基于 CLA 算法的跨社交平台用户身份匹配

王李冬[1] 胡克用[1] 周微微[1] 张 赟[2]
[1](杭州师范大学 浙江 杭州 310018)
[2](浙江传媒学院 浙江 杭州 310018)

摘 要 近几年,面向跨社交平台识别分布在不同社交网络上的同一用户依然是一个未解决的难题。该研究可以解决商业应用、资源整合、好友推荐等方面的相关问题。现有的算法如通过文本挖掘、单纯的用户属性无法取得良好的效果。提出 CLA(Combined Link and Attribute)算法实现用户身份匹配。通过好友亲密度获得候选用户,结合基于网络结构的链接信息和用户属性信息进行用户匹配度计算。其中,链接信息相似度利用朋友匹配度计算得到。将该算法应用于多种社交网络,实验结果表明,该算法效果优越于传统的算法效果。

关键词 跨社交网络 用户身份匹配 链接结构 CLA 方法

中图分类号 TP391 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.04.034

CROSS-SOCIAL PLATFORM USER IDENTIFICATION BASED ON CLA ALGORITHM

Wang Lidong[1] Hu Keyong[1] Zhou Weiwei[1] Zhang Yun[2]
[1](*Hangzhou Normal University*, *Hangzhou 310018*, *Zhejiang*, *China*)
[2](*Zhejiang University of Media and Communications*, *Hangzhou 310018*, *Zhejiang*, *China*)

Abstract In recent years, identifying the same users distributed on different social networks for cross-social platforms is still an unsolved problem. This research can solve the problems related to business applications, resource integration, friend recommendation and so on. Existing algorithms such as text mining and simple user attributes cannot achieve effective results. This paper proposed a combined link and attribute (CLA) algorithm for user identification. The candidate users were obtained through calculating friend intensity (FI), and then the user matching degree was calculated by combining the link information based on network structure and user attribute information. The similarity of link information was calculated by the friend matching degree. The algorithm was applied to a variety of social networks. The experimental results show that the CLA algorithm is superior to the traditional algorithm.

Keywords Cross-social network User identification Link structure CLA algorithm

0 引 言

社交网络可以让人们通过 Internet 进行联系和互动,类似现实社会当中的社交行为,典型的如美国的 Facebook、Twitter、Instagram,以及我国的人人网和微博。社交网络提供的服务越来越丰富,从最早期的文本发布到后期的图像与视频共享、用户间关注、评论等,使得越来越多的用户会在不同的社交网络上进行注册以获得不同的服务。

然而不同的社交网络之间信息完全孤立,网民的不同社交网络上的活动行为在不同网站上有不同的表现和侧重性。例如新浪微博以媒体属性为主,人人网以社交属性为主。如果要获得一个用户的完整图像,最大的困难就在于虚拟用户账号及其相应的用户行为分散在不同的社交网络中[3]。对网民在不同社交网络当中的虚拟用户账号进行匹配(如图 1 所示),是实现用户完整图像构建的前提基础。此外,跨社交网络的

收稿日期:2018 - 10 - 11。国家自然科学基金项目(61702151);浙江省自然科学基金项目(Y19F020080,LY17E070004);浙江省软科学研究计划项目(2017C35045);2017 年杭州市社会发展科研项目(20170533B05);杭州师范大学钱江学院科研项目。**王李冬**,副教授,主研领域:信息检索,数据挖掘。**胡克用**,讲师。**周微微**,副教授。**张赟**,副教授。

