软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(7):2175−2187 [doi: 10.13328/j.cnki.jos.005831]

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

一种基于最大公共子图的社交网络对齐方法*

冯 朔, 申德荣, 聂铁铮, 寇 月, 于 戈

(东北大学 计算机科学与工程学院,辽宁 沈阳 110819)

通讯作者: 冯朔, E-mail: fengshuo1989818@hotmail.com

摘 要: 随着 Internet 的普及,各类社交网络走进人们的视野,用户为满足不同的服务需求,往往不会局限于单一社交网络中,因此,跨社交网络环境下的用户识别问题成为研究者的热门话题.主要利用网络结构信息,针对社交网络对齐问题进行研究,主要包含以下研究点:首先,将网络对齐问题抽象为最大公共子图问题(α-MCS),并提出求解自适应参数α的方法,相比于传统的基于启发式定义参数α的方法,该方法可有效区分不同类型网络中匹配用户与非匹配用户;其次,为快速而准确地解决α-MCS,提出了基于最大公共子图的迭代式网络对齐算法 MCS_INA(α-MCS based iterative network alignment algorithm),该算法每次迭代过程主要包含两个阶段.第 1 个阶段,分别在两个社交网络中选取各自的候选匹配用户,第 2 个阶段,针对候选匹配用户进行识别.相比于其他算法,MCS_INA 时间代价低,且依据不同网络特征,通过参数估计,可保证较高的识别精度;最后,在真实数据集和合成数据集中验证了算法 MCS_INA 的有效性.

关键词: 社交网络;最大公共子图;用户识别;网络对齐

中图法分类号: TP311

中文引用格式: 冯朔,申德荣,聂铁铮,寇月,于戈.一种基于最大公共子图的社交网络对齐方法.软件学报,2019,30(7):2175−2187. http://www.jos.org.cn/1000-9825/5831.htm

英文引用格式: Feng S, Shen DR, Nie TZ, Kou Y, Yu G. Maximum common subgraph based social network alignment method. Ruan Jian Xue Bao/Journal of Software, 2019,30(7):2175−2187 (in Chinese). http://www.jos.org.cn/1000-9825/5831.htm

Maximum Common Subgraph Based Social Network Alignment Method

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

(School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China)

Abstract: With the popularization of Internet, plenty of social networks come into lives. To enjoy different services, users usually take part in multiple social networks simultaneously. Therefore, user identification across social networks has become a hot research topic. In this study, social network structure is used to solve the problem of network alignment. Firstly, the problem of network alignment is formalized as the problem of maximum common subgraph (α-MCS). A method is proposed to determine parameter α adaptively. Compared with the other heuristic methods on determiningα, the proposed method can distinguish matched users and unmatched users effectively on different kinds of social networks. Secondly, in order to fast answer α-MCS, algorithm MCS_INA (α-MCS based iterative network alignment algorithm) is proposed. MCS_INA mainly contains two steps in each iteration. In the first step, MCS_INA aims at selecting candidates in the two networks respectively. In the second step, a mapping algorithm is proposed to match candidates. Compared with other methods, MCS_INA has lower time complexity and higher identification accuracy on different networks. At last, experiments are conducted on real-world and synthetic datasets to demonstrate the effectiveness of the proposed algorithm MCS_INA.

Key words: social network; maximum common subgraph; user identification; network alignment

*基金项目: 国家重点基础研究发展计划(973)(2012CB316201); 国家自然科学基金(U1435216, 61672142, 61472070, 61602103); 国家重点研发计划(2018YFB1003404)

Foundation items: National Basic Research Program of China (973) (2012CB316201); National Natural Science Foundation of China (U1435216, 61672142, 61472070, 61602103); National Key R&D Program of China (2018YFB1003404)

收稿时间: 2018-07-20; 修改时间: 2018-11-22; 采用时间: 2019-01-26

随着 Internet 的广泛普及,各类社交网络走进人们的视野,不同社交网络为用户提供了不同的社交服务,例如,豆瓣为用户提供了图书、电影、音乐交流分享服务,知乎为用户提供了问答服务,微博为用户提供了分享动态的服务,用户为满足不同的服务需求,往往不会局限于单一社交网络中,而是在多个社交平台中注册账号,因此,社交网络对齐问题(通常也称为用户识别问题、用户匹配问题、网络反匿名化问题以及锚链接预测问题)逐渐引起了学者的关注.网络对齐将有效集成分散于各个网络中的用户资源,将大大提高用户推荐、广告投放、用户组形成等以用户为中心的服务质量.

针对网络对齐问题,在早期研究工作中,研究者们主要利用用户 E-mail[1]、用户真实姓名[2]等信息进行识别,虽然依据 E-mail 和真实姓名能够精确匹配用户,但在真实社交网络中,E-mail 和真实姓名由于隐私保护的原因,通常难以获取[3].因此,现阶段工作主要集中于利用:(1) 用户属性信息,如用户头像、用户喜好等[4,5];(2) 用户行为信息,如发帖关键字、用户行为轨迹等[6,7];(3) 用户结构信息,如用户朋友关系、用户与帖子的评论关系等[3,8,9].虽然现有方法具有良好的准确性,但真实社交网络通常面临用户数据匿名化严重、部分用户数据难以获取等问题,且现有公开数据也面临数据缺失、数据不一致、数据分布不均、数据异构等问题.

本文利用用户结构信息研究网络对齐问题,相比于用户属性信息与行为信息,用户结构信息具有易获取、易识别的特点,例如,人人网中用户的朋友关系、微博用户的关注与被关注关系以及大部分社交网络中用户之间的互动关系,均可作为标识用户的重要依据.但这并不意味着用户属性信息以及用户行为信息是无用的,在处理真实数据的过程中,本文可结合用户属性信息和用户行为信息,以取得更准确的用户识别效果.

网络对齐问题可抽象为最大公共子图问题(α-MCS)[10−12],如图 1 所示,G,G'表示两个不同的社交网络,节点表示用户,实线边表示图中用户之间的朋友关系,α-MCS 的目标是求取 G 到 G'的映射函数 F,使得 $Sco(F)$=#重叠边数量−α#非重叠边数量,取值最大,其中,α表示平衡重叠边与非重叠边数量的参数.例如,当α=0 时,对于 F_1={(1,b),(2,a),(3,f),(4,e),(5,d),(6,c)},$Sco(F_1)$=8,不存在其他映射关系 F_2 使得 $Sco(F_2)>Sco(F_1)$,因此,称由 F_1 形成的公共子图为 G,G'的最大公共子图.

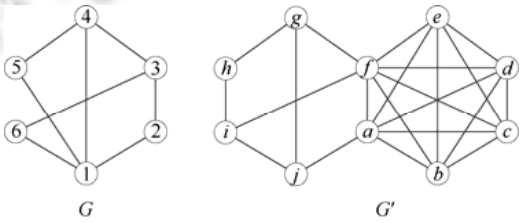

Fig.1 A pair of toy social networks for network alignment

图 1 社交网络对齐问题示意图

传统的最大公共子图问题在解决网络对齐问题时,存在以下几点不足:首先,其参数α无法自适应于不同类型网络,传统方法主要依赖于启发式的方法确定α,然而,由于不同网络结构的不同,使得该类方法具有一定的局限性;其次,传统方法计算复杂度较高,这类算法通常只能处理数据量较小的网络,随着社交网络规模的扩大,现有算法已不再适合处理大规模数据的社交网络;此外,现有方法的目标通常追求代价函数最小的匹配结果,而非社交网络用户的真实匹配关系,并没有结合社交网络结构特征有效地解决问题.因此,本文针对现有方法的不足,提出基于最大公共子图的社交网络对齐方法,主要有以下贡献点.

