2015. 1263−1277. [doi: 10.1145/2723372.2723739]

[12] Whang SE, Lofgren P, Garcia-Molina H. Question selection for crowd entity resolution. Proc. of the VLDB Endowment, 2014,6(6): 349−360. [doi: 10.14778/2536336.2536337]

[13] Gokhale C, Das S, Doan AH, Narasimhan JF, Rampalli N, Shavlim J, Zhu XJ. Corleone: Hands-Off crowdsourcing for entity matching. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2015. 601−612. [doi: 10.1145/2588555.2588576]

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

[15] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. In: Proc. of the 20th ACM SIGKDD. New York: ACM Press, 2014. 701−710. [doi: 10.1145/2623330.2623732]

[16] Rodriguez A, Laio A. Machine learning: Clustering by fast search and find of density peaks. Science, 2014,344(6191):1492−1496. [doi: 10.1126/science.1242072]

汪潜(1993－),男,安徽合肥人,硕士生,主要研究领域为社交网络.

寇月(1980－),女,博士,副教授,CCF 专业会员,主要研究领域为实体搜索,数据挖掘.

申德荣(1964－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为分布式数据管理,数据集成.

聂铁铮(1980－),男,博士,副教授,CCF 专业会员,主要研究领域为数据质量,数据集成.

冯朔(1989－),男,博士生,CCF 学生会员,主要研究领域为社交网络.

于戈(1962－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为数据库,分布式系统,嵌入式系统.