用户身份匹配对商业应用,好友推荐,网络安全,通信录合并等有重要的意义。

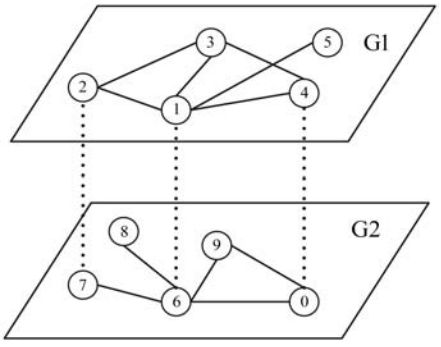

图 1 身份匹配示例

目前的跨社交平台用户匹配技术可以大致分为基于用户图像的方法[4-5],基于内容的方法[6-7]和基于网络结构的方法[8-9]。在基于用户图像的方法中,Perito 等[4]计算用户名的相似度并通过二分类器进行识别。Motoyama 等[5]通过计算用户的属性(教育、职业等)相似度进行匹配。属性信息虽然很容易获得,但在大型社交网络中存在较大的重复性,单纯依靠用户属性方法并无法解决大社交网络的用户匹配问题,而且多数社交网络将用户的属性信息设定为隐私数据。在基于内容的方法中,Kong 等[6]将用户的时间信息、空间信息和文本信息综合起来计算相似度。Goga 等[7]利用用户的地理位置、时间戳和书写风格进行识别。虽然基于内容的方法具备一定的有效性,但上述的地理和空间信息在社交网络中存在稀疏特性,方法难以适用大范围社交网络。在基于网络结构的方法中,Bartunov 等[8]提出基于 JLA(Joint Link-Attribute)的识别方法,该方法共同考虑了用户属性和链接信息。Narayanan[9]等提出完全基于链接属性的识别方法。现有研究表明,融合用户图像信息、内容信息和网络信息的方法往往比单个方法的效果要好[10,12]。

本文主要研究跨社交网络的局部身份匹配问题,在给定先验种子节点集的基础上,在两个社交网络中推断出所有潜在的匹配用户对集合。基于上述思想,本文提出融合链接信息和属性信息的 CLA(Combined Link and Attribute)算法。该算法首先通过特定的属性信息和人工选取得到先验种子用户,然后利用好友亲密度获得候选用户进行匹配。针对候选匹配用户,通过融合链接信息和属性相似度匹配准则计算得到匹配值最高的用户作为匹配用户,然后将匹配用户作为种子用户迭代运行,直到没有新的匹配用户产生。

1 CLA 算法

CLA 算法的目的是为了准确匹配尽可能多的用户,算法的大致流程如图 2 所示。算法首先根据用户的用户名、email 等信息选取先验种子用户,再从种子用户的好友集中选择候选用户进行匹配,根据属性相似度和链接相似度计算匹配准则,将匹配得到的用户作为新的种子用户,迭代运行,直到没有新的匹配用户产生。

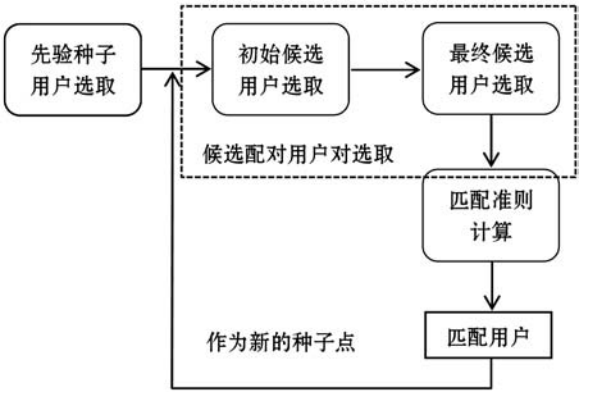

图 2 CLA 算法流程

1.1 先验配对用户对选取

定义 1(配对用户对) 给定两个社交网络,分别表示为 $GA=(U_A,E_A,S_A)$、$GB=(U_B,E_B,S_B)$。U_A表示网络 GA 的用户实体集合,E_A为网络 GA 的用户关系(链接关系),U_B表示网络 GB 的用户实体集合,u_{Ai}代表用户集合 U_A中的第 i 个用户,u_{Bj}代表用户集合 U_B中的第 j 个用户。S_A表示网络 GA 的用户属性集合,s_{Ai}代表用户 u_{Ai}属性向量。若用户 u_{Ai}和用户 u_{Bj}在现实生活中属于同一个体,则 u_{Ai}和 u_{Bj}被认定为配对用户对,记为 UMP。