1) 本文利用最大公共子图问题(α-MCS)对网络对齐问题进行求解,并针对参数α的取值进行讨论,使其自适应于不同社交网络结构.

2) 为快速地解决α-MCS,本文提出基于最大公共子图的迭代式网络对齐算法 MCS_INA,主要分为两个阶段:第 1 个阶段,分别在两个社交网络中选取易于识别的候选匹配用户集合,第 2 个阶段,针对候选匹配用户集合进行识别.相比于传统方法,该方法结合社交网络特征,通过参数估计,快速定位候选集,降低了算法复杂度,该算法复杂度为 $O(DD'(D+D')(|V|+|V'|))$,远小于现有方法,其中,D,D'分别表示 G,G'的平均度数,V,V'表示 G,G'的用户

集合.

3) 为验证 MCS_INA 的有效性,本文选择在真实数据集和合成数据集上进行实验,实验结果表明,MCS_INA 在识别准确率与召回率上明显优于现有算法.

本文第 1 节简述相关工作.第 2 节简述相关预备知识,包括符号定义以及网络对齐模型.第 3 节针对α-MCS中参数α进行讨论,求取自适应的参数α.第 4 节提出算法 MCS_INA,以有效解决α-MCS.第 5 节设计并展示实验结果,分析算法的有效性.第 6 节对本文进行总结.

1 相关工作

现阶段网络对齐方法可分为 3 类:基于用户属性信息的网络对齐方法、基于用户行为信息的网络对齐方法、基于用户结构信息的网络对齐方法.

在网络对齐领域中,最传统的方法是利用用户真实姓名和 E-mail 进行用户识别,该类方法通过衡量字符串之间的转换规则以及相似性进行用户识别.然而,用户名和 E-mail 在具有较高识别准确性的同时,大大降低了召回率.因此,部分学者利用额外的用户属性信息进行网络对齐.虽然额外属性信息的加入提高了用户识别的召回率,但真实网络数据中用户属性信息往往具有隐私性、异构性,研究者获取到的数据往往是不全面的,甚至是错误的.

基于用户行为信息的对齐方法认为用户在不同社交网络中表现出相似的用户行为特征,例如用户写作习性、用户登录时间地点特征、用户主题偏好等.传统方法主要通过提取用户行为特征相似性进行用户识别,例如文献[13]针对用户写作特征,从用户词汇特征、语义特征、文章结构特征、文章主题特征进行特征抽取;文献[6]分别从“用户-时间”“用户-空间”“用户-关键字”3 个方面进行用户特征提取.通过计算用户行为特征相似性,大部分已有方法将网络对齐问题转化为二分类问题,并利用 SVM、逻辑回归、稳定婚姻匹配等分类方法进行识别.

基于用户结构信息的网络对齐问题,可抽象为最大公共子图问题.该问题最早在文献[14]中提出,作者认为重叠边数量是衡量最大公共子图问题的唯一条件,而文献[10]认为最大公共子图问题需要综合考虑重叠边数量与非重叠边数量.传统的最大公共子图问题均为 NP 完全问题,后续大量文献针对最大公共子图问题进行了近似求解,但现有近似求解方法的复杂度均大于 $O(n\log n)$,并不适用于真实网络环境.

为了解决真实网络环境下的网络对齐问题,文献[15]认为,少部分已知匹配用户可显著提升网络对齐的准确性,该文献通过部分节点迭代地识别其余用户;之后,文献[16]在有权重图上提出基于随机游走的用户相似性算法;文献[17]针对大规模网络上时间代价过高的问题,以牺牲部分召回率为代价,有效降低了时间复杂度,文献[3]有效解决了无已知匹配用户情况下的网络对齐问题.这些方法均利用网络结构特征,通过求取用户相似性,选择匹配用户.

此外,部分学者提出了基于网络表示的网络对齐算法.文献[18]将用户映射到多维空间中,他们认为不同网络中相同用户在该空间中距离相近;文献[19]对有向网络中的用户进行识别,其用户映射函数与用户自身属性、父母属性、孩子属性相关.此外,文献[20]认为,映射函数同样与用户所处社群相关.

然而,以上算法均为启发式算法,文献[8]首次提出了网络对齐模型,网络对齐模型是描述现实社交网络用户关系的数学模型,文献[8]中证明了其算法的正确性,并在现实网络对齐问题中取得了较高的准确率.之后,文献[9]在此基础上,通过增加匹配迭代次数,显著提升了准确率与召回率,文献[21]证明了其方法在无标度网络对齐模型中的正确性,文献[22]针对已知匹配用户过少的情况,以降低准确率为代价,提升了部分召回率.

本文方法与已有方法不同之处有二.首先,本文利用网络对齐模型,对最大公共子图问题中参数取值进行讨论,给出严谨的理论推导过程,相比于传统启发式确定参数的方法,本文方法可自适应于不同类型的网络;其次,本文在解决网络对齐的过程中,借鉴最大公共子图思想与网络对齐模型,结合社交网络特征,通过参数估计,快速定位候选集,有效降低了算法复杂度,本文算法的计算复杂度远小于现有方法.

2 预备知识

2.1 符号和定义

定义 1(网络). 给定网络 $G(V,E)$,其中,V 表示网络 G 中用户集合,E 表示网络 G 中用户关系集合,若用户 i 与用户 j 之间存在边,则表示 $V_{i,j}=V_{j,i}=1$,否则,$V_{i,j}=V_{j,i}=0$,对于网络中节点 i,本文使用 $\overline{i}$ 表示该节点 i 所对应的用户个体.

特别地,本文假设对于任意用户 $i\in V,j\in V$,有 $\overline{i}\neq\overline{j}$.

定义 2(对齐网络). 给定网络 $G(V,E)$和 $G'(V',E')$,若存在节点 $i\in V,i'\in V'$满足 $\overline{i}\neq\overline{i'}$,则称 G,G'为对齐网络.在对齐网络中,为区分不同网络,本文统一使用右上角标进行区分,如 $G(V,E)$与 $G'(V',E')$表示不同网络,$i\in V$ 与 $i'\in V'$ 表示不同网络中的用户.

定义 3(最大公共子图问题α-MCS). 给定对齐网络 $G(V,E)$和 $G'(V',E')$,以及初始映射函数 $F_0:V_0\to V_0'$,其中,$V_0\subseteq V,V_0'\subseteq V'$. α-MCS 的目标是找到从 V 到 V'的一一映射函数 $F:V\to V'$,满足:(1) 对于任意节点 $i\in V_0$,有 $F(i)=F_0(i)$;(2) 对于其他任意满足条件 1 的映射函数 F',满足 $Sco(F)>Sco(F')$,其中,

$$Sco(F)=\sum_{i\in V_F,j\in V_F}V_{i,j}V'_{F(i),F(j)}-\alpha\left|V_{i,j}-V'_{F(i),F(j)}\right| \tag{1}$$

其中,$V_F\subseteq V$ 表示节点集合 V 中通过映射 F 所匹配的节点集合,对于节点 i,j,若 $V_{i,j}V'_{F(i),F(j)}=1$,则称边 $V_{i,j}$ 为重叠边,若 $\left|V_{i,j}-V'_{F(i),F(j)}\right|=1$,则称边 $V_{i,j}$ 为非重叠边.由此可知,公式(1)为图 G 与 G'的重叠边与非重叠边数量之差,α 为平衡重叠边与非重叠边数量的参数.为方便表示,后文中称映射函数 F 为公共子图 F,称 $Sco(F)$为公共子图 F 的 Sco 得分.

