```
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 u1∈U1,u2∈U2 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.

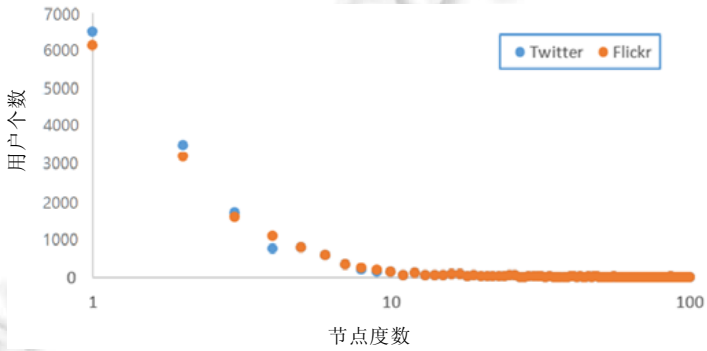

Fig.5 Degrees distribution of users

图 5 用户节点度数分布

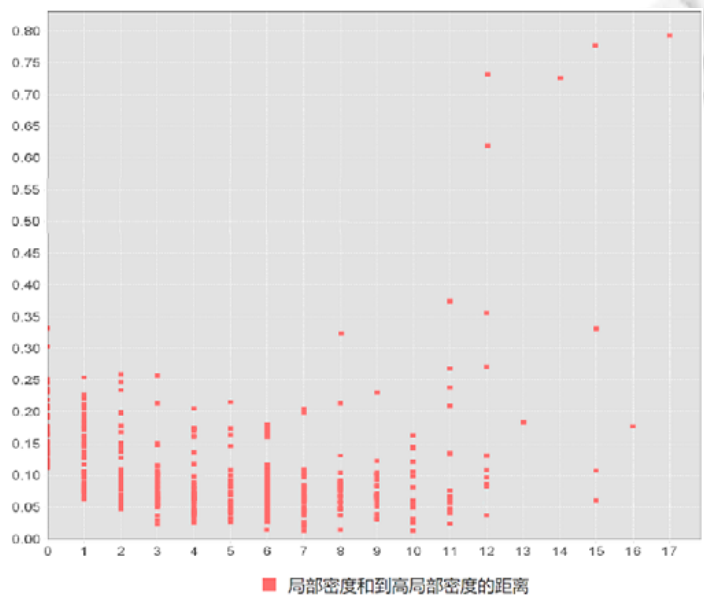

Fig.6 Clustering decision diagram

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.

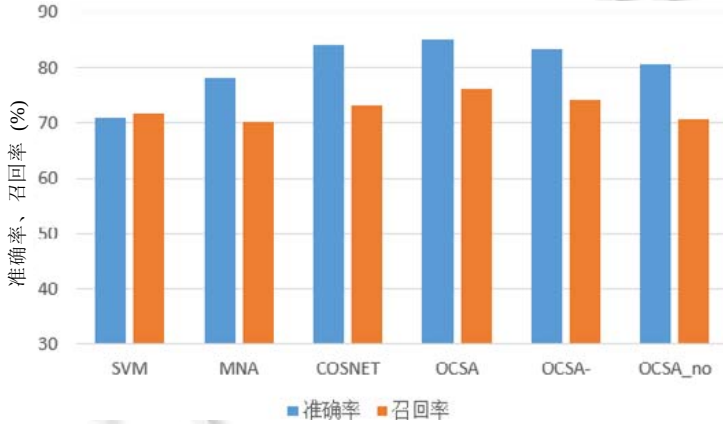

Fig.7 Perfomance of different methods

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.

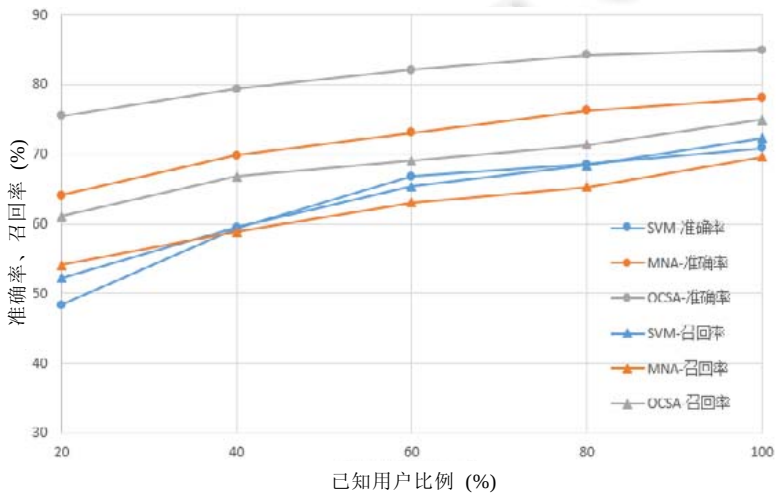

Fig.8 Impact of number of aligning users

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.

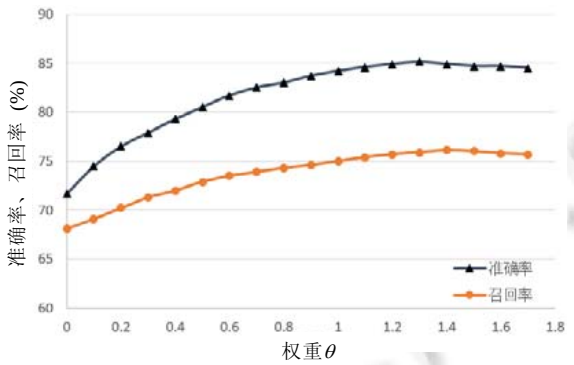

Fig.9 Impact of θ on precision and recall

图 9 θ对准确率、召回率的影响

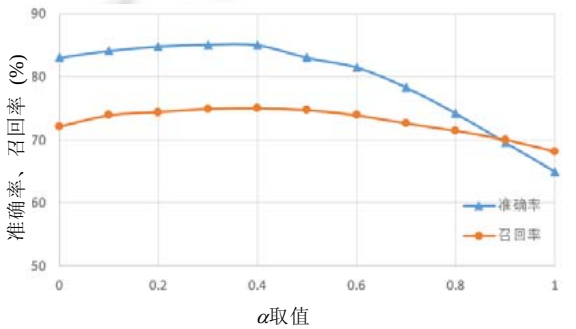

Fig.10 Impact of α on precision and recall

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.

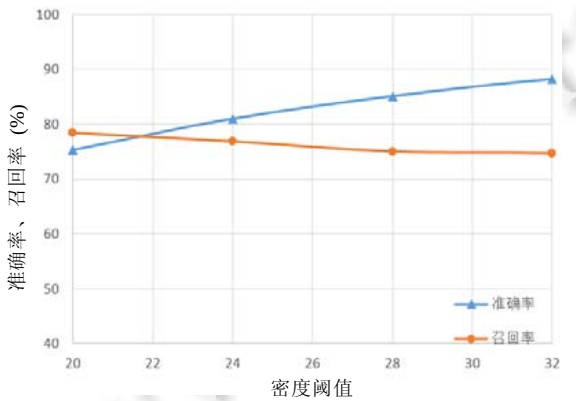

Fig.11 Impact of β on precision and recall

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.

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.

冯朔(1989－),男,博士生,CCF 学生会员,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.

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.

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.