定义 2(先验种子用户) 先验种子用户是指通过特定方法找出的先验配对用户,由此找到更多的配对用户对,本文记为 $PUMP$。

目前没有通用的方法适用于获取任意两个社交网络的先验种子用户,一般需要针对特定社交网络选取特定的属性信息进行识别。Balduzzi 等[1]提出利用 email 对用户进行判定。由于 email 的唯一性,利用 email 进行判定是较好的方法。

此外,文献[3]指出,同个用户往往在不同的社交网络使用同一个昵称(nickname)。因此,若两个社交网络中用户的用户名完全一样,可认定为该对用户为同一对象,匹配用户对。然而,部分社交网络允许不同的用户以相同用户名进行注册,如人人网。单单通过用户名无法直接判断两用户是否属于同一人,一种简单的解决方法就是通过其他可获取的因素,如地理位置、生日、工作单位、性别等属性信息进行进一步确认。此外,部分网络会提供额外的信息,如 twitter,该网络提供独特的 URL 地址用于用户自识别,因此针对 twit-

ter 的先验配对用户获取可直接利用该 URL 与 Facebook 进行用户身份识别。本文拟综合利用 Email、URl 等信息的相等匹配获得先验种子用户,并结合生日、工作单位、性别等属性信息通过人工挑选对用户进行准确性的甄别。

1.2　候选配对用户对选取

文献[11]对 129 个同时在新浪网和人人网注册的用户进行调研,发现这些用户的好友集中有 67.5% 同时存在于新浪网和人人网。基于此,本文假设:若两个用户相配对,则他们的好友中存在匹配用户对的概率较大。为了选取候选配对用户,可以从种子用户的好友集出发。本文将候选用户选取分为初始候选用户选取和最终用户选取。其中,初始候选用户选取的规则定义如下:

定义 3(初始候选用户对)　若 u_{Ai} 和 u_{Bj} 为两个社交网络中的种子用户,$u_{Ak} \in friend(u_{Ai})$,$u_{Bl} \in friend(u_{Bj})$,则 (u_{Ak}, u_{Bl}) 属于初始候选用户对 CMP_P,定义为:

$$CMP\_P = \{(u_{Ak}, u_{Bl}) \mid u_{Ak} \in friend(u_{Ai}) \wedge u_{Bl} \in friend(u_{Bj})\} \tag{1}$$

式中:u_{Ak} 和 u_{Bl} 分别代表两个社交网络的种子用户;$friend(u_{Ai})$ 代表用户 u_{Ai} 的好友集。

在大规模的社交网络环境下,部分种子用户的好友集数量较多。若对两个网络的种子用户的好友集用户进行两两比对,则需要耗费较长的运行时间,尤其在海量社交网络的大数据环境下无法保证运行效率。为了提高匹配效果,本文通过特定的机制预先获得在另一个网络中更有可能存在匹配节点的用户,然后将此类用户与另一个网络中的种子点用户的好友集逐一进行匹配度计算。为了对初始候选用户对进一步缩小范围以提高匹配效率,我们拟利用用户聚簇特性对某一个网络中的用户是否可能在另外一个网络中存在匹配用户进行评估。文献[3]提出,如果一个用户与已经识别的用户为好友关系且具备较高的亲密度,则该用户可以被优先选择。本文将好友亲密度定义如下:

定义 4(好友亲密度(Friend Intensity,FI))　若 u_{Ai} 和 u_{Aj} 代表社交网络 G_A 中的两用户,u_{Am} 属于用户 u_{Ai} 和 u_{Aj} 的共同好友。F_{Ai}、F_{Aj} 和 F_{Am} 分别表示用户 u_{Ai}、u_{Aj} 和 u_{Am} 的好友集合。好友亲密度函数定义如下:

$$FI(u_{Ai}, u_{Aj}) = \frac{2 \times \sum_{f_{Am} \in F_{Am}} \deg(f_{Am})^{-1} + (\deg(u_{Ai}) + \deg(u_{Aj}) - \deg(u_{Am}))^{-1}}{\sum_{f_{Ai} \in F_{Ai}} \deg(f_{Ai})^{-1} + \sum_{f_{Aj} \in F_{Aj}} \deg(f_{Aj})^{-1}} \tag{2}$$