例如,如图 1 所示,令α=1,初始映射函数 $F_0=\{(2,a),(3,f)\}$,则 G,G'的最大公共子图 F 为 $F=\{(1,j),(2,a),(3,f),(4,g),(6,i)\}$,其中,重叠边为(4,1),(4,3),(3,6),(3,2),(2,1),(6,1),无非重叠边,因此,公共子图 F 的得分为 $Sco(F)=6$,而对于其他任意公共子图,其 Sco 得分均不可能超过 6.因此,公共子图 F 为 G,G'的最大公共子图.

2.2 网络对齐模型

网络对齐模型[8]在跨社交网络分析的过程中具有重要意义,该模型认为现实中不同的社交网络均源起于相同的网络,即对于任意对齐网络 G 和 G',均源起于一个潜在网络 G^*,其中,G^*描述了用户之间的全部社交关系.依据此假设,网络对齐模型可描述为两个独立的过程:(1) G^*的模型化,本文假设 $G^*(V^*,E^*)$满足,对于任意 $i^*,j^*\in V^*,i^*$与 j^*之间存在边的概率为 p_{i^*,j^*},本文不针对 p_{i^*,j^*}附加任何约束条件,即网络 G^*可满足均匀分布(如 ER 模型[23]),也可满足幂律分布(如 Stochastic blockmodels[24]).(2) 真实网络 G,G'的产生:本文假设对齐网络 G 和 G'均为网络 G^*的子图:对于 G^*中任意一个节点,其有 S_V 的概率出现在网络 G 中,有 $S_{V'}$的概率出现在网络 G'中;而对于网络 G^*中任意一条边 $V_{i^*,j^*}=1$,若节点 i^*,j^*已存在于在 G 中,则该边存在于 G 中的概率为 S_E.同理,若节点 i^*,j^*已存在于在 G'中,则该边在 G'中存在的概率为 $S_{E'}$.

3 最大公共子图自适应参数分析

本节着重针对最大公共子图问题中参数α的取值进行分析,依据公式(1),公共子图 F 得分的期望值为

$$E[Sco(F)]=\sum_{i\in V_F,j\in V_F}\Pr(V_{i,j},V'_{F(i),F(j)})-\alpha\Pr(V_{i,j},\neg V'_{F(i),F(j)})-\alpha\Pr(\neg V_{i,j},V'_{F(i),F(j)}),$$

其中,$\Pr(V_{i,j},V'_{F(i),F(j)})$表示边 $V_{i,j}$ 与边 $V'_{F(i),F(j)}$均存在的联合概率,$\Pr(V_{i,j},\neg V'_{F(i),F(j)})$表示边 $V_{i,j}$ 存在且边 $V'_{F(i),F(j)}$不存在的联合概率.

定理 1. 给定对齐网络 $G(V,E)$和 $G'(V',E')$,令映射函数 $F:V\to V'$为 V 到 V'的一一映射,且满足:(1) 对于任意 $i\in V_F$,有 $\overline{F(i)}=\overline{i}$;(2) 对于任意 $i\notin V_F$,不存在 $i'\in V'$,使得 $\overline{i}=\overline{i'}$,则存在$\alpha$,使得对于其他任意映射关系 F',满足 $E[Sco(F)]>E[Sco(F')]$.

证明:依据定理 1 的描述,仅需证明存在α,使得以下两个条件成立:

① 对于任意 $i\in V,j\in V,i'\in V',j'\in V'$,若 $\bar{i}=\bar{i}',\bar{j}=\bar{j}'$ 或 $\bar{i}=\bar{j}',\bar{j}=\bar{i}'$, 则有

$$\Pr(V_{i,j},V'_{i',j'})-\alpha\Pr(V_{i,j},\neg V'_{i',j'})-\alpha\Pr(\neg V_{i,j},V'_{i',j'})\geqslant 0 \tag{2}$$

② 对于任意 $i\in V,j\in V,i'\in V',j'\in V'$,若不满足条件①,则有

$$\Pr(V_{i,j},V'_{i',j'})-\alpha\Pr(V_{i,j},\neg V'_{i',j'})-\alpha\Pr(\neg V_{i,j},V'_{i',j'})<0 \tag{3}$$

首先,对于条件①,有 $\Pr(V_{i,j},V'_{i',j'})=\Pr(V_{i,j})\Pr(V'_{i',j'}\mid V_{i,j})=\Pr(V_{i,j})S^2_{V'}S_{E'}$, 其中, $\Pr(V'_{i',j'}\mid V_{i,j})$ 表示边 $V'_{i',j'}$ 在边 $V_{i,j}$ 存在的情况下存在的概率,依据网络对齐模型,有 $\Pr(V'_{i',j'}\mid V_{i,j})=\Pr(V'_{i',j'}\mid V^*_{\bar{i},\bar{j}})=S^2_{V'}S_{E'}$, 其中, $V^*_{\bar{i},\bar{j}}$ 表示用户 i,j 在图 G^* 中存在边.同理, $\Pr(V_{i,j},V'_{i',j'})=\Pr(V'_{i',j'})S^2_VS_E$. 另外,对于 $\Pr(V_{i,j},\neg V'_{i',j'})$ 和 $\Pr(\neg V_{i,j},V'_{i',j'})$, 依据全概率公式,有 $\Pr(V_{i,j},\neg V'_{i',j'})=\Pr(V_{i,j})-\Pr(V_{i,j},V'_{i',j'}),\Pr(\neg V_{i,j},V'_{i',j'})=\Pr(V'_{i',j'})-\Pr(V_{i,j},V'_{i',j'})$, 因此公式(2)可化简为

$$\Pr(V_{i,j})\left(\left(\frac{1}{2}+\alpha\right)S^2_{V'}S_{E'}-\alpha\right)+\Pr(V'_{i',j'})\left(\left(\frac{1}{2}+\alpha\right)S^2_VS_E-\alpha\right)\geqslant 0.$$

令 $\left(\frac{1}{2}+\alpha\right)S^2_{V'}S_{E'}-\alpha\geqslant 0,\left(\frac{1}{2}+\alpha\right)S^2_VS_E-\alpha\geqslant 0$, 即 $\alpha\leqslant\frac{\mathrm{Min}(S^2_VS_E,S^2_{V'}S_{E'})}{2-2\mathrm{Min}(S^2_VS_E,S^2_{V'}S_{E'})}$, 可使得条件①满足.

其次,对于条件②,由于 i,j 和 i',j' 为不同节点对,有 $\Pr(V_{i,j},V'_{i',j'})=\Pr(V_{i,j})\Pr(V^*_{\bar{i},\bar{j}}\mid V^*_{\bar{i},\bar{j}})S^2_{V'}S_{E'}$. 同理, $\Pr(V_{i,j},V'_{i',j'})=\Pr(V'_{i',j'})\Pr(V^*_{\bar{i},\bar{j}}\mid V^*_{\bar{i},\bar{j}})S^2_VS_E$, 因此,公式(3)可化简为

$$\Pr(V_{i,j})\left(\left(\frac{1}{2}+\alpha\right)S^2_{V'}S_{E'}\Pr(V^*_{\bar{i},\bar{j}}\mid V^*_{\bar{i},\bar{j}})-\alpha\right)+\Pr(V'_{i',j'})\left(\left(\frac{1}{2}+\alpha\right)S^2_VS_E\Pr(V^*_{\bar{i},\bar{j}}\mid V^*_{\bar{i},\bar{j}})-\alpha\right)<0.$$

假设对于 G^* 中任意两条边 $V^*_{\bar{i},\bar{j}}$ 和 $V^*_{\bar{i}',\bar{j}'}$, 有 $\Pr(V^*_{\bar{i}',\bar{j}'}\mid V^*_{\bar{i},\bar{j}})\leqslant\lambda<1$, 则令 $\left(\frac{1}{2}+\alpha\right)S^2_{V'}S_{E'}\lambda-\alpha<0$, $\left(\frac{1}{2}+\alpha\right)S^2_VS_E\lambda-\alpha<0$, 即当 $\alpha>\frac{\mathrm{Min}(S^2_VS_E\lambda,S^2_{V'}S_{E'}\lambda)}{2-2\mathrm{Min}(S^2_VS_E\lambda,S^2_{V'}S_{E'}\lambda)}$ 时,可使得条件②满足.

综上,当 $\frac{\mathrm{Min}(S^2_VS_E\lambda,S^2_{V'}S_{E'}\lambda)}{2-2\mathrm{Min}(S^2_VS_E\lambda,S^2_{V'}S_{E'}\lambda)}<\alpha\leqslant\frac{\mathrm{Min}(S^2_VS_E,S^2_{V'}S_{E'})}{2-2\mathrm{Min}(S^2_VS_E,S^2_{V'}S_{E'})}$ 时,定理 1 成立. □

由定理 1 可知,通过选取合适的α,可以有效地区分网络中匹配用户与非匹配用户,为便于求解,本文取 $\alpha=\frac{\mathrm{Min}(S^2_VS_E,S^2_{V'}S_{E'})}{2-2\mathrm{Min}(S^2_VS_E,S^2_{V'}S_{E'})}$, 其中,参数 $S_V,S_E,S_{V'},S_{E'}$均与 G 和 G'的网络结构相关,本文将在第 4 节中详细介绍如何进行参数估计.

4 基于最大公共子图的迭代式网络对齐算法

在第 3 节中,本文讨论了最大公共子图问题中参数α的取值,本节将提出基于最大公共子图的迭代式网络对齐算法 MCS_INA(α-MCS based iterative network alignment algorithm),如算法 1 所示.

算法 1. *MCS_INA*(G,G',F_0).

输入:G,G',F_0;

输出:F.

1 $F\leftarrow F_0$

2 Do

3 $C\leftarrow$*Candidate*(G,F)

4 $C'\leftarrow$*Candidate*(G',F)

5 $M\leftarrow$*Match*(G,G',F,C')

6 $M'\leftarrow$*Match*(G,G',F,C)

7 $F\leftarrow M\cap M'$

8 While ($M\cap M'$ is not empty)

9 Output F

给定对齐网络 G 和 G'以及初始匹配用户关系 F_0,MCS_INA 的输出是包含 F_0 的用户对应关系 F.首先,算法利用初始匹配用户集合 F_0 初始化输出集合 F(第 1 行),之后,迭代地利用已匹配用户识别其他用户.在每次迭代过程中,主要分为两步:第 1 步,分别从 G 和 G'中选取识别度较高的候选匹配用户集合 C,C'(第 3 行~第 4 行),为降低计算复杂度,该步骤分别针对两个网络 G,G'单独进行处理,选取与已匹配用户关系较近的用户集合;第 2 步,分别针对候选匹配用户集合 C,C'进行用户匹配,借鉴最大公共子图思想,构建匹配用户映射关系 M:$V\to C'$和 M':$V'\to C$(第 5 行~第 6 行),并将 M 与 M'重叠的部分作为匹配结果(第 7 行),加入到输出集合 F 中,并执行下一次循环,若连续两次迭代匹配结果 F 未发生改变,则停止迭代,将 F 作为算法的输出.

本文在第 4.1 节中将深入讨论候选用户集合选取问题;在第 4.2 节中将深入讨论候选用户集合匹配问题.

4.1 候选匹配用户选取策略

为便于描述,本节仅针对 G 中候选匹配用户选取问题进行讨论.给定网络 $G(V,E)$以及已匹配用户映射关系 F,候选匹配用户选取算法 $Candidate(G,F)$的目标是,在 G 中选取与匹配用户集合 V_F 关系紧密的用户群体 C.结合最大公共子图思想,对于用户 k,构建其代价函数如下:

$$\Delta Sco_E(k)=\sum_{i\in V_{\bar{F}}}\Pr(V_{i,k},V'_{F(i),k'})-\alpha\Pr(V_{i,k},\neg V'_{F(i),k'})-\alpha\Pr(\neg V_{i,k},V'_{F(i),k'}) \tag{4}$$

其中,k'为与 k 匹配的用户;$\Delta Sco_E(k)$表示通过匹配用户 k,最终匹配结果 Sco 值提升幅度的期望.由于 k'未知,因此下面着重讨论如何估计$\Delta Sco_E(k)$的取值.为方便起见,本文假设 V_F 中用户以及 k'均匹配正确,即 $\bar{k}=\bar{k}'$, 且对于任意 $i\in V_F$,有 $\bar{i}=\overline{F(i)}$.

对于用户 $i\in V_F$,由于 $\bar{i}=\overline{F(i)}$ [0,1], $\bar{k}=\bar{k}'$, 有 $\Pr(V_{i,k},V'_{F(i),k'})=S_{E'}\Pr(V_{i,k})$,$\Pr(V_{i,k},\neg V'_{F(i),k'})=(1-S_{E'})\Pr(V_{i,k})$, $\Pr(\neg V_{i,k},V'_{F(i),k'})=\Pr(V^*_{\bar{i},\bar{k}})(1-S_E)S_{E'}$,其中,$\Pr(V^*_{\bar{i},\bar{k}})$ 表示 $\bar{i}$ 与 $\bar{k}$ 之间有边的概率,$\Pr(V^*_{\bar{i},\bar{k}})=\Pr(V_{i,k})/S_E$,因此,公式(4)可化简为

$$\Delta Sco_E(k)=\sum_{i\in V_{\bar{F}}}S_{E'}\Pr(V_{i,k})-\alpha(1-S_{E'})\Pr(V_{i,k})-\alpha\Pr(V_{i,k})(1-S_E)S_{E'}/S_E \tag{5}$$

公式(5)中,S_E(本文仅对 S_E 进行分析,$S_{E'}$同理)、$\Pr(V_{i,k})$、α均为未知参数,下面将对这些参数进行分析估计.

首先,S_E 表示网络对齐模型中网络 G^*中的边保留到 G 的概率,可通过以下公式进行估计:

$$S_E=\sum_{i\in V_F,j\in V_F}V_{i,j}V'_{F(i),F(j)}\Big/\sum_{i\in V_F,j\in V_F}V'_{F(i),F(j)} \tag{6}$$

其中,分子部分 $\sum_{i\in V_F,j\in V_F}V_{i,j}V'_{F(i),F(j)}=\sum_{i\in V_F,j\in V_F}\Pr(V^*_{\bar{i},\bar{j}})S_{E'}S_E$ 为 G 与 G'中重叠边数量,分母部分 $\sum_{i\in V_F,j\in V_F}V'_{F(i),F(j)}=\sum_{i\in V_F,j\in V_F}\Pr(V^*_{\bar{i},\bar{j}})S_{E'}$ 为 G'中已匹配用户间边的数量.

其次,$\Pr(V_{i,k})$表示 G 中用户 i 与 k 之间存在边的可能性,可通过 i,k 之间的间接关系进行预测,本文认为 $\Pr(V_{i,k})$与用户 i 与 k 的共同邻居数量相关.

$$\Pr(V_{i,k})=\sum_{p\in V,q\in V}\left(\delta(p,q,|N(i)\cap N(k)|)V_{p,q}\right)\Big/\sum_{p\in V,q\in V}\delta(p,q,|N(i)\cap N(k)|) \tag{7}$$

其中,$N(i)$表示用户 i 的邻居集合,$\delta(p,q,|N(i)\cap N(k)|)$表示用户对 p,q 的共同邻居数量是否等于$|N(i)\cap N(k)|$的判断.若相等,则$\delta(p,q,|N(i)\cap N(k)|)$取值为 1;否则,$\delta(p,q,|N(i)\cap N(k)|)$取值为 0.公式(7)的分子部分表示网络中公共邻居数量为$|N(i)\cap N(k)|$且存在边的节点对数量,分母为网络中公共邻居数量为$|N(i)\cap N(k)|$的节点对数量.显然,若网络中大部分公共邻居数量为$|N(i)\cap N(k)|$的节点对之间存在边,则 $\Pr(V_{i,k})$取值应该较高,否则,$\Pr(V_{i,k})$取值较低.