式中,deg( )代表用户的度。

基于某用户的所有好友亲密度,我们通过用户的好友评分对该用户是否有可能在另外一网络中存在匹配用户进行评估。假设已经识别的用户集表示为 u_{match},本文将某用户 u_{Ai} 的好友评分 $FC(u_{Ai})$ 定义如下:

$$FC(u_{Ai}) = \sum_{u_{Aj} \in u_{match}} FI(u_{Ai}, u_{Aj}) \tag{3}$$

若用户 u_{Ai} 的 FC 值越大,则该用户在另一个网络中存在匹配用户的可能性也越大。基于上述定义,先将先验种子集的朋友集设定为初始候选用户,再根据好友亲密度从初始候选用户集中选择特定用户作为最终候选用户账号集 u_{select}。该类用户具备较大概率在另外一个网络中存在匹配用户,从而大大提升匹配效率。具体算法如下:

算法 1　候选用户选取

输入　社交网络 G_A,G_B,先验种子用户集 $PUMP$。

输出　候选用户账号集合 u_{select}。

1 根据式(1)得到初始候选用户对集合 $CMP\_p = \{u_A, u_B\}$;

2 for each $u_{Ai} \in u_A, u_{Bj} \in u_B$

3　　根据式(3)计算 $FC(u_{Ai})$,$FC(u_{Bj})$

4 end for

5 根据 $FC(u_{Ai})$ 和 $FC(u_{Bj})$ 对 u_A,u_B 的元素进行排序;

6 $u_{select} = \{u_A[0], u_A[1], u_B[0], u_B[1]\}$

1.3　用户链接匹配度计算

为了实现两个用户之间的匹配,需要计算链接匹配度 CR,现有的 NS 算法[2]通过度数计算实现:

$$CR(u_{Ai}, u_{Bj}) = \frac{s_{in}}{\sqrt{d_{in-Bj}}} + \frac{s_{out}}{\sqrt{d_{out-Bj}}} \tag{4}$$

式中,s_{in} 和 s_{out} 分别代表用户 u_{Ai} 和用户 u_{Bj} 的共享入度邻居好友和共享出度邻居好友的数目。入度好友指社交网络中好友的单向关注关系。d_{in-Bj} 和 d_{out-Bj} 分别代表用户 u_{Bj} 的入度和出度。然而,该算法需要假设不同社交网络中的同一用户具备相同的入度和出度,该假设与现实世界的社交网络不符。在本文方法中,我们拟引入朋友关系实现用户链接匹配度计算。

定义 5(朋友匹配度(Friend Match Degree,FMD))　F_{Ai} 和 F_{Bj} 分别代表用户 u_{Ai} 和用户 u_{Bj} 的好友集,$F_{Ai} \cap F_{Bj}$ 代表两用户的共同好友,则两用户的朋友匹配度定义如下:

$$FMD(u_{Ai}, u_{Bj}) = \frac{2 \times w(F_{Ai} \cap F_{Bj})}{w(F_{Ai}) + w(F_{Bj})} \tag{5}$$

式中:

$$w(F_{Ai}) = \sum_{u \in F_{Ai}} 1/\deg(u) \tag{6}$$

若 $F_{Ai} = F_{Bj}$,则 $FMD(u_{Ai}, u_{Bj}) = 1$。然而,该方法存在一个弊端,若 F_{Ai} 的数目较少,则会出现错误匹配的情况。如图 3 所示,$F_{A1} = F_{B7} = \{u_3\}$,则 $FMD(u_{A1},$

u_{B7}) = 1,使得 u_{A1} 和 u_{B7} 被错误配对。

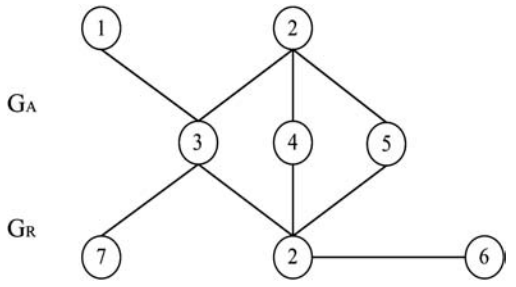

图 3 网络对示例

为了解决式(5)带来的错误匹配问题,我们加入共同好友因子进行调整,将式(5)调整如下:

$$FMD(u_{Ai},u_{Bj}) = |F_{Ai} \cap F_{Bj}| + \frac{2 \times w(F_{Ai} \cap F_{Bj})}{w(F_{Ai}) + w(F_{Bj})} \quad (7)$$

$|F_{Ai} \cap F_{Bj}|$ 代表已经识别的共同好友个数,包括初始种子点。$FMD(u_{Ai},u_{Bj})$ 值越高,代表两用户为匹配用户的概率就越大。

1.4 属性相似度计算

用户之间的属性相似度拟采用用户名、姓名、URL 和 Email 信息。

用户名和姓名都可表示为字符串。部分文献采用 Levenshtein 距离进行度量[14]。Levenshtein 距离作为计算两个字符串间的差异程度的字符串度量,曾被多次应用于用户名的差异度量并取得较好的效果[11]。两个用户名之间的用户名相似度计算如下:

$$Sim_{na} = 1 - \frac{lev(n_1,n_2)}{\max(l(n_1),l(n_2))} \quad (8)$$

式中:$lev(n_1,n_2)$ 表示用户名 n_1 和用户名 n_2 之间的 Levenshtein 距离;$l(n_i)$ 表示 n_i 的字符数。该方法可针对用户名信息进行相似度度量,但针对姓名信息无法实现有效衡量。

姓名信息(可选),在多数的网络中都会出现,例如 Facebook 和 Twitter。该信息可作为与用户名同等重要的属性字段进行身份匹配,但无法作为身份识别的唯一判定信息[13]。Levenshtein 距离对顺序较敏感,完全相同的名字,若“姓”和“名”的顺序倒置,将产生完全不一样的计算结果。鉴于国外社交网络的姓名中,“姓”和“名”的顺序并无统一规则,本文利用 VMN 算法[6]对姓名进行度量。VMN 是一种非常有效的名字匹配技术,可以对姓名等信息实现模糊匹配,得到 0 或 1 的匹配结果值。在 VMN 算法中,名字“Tony Xie”和“Xie Tony”的相似度为 1。

URL 信息(可选),若某社交网络提供 URL 信息助于身份识别,则根据 URL 信息与相应社交网络的链接地址进行比对,若完全相同,则返回 1,否则为 0。

Email 信息(可选),若两个用户的 Email 完全相同,则该属性相似度为 1,否则为 0。不同社交网络上的两用户若具备相同的 Email,则他们为同一个体的概率非常大。Email 信息是进行身份识别的有效属性字段。

上述信息中,姓名信息、URL 信息以及 Email 信息需要根据特定的社交网络做相应的选择。通过上述的属性相似度计算可以得到用户 u_1 和用户 u_2 之间的相似度向量 $H(s_1,s_2)$,s_1 和 s_2 为其各自的属相向量。将已知的匹配用户对的属性相似度向量作为训练向量,不同属性信息的相似度作为不同的向量维度值。基于此,用户身份是否匹配转化为一个二分类问题,即 $C(H(s_1,s_2)) \in [0,1]$,C 代表分类器,1 代表 u_1 和 u_2 为同个用户,否则为不同用户。本文利用 SVM 分类器将向量集合进行监督学习训练,根据得到的 SVM 分类器来对新的向量进行分类。

1.5 用户匹配准则

定义 6(用户匹配度) 给定不同社交网络的两个用户 u_{Ai} 和 u_{Bj},两者的账号属性向量分别为 s_{Ai} 和 s_{Bj}。他们之间的用户匹配度计算如下:

$$Mat(u_{Ai},u_{Bj}) = C(H(s_{Ai},s_{Bj})) \times \alpha + FMD(u_{Ai},u_{Bj}) \quad (9)$$