最后,对于α,依据定理 1,当 $\alpha=\mathrm{Min}(S_V^2S_E,S_{V'}^2S_{E'})\big/(2-2\mathrm{Min}(S_V^2S_E,S_{V'}^2S_{E'}))$ 时,可有效区分匹配用户与非匹配用户.然而,由于参数 S_V 与 $S_{V'}$无法预先确定,故无法准确估计α的取值.因此,在实验过程中,本文逐渐降低α取值,优先计算α较大时的匹配用户.本文令$\alpha=\beta\mathrm{Min}(S_E,S_{E'})/(2-2\beta\mathrm{Min}(S_E,S_{E'}))$,初始情况下,$\beta$=1,随着迭代的进行,逐渐降低$\beta$的取值.当$\beta=(|V_F|/\mathrm{Max}\{|V|,|V'|\})^2$ 时,$\alpha\leqslant\mathrm{Min}(S_V^2S_E,S_{V'}^2S_{E'})\big/(2-2\mathrm{Min}(S_V^2S_E,S_{V'}^2S_{E'}))$, 已小于最优取值,则迭代停止.由此,候选匹配选取算法 $Candidate(G,G',F)$如算法 2 所示.

算法 2. *Candidate*(G,G',F).

输入:G,G',F;

输出:C.

```
1  Compute S_E,S_E',β=1
2  Tmp←∅
3  For i∈V_F do
4     Tmp←N(i)
5  End For
6  For β∈[(|V_F|/Max{|V|,|V'|})^2,1]
7     Compute α
8     For i∈Tmp−V_F do
9        Compute ΔSco_E(i)
10       If ΔSco_E(i)>0
11          C.put(i)                    //C is ordered by ΔSco_E
12       End If
13    End For
14    If C is empty
15       β←β−0.1
16    End If
17 End For
18 Output C
```

在算法 2 中,首先,利用公式(6)对参数 S_E,$S_{E'}$进行估计,并赋值 β=1(第 1 行).之后,利用已识别用户集合 V_F,获取待分析的候选匹配用户集合 *Tmp*(第 2 行~第 5 行),从第 6 行开始,迭代地分析 *Tmp* 中用户是否适合作为候选匹配用户.若对于用户 $i\in Tmp-V_F$,其$\Delta Sco_E(i)>0$,则认为用户 i 适合作为候选匹配用户,并将其放入候选匹配用户集合 C 中(第 9 行~第 12 行),并输出结果集 C.若结果集 C 为空集,则有可能参数β取值过高,降低参数β,并重新计算候选匹配用户集合(第 15 行).在算法 2 中,通过迭代降低参数β,可有效提高算法识别精度,初始情况下,β取值为 1,相对应的α取值较高,候选匹配用户集合选取相对严格.而随着迭代的进行,β取值逐渐降低,进而α取值随之降低,候选匹配用户集合选取逐渐宽松,最终,当β取最低值时,若候选匹配用户集合依然为空,则无适合匹配的用户.

通过算法 2,可获取有序的候选匹配用户集合 C,集合 C 中用户依据与已匹配用户之间的关系强度进行排序,与已匹配用户关系紧密的用户具有较高排名,相反地,关系疏远的用户具有较低排名.另外,对于每个候选匹配用户,计算其代价函数的时间复杂度为$O(D^2)$.因此,在 MCS_INA 中,候选匹配选取算法 *Candidate*(G,G',F)的时间复杂度为 $O(D^2|V|+D'^2|V'|)$.

4.2 用户匹配策略

给定 G 中候选匹配用户集合 C,用户匹配算法 *Match*(G,G',F,C)的目标是,构建映射函数 $M':V'\to C$.由于最大公共子图问题为 NP 完全问题,为降低计算复杂度,本节利用第 4.1 节中候选匹配用户排名,结合贪婪思想,提出近似求解算法.

对于候选匹配用户集合 C 中任意用户 $k\in C$,令 k'为 G'中未匹配用户,借鉴最大公共子图思想(见公式(1)),则 k'与 k 的匹配度可表示为

$$\Delta Sco(k,k')=\sum_{i\in V_F}V_{i,k}V'_{F(i),k'}-\alpha\left|V_{i,k}-V'_{F(i),k'}\right| \tag{8}$$

公式(8)表示,若匹配用户 k 与 k',可提升匹配结果 *Sco* 得分$\Delta Sco(k,k')$.

至此,用户匹配算法 *Match*(*G*,*G*′,*F*,*C*)如算法 3 所示.

算法 3. *Match*(*G*,*G*′,*F*,*C*).

输入:*G*,*G*′,*F*,*C*;

输出:*M*.

1 $M\leftarrow\varnothing$
2 For $k\in C$ //*C* is ordered by ΔSco_E
3 $Tmp\leftarrow\varnothing$, $k'\leftarrow$null
4 For $i\in N(k)\cap V_F$
5 $Tmp\leftarrow N(F(i))$
6 End For
7 For $t'\in Tmp-V'_F$
8 If $\Delta Sco(k,t')>\Delta Sco(k,k')$ & $\Delta Sco(k,t')>0$
9 $k'\leftarrow t'$
10 End If
11 End For
12 $M\leftarrow(k,k')$
13 End For
14 Output *M*

算法 3 中采用贪婪思想有序地对用户集合 *C* 中用户进行识别,优先识别与已匹配用户关系紧密的用户,可有效降低识别错误的发生.

在算法 3 中,对于每个候选匹配用户 *k*,其对应的 *Tmp* 集合中用户的个数为 $O(DD')$,而对于 *Tmp* 中每个用户 *t*′,计算 *k* 与 *t*′匹配度的时间复杂度为 $O(D+D')$,因此,识别每个候选匹配用户 *k* 的时间复杂度为 $O(DD'(D+D'))$,且在 MCS_INA 中,用户匹配算法 *Match*(*G*,*G*′,*F*,*C*)的时间复杂度为 $O(DD'(D+D')(|V|+|V'|))$.

综上,算法 MCS_INA 的时间复杂度为 $O(DD'(D+D')(|V|+|V'|))$.

5 实 验

5.1 实验环境

实验环境:本文采用 Java 编程语言实现相关算法,实验主机采用 Intel i5-4590 处理器,主频 3.30GHz,8GB 内存,操作系统为 64 位 Windows 7.

数据集:本文所使用数据集见表 1.首先,Facebook 表示匿名化的真实 Facebook 数据集,其中,第 1 个网络(FL)为 Facebook 新奥尔良市用户关系网络,另一个网络(FW)为 Facebook 新奥尔良市用户信息墙交互网络,其中,重叠的用户数量为 25 538,重叠的边的数量为 151 580,该数据集可参考文献[25];其次,本文利用真实社交网络生成部分数据集,其中,Twitter 表示真实 Twitter 用户数据集,$T_{1.0}$ 表示原始的 Twitter 数据规模,$T_{0.8}$ 表示 $T_{1.0}$ 中 80%的边以及 80%的节点被保留到数据集 $T_{0.8}$ 中,$T_{0.7}$ 和 $T_{0.9}$ 同理.在生成 $T_{0.7}$、$T_{0.8}$ 和 $T_{0.9}$ 时,均采用随机概率保留点和边.另外,本文采用不同随机图生成算法生成合成数据集 ER 和 PA,其中,ER 表示该网络分布满足 ER 随机图模型[23],PA 表示该网络节点关系分布满足幂律分布[26],所有随机图均通过 igraph 生成.最后,本文随机地选取匹配用户作为已知匹配用户,该方式将有效降低识别瓶颈的发生,若已知匹配用户集中于单一社区内,将造成社区外部节点识别准确率的下降.