式中:$C(H(s_{Ai},s_{Bj})) \in [0,1]$ 代表用户 u_{Ai} 和 u_{Bj} 的属性相似度分类结果,$FMD(u_{Ai},u_{Bj})$ 代表用户 u_{Ai} 和 u_{Bj} 的链接匹配度结果(见式(7)),参数 α 用于平衡属性相似度和基于链接的用户链接匹配度,本文将其定义为已经识别出的用户个数。

具体的匹配算法如算法 2。该算法将式(1)得到的初始候选用户对和算法 1 得到的候选用户账号集合 u_{select} 作为输入,然后针对候选用户账号集合中的每一个用户,在另一个网络中的初始候选用户对范围内搜索匹配用户。若两者之间的用户匹配度最高,则视为配对。

算法 2 匹配准则计算

输入:初始候选用户对集合 u_A,u_B;
候选用户账号集合 u_{select}。
输出:匹配用户账号 u_{match}。
1 $u_{match} = \varnothing$
2 for each $u_{select(i)} \in u_{select}$
3 if $u_{select(i)} \in u_A$ then
4 for each $u_{Bi} \in u_B$ do
5 根据式(9)计算 $Mat(u_{select(i)},u_{Bi})$
6 end for
7 $u_{match} = u_{match} \cup \text{argmax}_{u_{Bi} \in u_B}(Mat(u_{select(i)},u_{Bi}))$
8 else
9 若 $u_{select(i)} \in u_B$ 用同样方法执行
10 end for

2　实验与讨论

本文使用文献[8]的 Facebook 和 Twitter 基准数据集。Facebook 和 Twitter 基准数据集共包含 16 个来自 Facebook 和 Twitter 的网络对。帐号的属性信息包含用户名、姓名和 URL 信息。由于在 ego-network 上研究,因此在样例网络中,任意两个用户之间的距离不会超过两个人。数据集已经标注两个网络中的匹配用户对,并同时标注了种子用户,具体相关信息如表 1 所示。图 4 显示了 Facebook-Twitter 的一个数据集示例,该数据集共 10 对匹配用户,图中标示出了 5 对。