对比算法:由于启发式的解决方法适用性较低,与本文研究内容有差异;基于网络表示的网络对齐算法,需要大量训练数据,而本文方法仅基于预先匹配的少量用户节点数量(占比通常为 10%以下),两种方法环境不同.为此,本文仅选取与本文研究方法密切相关的两种经典算法 CN 和 CNR 进行比较:(1) CN[8]:CN 算法为迭代算法,每次迭代过程中,选取共同邻居数量最多的用户对作为匹配用户;(2) CNR[9]:与 CN 算法类似,但 CNR 算法在

每次迭代过程中优先匹配度数较高的用户;(3) MCS_INA:本文提出的算法.

效果评价:本文采用准确率(precision)、召回率(recall)、*F*-measure 以及运行时间(runtime)这 4 个方面进行评估.

Table 1 Datasets

表 1 数据集

数据集		节点数	边数
Facebook	FL (FacebookLinks)	49 247	797 470
	FW (FacebookWall)	25 622	157 236
Twitter	$T_{1.0}$	11 473	191 626
	$T_{0.9}$	10 331	138 970
	$T_{0.8}$	9 149	97 360
	$T_{0.7}$	8 008	66 578
ER		20 000	400 000
PA		20 000	399 791

5.2 真实数据集中的实验效果

首先,为比较不同算法在不同数据集中的识别准确率,该组实验采用 FL&FW、$T_{0.7}$&$T_{0.8}$、$T_{0.7}$&$T_{0.9}$ 和 $T_{0.8}$&$T_{0.9}$ 这 4 个数据集,对于每组数据集,随机地选取 10%的匹配用户作为已知,实验结果如图 2 所示.在 3 种不同算法中,本文算法 MCS_INA 的准确率最高,而 CN 算法准确率最低.另外,对比 Twitter 的 3 组数据集,3 种算法在数据集 $T_{0.8}$&$T_{0.9}$ 上具有较高准确率,而在 $T_{0.7}$&$T_{0.8}$ 数据集上具有较低准确率.原因是,$T_{0.8}$&$T_{0.9}$ 重叠用户数量以及重叠边较多,期望情况下其重叠边为 $T_{1.0}$ 边数量的 37%,而 $T_{0.7}$&$T_{0.8}$ 的重叠边比率仅为 17%.因此,$T_{0.8}$&$T_{0.9}$ 相对更容易识别.

其次,对比不同算法在不同数据集中的召回率,如图 3 所示.在 3 种算法中,本文算法 MCS_INA 的召回率依然最高,其次为 CNR,算法 CN 召回率最低;对比图 2 中的准确率,算法 CNR 在数据集 $T_{0.7}$&$T_{0.8}$、$T_{0.7}$&$T_{0.9}$、$T_{0.8}$&$T_{0.9}$ 中的召回率略高于准确率,这是由于 Twitter 数据集中包含大量单独存在于单个网络中的用户,算法 CNR 错误地识别了这一部分用户.而在数据集 FL&FW 中,算法 CNR 的准确率略高于召回率,这是由于 FW 数据集几乎为 FL 数据集的子集.

最后,综合准确率与召回率,*F*-measure 的比较结果如图 4 所示,MCS_INA 的综合性能明显优于算法 CN 和 CNR.

为测试不同算法在不同数据集中的运行时间,本组实验记录了不同算法的运行时间,见表 2.由表 2 可知,算法 CN 的运行时间最短,其次为 MCS_INA,CNR 的运行时间最长.虽然算法 CN 具有最短的运行时间,但综合图 4 中 *F*-measure 的比较结果来看,MCS_INA 依然具有最高的综合性能.另外,对于算法 CNR,无论从算法执行时间还是算法精度,MCS_INA 均优于 CNR.

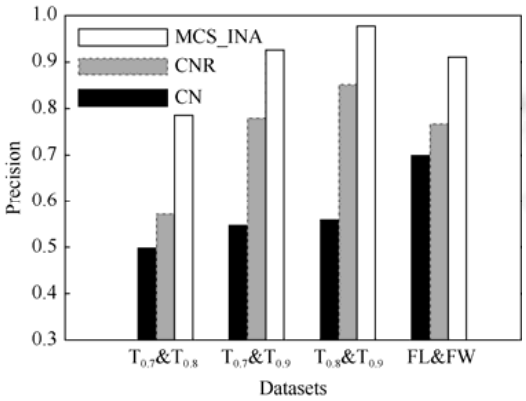

Fig.2 Precision on real-world datasets

图 2 真实数据集中的准确率比较结果

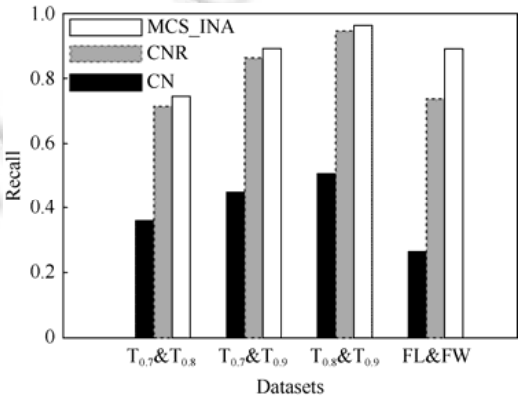

Fig.3 Recall on real-world datasets

图 3 真实数据集中的召回率比较结果

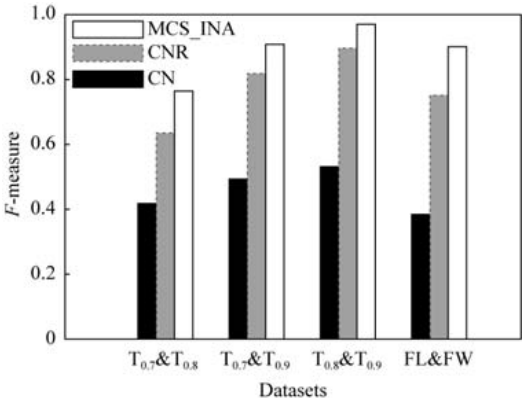

Fig.4 *F*-measure on real-world datasets

图 4 真实数据集中的 *F*-measure 比较结果

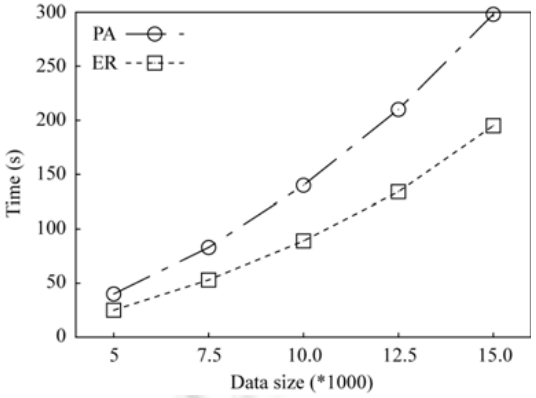

Fig.5 Running time regarding the nodes

图 5 运行时间随节点数变化实验图

Table 2 Runnin time on real-world datasets (min)

表 2 真实数据集中运行时间比较结果 (分钟)

	$T_{0.7}$&$T_{0.8}$	$T_{0.7}$&$T_{0.9}$	$T_{0.8}$&$T_{0.9}$	FL&FW
CN	0.53	0.58	0.65	2.43
CNR	1.32	1.95	2.27	10.21
MCS_INA	0.95	1.22	1.43	7.55

5.3 合成数据集中的实验效果

在第 5.2 节中,本文针对真实数据集进行了实验,虽然在真实数据集中算法 MCS_INA 具有较优性能,但并不代表在所有数据集中算法 MCS_INA 均表现优异,为此,在第 5.3 节中,本文利用不同类型的合成数据集,测试算法的性能.

1) MCS_INA 在不同类型网络中的性能实验

为验证算法 MCS_INA 在不同类型数据集中的表现,本节分别在 ER 数据集与 PA 数据集中测试 MCS_INA 算法的性能,见表 3 和表 4.以 ER 数据集为例,数据集 $ER_{0.5}$、$ER_{0.6}$、$ER_{0.7}$、$ER_{0.8}$、$ER_{0.9}$ 分别表示从数据集 ER 中以概率[0.5,0.6,0.7,0.8,0.9]提取点和边,对于每组数据集,本实验选取 10%的用户作为已知匹配用户.由表 3 和表 4 可知,当数据集重叠部分较大时($ER_{0.6}$&$ER_{0.7}$、$ER_{0.7}$&$ER_{0.8}$、$ER_{0.8}$&$ER_{0.9}$、$PA_{0.7}$&$PA_{0.8}$、$PA_{0.8}$&$PA_{0.9}$),MCS_INA 具有较高的准确率与召回率,而当数据集重叠部分较小时($ER_{0.5}$&$ER_{0.6}$、$PA_{0.5}$&$PA_{0.6}$、$PA_{0.6}$&$PA_{0.7}$),MCS_INA 具有较低的准确率与召回率,其原因是,当数据集重叠部分较小时,非匹配用户之间相似性相对较强,从而错误地匹配非匹配用户,降低了准确率与召回率.另外,对比表 3 与表 4,MCS_INA 在 ER 数据集中的表现明显强于 PA 数据集,其原因是,ER 数据集中用户之间相似程度较低,而 PA 数据集中,尤其是度数较低用户之间,相似程度较高,当删除部分用户以后,MCS_INA 错误地将这些相似度较高的用户进行匹配,从而降低了准确率与召回率.

Table 3 Performance of MCS_INA on synthetic ER datasets

表 3 MCS_INA 在合成 ER 数据集中的运行结果

	$ER_{0.5}$&$ER_{0.6}$	$ER_{0.6}$&$ER_{0.7}$	$ER_{0.7}$&$ER_{0.8}$	$ER_{0.8}$&$ER_{0.9}$
Precision	0.49	0.98	1.0	1.0
Recall	0.38	0.96	0.99	1.0
F-measure	0.42	0.97	0.99	1.0

Table 4 Performance of MCS_INA on synthetic PA datasets

表 4 MCS_INA 在合成 PA 数据集中的运行结果

	$PA_{0.5}$&$PA_{0.6}$	$PA_{0.6}$&$PA_{0.7}$	$PA_{0.7}$&$PA_{0.8}$	$PA_{0.8}$&$PA_{0.9}$
Precision	0.27	0.81	0.95	0.98
Recall	0.35	0.87	0.96	0.99
F-measure	0.30	0.83	0.95	0.98

2) MCS_INA 运行时间随网络规模变化的实验

为测试 MCS_INA 运行时间随网络规模的变化趋势,本实验利用 ER 与 PA 数据集进行实验.首先,固定合成网络平均度数为 15,变化网络中节点数量,生成不同的原始网络.之后,采用参数 S_E=$S_{E'}$=S_V=$S_{V'}$=0.8,生成对齐网络,实验结果如图 5 所示.横轴表示生成网络中节点数量,纵轴表示算法运行时间,可知,随着网络中节点数量的增多,MCS_INA 算法的运行时间随网络中节点数量的增加基本呈线性增长.然后,固定合成网络节点数量为 5 000,变化网络中平均节点度数,生成不同的原始网络.之后,采用参数 S_E=$S_{E'}$=S_V=$S_{V'}$=0.8,生成对齐网络,实验结果如图 6 所示.通过实验可知,MCS_INA 算法的运行时间随网络中节点度数的增加呈指数型增长,且 MCS_INA 处理 ER 数据集的能力要高于处理 PA 数据集的能力.

3) MCS_INA 性能随已知匹配用户数量变化的实验

本实验数据集采用 $ER_{0.8}$&$ER_{0.8}$,即依据 ER 数据集,采用参数 S_E=$S_{E'}$=S_V=$S_{V'}$=0.8 生成两组不同 $ER_{0.8}$ 并对其进行匹配,本实验随机抽取不同数量百分比的用户作为已知匹配用户对,实验结果如图 7 所示.由图 7 可知,随着已知匹配用户的减少,实验准确率与召回率逐渐降低,当已知匹配用户数量减少至 0.3%时,准确率与召回率实现断崖式降低.这是由于,当已知匹配用户数量降低至 0.3%时,这些已知匹配用户之间几乎不存在直接关系,从而使得 MCS_INA 的准确率与召回率基本降至 0.

4) 参数分析

首先,对自适应参数α进行实验分析,如图 8 所示,1-MCS_INA 表示在每次迭代中不对参数α进行估计,并设定α为 1,同理于 0.5-MCS_INA.该实验分别在 3 个不同数据集 $ER_{0.7}$&$ER_{0.7}$、$ER_{0.8}$&$ER_{0.8}$、$ER_{0.9}$&$ER_{0.9}$ 中运行 MCS_INA、1-MCS_INA 和 0.5-MCS_INA.由图 8 可知,通过自适应调节参数α,在 3 个不同数据集中均取得最优性能.另外,0.5-MCS_INA 的表现优于 1-MCS_INA,其原因是,对于 1-MCS_INA,其节点对相似性函数(见公式(8))中参数α过大,很多匹配用户无法达到阈值,使得召回率降低.

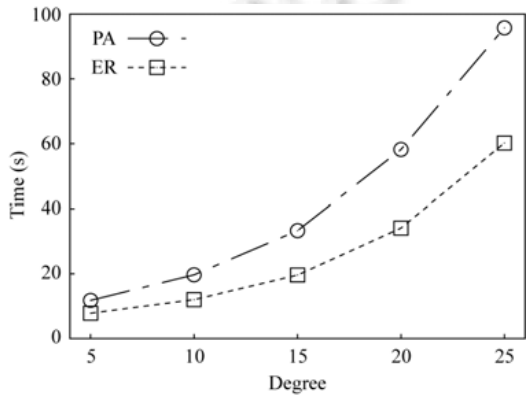

Fig.6 Running time regarding node degree

图 6 运行时间随节点度数变化实验图

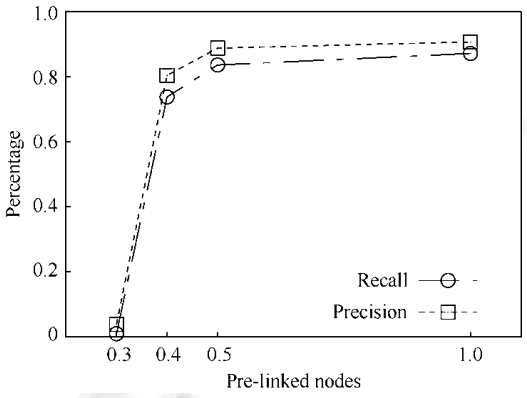

Fig.7 Performance regarding pre-linked nodes

图 7 性能随已知匹配用户变化实验图

其次,针对每次迭代过程中参数α、S_E 和 $S_{E'}$的估计准确性进行分析,本实验采用数据集 $ER_{0.8}$&$ER_{0.8}$,并记录每次迭代过程中 3 个参数值的大小,如图 9 所示.对于参数 S_E 和 $S_{E'}$,其取值随迭代过程逐渐降低,并维持在 0.8 左右.对于参数α,其波动范围较大,在前几次迭代过程中,参数α的取值范围较大,优先对识别度较高的用户进行识别,之后,参数α的取值随迭代过程逐渐降低,并最终稳定在 0.4 左右.通过理论计算参数α可知,当参数α理论取值 0.52 时最优(通过定理 1 可知),之所以会导致实际参数取值与理论取值不一致的情况,是因为在实际情况下,通常有少部分匹配用户,其结构相似性较低,需适量降低参数α的取值.

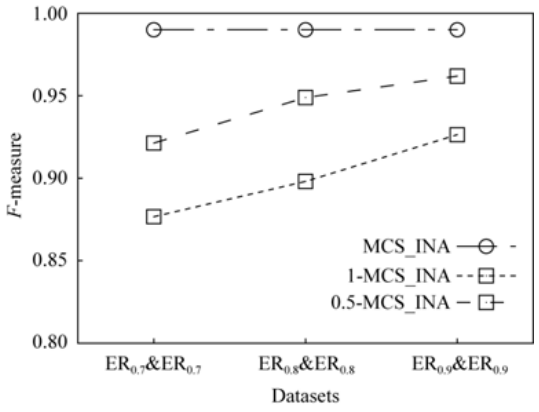

Fig.8 Performance regarding different α

图 8 不同参数α对识别准确性的影响

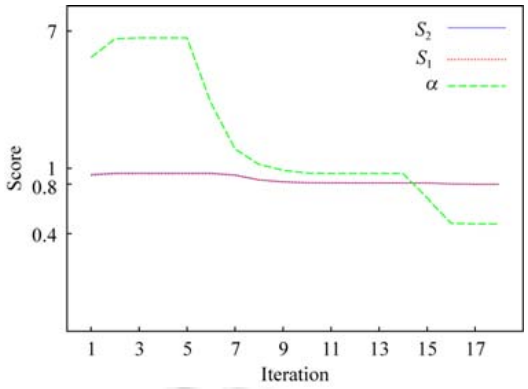

Fig.9 Estimation of the parameters

图 9 实验参数估计

6 结束语

本文主要针对基于用户结构信息的跨社交网络用户识别问题进行研究.首先,借鉴传统最大公共子图问题,提出了求解自适应参数的方法,使得最大公共子图问题可适用于求解不同类型的网络对齐问题;其次,针对最大公共子图计算复杂度过高的问题,本文提出了基于最大公共子图的迭代式网络对齐算法MCS_INA,相比于传统算法,MCS_INA 在每次迭代过程中,仅针对部分候选匹配用户进行匹配,且本文所提出的候选匹配算法有效结合了网络对齐模型,具有严格的理论支持;最后,本文在真实数据集和合成数据集上进行了实验,实验结果表明本文所提出算法具有较高的识别准确率与较低的时间代价.在未来的工作中,将着重针对初始匹配用户过于集中的问题,同时结合用户属性信息、用户行为信息以处理跨网络用户识别问题.

References:

[1] Ding L, Zhou L, Finin T, *et al*. How the semantic Web is being used: An analysis of FOAF documents. In: Proc. of the 38th Annual Hawaii Int'l Conf. on System Sciences. IEEE, 2005. 113c.

[2] Mika P. Flink: Semantic Web technology for the extraction and analysis of social networks. Web Semantics: Science, Services and Agents on the World Wide Web, 2005,3(2-3):211–223.

[3] Zhou X, Liang X, Du X, *et al*. Structure based user identification across social networks. IEEE Trans. on Knowledge and Data Engineering, 2018,30(6):1178–1191.

[4] Raad E, Chbeir R, Dipanda A. User profile matching in social networks. In: Proc. of the 13th Int'l Conf. on Network-Based Information Systems (NBiS). IEEE, 2010. 297–304.

[5] Malhotra A, Totti L, Meira Jr W, *et al*. Studying user footprints in different online social networks. In: Proc. of the 2012 Int'l Conf. on Advances in Social Networks Analysis and Mining (ASONAM 2012). IEEE Computer Society, 2012. 1065–1070.

[6] 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 & Knowledge Management. ACM, 2013. 179–188.

[7] Liu S, Wang S, Zhu F, *et al*. 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, 2014. 51–62.

[8] Yartseva L, Grossglauser M. On the performance of percolation graph matching. In: Proc. of the 1st ACM Conf. on Online Social Networks. ACM, 2013. 119–130.

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

[10] Feizi S, Quon G, Recamonde-Mendoza M, *et al*. Spectral alignment of graphs. arXiv Preprint arXiv: 1602.04181, 2016.

[11] Islam MS, Liu C, Li J. Efficient answering of why-not questions in similar graph matching. IEEE Trans. on Knowledge and Data Engineering, 2015,27(10):2672–2686.

[12] Zhu Y, Qin L, Yu J X, *et al*. High efficiency and quality: Large graphs matching. The Int'l Journal on Very Large Data Bases, 2013,22(3):345–368.

[13] Zheng R, Li J, Chen H, *et al*. 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.

[14] Bokhari SH. On the mapping problem. IEEE Trans. on Computers, 1981,(3):207–214.

[15] Narayanan A, Shmatikov V. De-anonymizing social networks. In: Proc. of the 30th IEEE Symp. on Security and Privacy. IEEE, 2009. 173–187.

[16] Zhang Z, Gu Q, Yue T, *et al*. Identifying the same person across two similar social networks in a unified way: Globally and locally. Information Sciences, 2017,394:53–67.

[17] Zhou X, Liang X, Zhang H, *et al*. Cross-platform identification of anonymous identical users in multiple social media networks. IEEE Trans. on Knowledge and Data Engineering, 2016,28(2):411–424.

[18] Man T, Shen H, Liu S, *et al*. Predict anchor links across social networks via an embedding approach. In: Proc. of the IJCAI. 2016,16:1823–1829.

[19] Tan S, Guan Z, Cai D, *et al*. Mapping users across networks by manifold alignment on hypergraph. In: Proc. of the AAAI. 2014,14:159–165.

[20] Liu L, Cheung W K, Li X, *et al*. Aligning users across social networks using network embedding. In: Proc. of the IJCAI. 2016. 1774–1780.

[21] Fabiana C, Garetto M, Leonardi E. De-anonymizing scale-free social networks by percolation graph matching. In: Proc. of the 2015 IEEE Conf. on Computer Communications (INFOCOM). IEEE, 2015. 1571–1579.

[22] Kazemi E, Hassani SH, Grossglauser M. Growing a graph matching from a handful of seeds. Proc. of the VLDB Endowment, 2015, 8(10):1010–1021.

[23] Erd P, Rényi A. On random graphs I. Publicationes Mathematicae Debrecen, 1959,6:290–297.

[24] Karrer B, Newman MEJ. Stochastic blockmodels and community structure in networks. Physical Review E, 2011,83(1 Pt 2): 016107.

[25] Viswanath B, Mislove A, Cha M, *et al*. On the evolution of user interaction in Facebook. In: Proc. of the 2nd ACM Workshop on Online Social Networks. ACM, 2009. 37–42.

[26] Barabási AL, Albert R. Emergence of scaling in random networks. Science, 1999,286(5439):509–512.

冯朔(1989－),男,辽宁沈阳人,学士,CCF 学生会员,主要研究领域为社交网络用户识别,网络对齐.

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.

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,大数据管理.

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.