表 1　Facebook-Twitter 数据集信息

信息项	Twitter	Facebook
用户数量	398	977
链接数目	1 728	10 256
匹配用户对	141	
已知种子用户	71	

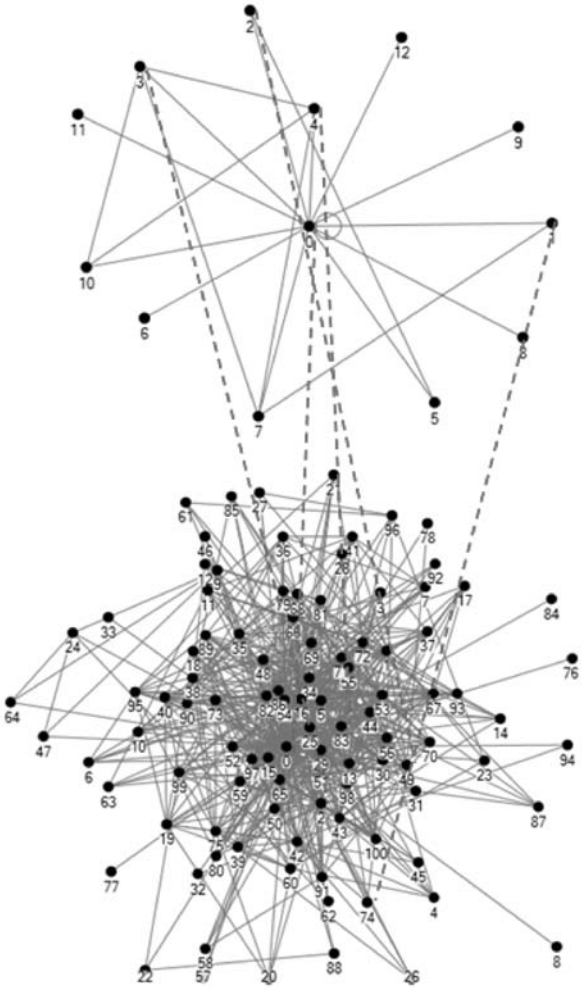

图 4　Facebook-Twitter 数据集示例

此外我们从人人网和新浪网上搜集了相应的数据组成数据集。新浪微博数据库是从新浪微博的搜索界面上抓取,而人人网的数据库则是从其开放的 API 获取。人人网和新浪网的数据如表 2 所示。其中,新浪爬取的用户信息包括用户 ID、用户名、微博数、粉丝数、关注数,以及关注关系等。人人网用户信息包括用户名,户主的所有好友、户主及好友关注的公共主页等。

表 2　新浪-人人数据集信息

网络	节点	边	平均度
Sina 微博	1.23×10^4	2.1×10^4	3.3
人人网	5.7×10^4	13.7×10^4	5.4

在评价方案中,本文利用传统的准确率、召回率以及 F1-measure 综合进行衡量。具体如下:

$$recall = tp/(tp + fn) \tag{10}$$

$$precision = tp/(tp + fp) \tag{11}$$

$$F1\text{-}measure = \frac{2 \times recall \times precision}{precision + recall} \tag{12}$$

式中:tp 代表被正确匹配的用户对;fp 代表被错误匹配的用户对,fn 代表未被匹配的用户对。

采用不同的基准算法与本文方法进行比较,分别为 JLA 方法[8]、NS 方法[2]。算法实验在 Windows7 环境下采用 Matlab 编写。由于 Facebook-Twitter 数据集已经具备种子用户,则本文算法 CLA 在执行过程中省略了先验种子用户选取的步骤。

其中,JLA 方法使用监督分类器的剪枝手段。所有方法的结果都是取 16 对网络对的均值。由表 3 可得,无论从 Facebook 到 Twitter 的匹配,还是 Twitter 到 Facebook 的匹配,JLA 算法可以保持最高的准确率,但是该方法的召回率并不十分理想。其中,NS 方法效果最差,可见,仅仅以网络拓扑为计算依据的方法远比综合属性因素和链接关系的方法要差。CLA 方法的效率和 JLA 相比稍显优越,而且 JLA 中基于条件随机场的最优用户映射实现方法比 CLA 方法中基于判定准则的匹配方法比更加复杂,JLA 方法需额外利用基于监督分类器的剪枝操作才可获得相对满意的效果。若应用于海量用户的跨社交平台,JLA 的繁琐步骤无法保证效率。

表 3　Facebook-Twitter 数据集的身份匹配效果

方法名称	准确率	召回率	F1
JLA (Twitter -> Facebook)	1.00	0.39	0.561
JLA (Facebook -> Twitter)	1.00	0.58	0.734
NS	0.759	0.632	0.690
CLA	0.792	0.773	0.782

针对新浪-人人数据集,我们抽取四个子网络进行实验。由于获得的用户数据并不包含具体属性信息,本文的 CLA 方法在进行匹配度计算时仅计算用户链接匹配度。抽取子网络时,针对两个网络中的相同用户对进行二层好友的深度遍历。然后,针对每个子网络人工标注 20 个种子用户并分别执行 CLA 算法和 NS 算法。由于子网络中的匹配用户对的总数目未知,我们仅计算准确率进行效果分析。由表 4 数据可得,Sina 微博抽取的子网络规模约 1 000 节点,人人网抽取的子网络规模约 3 000 节点。CLA 方法和 NS 方法都可以获得一定数量的匹配用户对,但是 CLA 方法在所有的实验中都保持较高的准确率。该结果表明,在仅考虑网络链接结构的情况下,本文提出的方法性能依然比 NS 方法更加优越。

表 4 新浪-人人数据集的身份匹配效果

	子网络用户对		匹配的用户对数目		准确率	
			CLA	NS	CLA	NS
1	Sina RenRen	1 134 3 774	249	95	0.432	0.213
2	Sina RenRen	1 025 3 436	221	83	0.507	0.237
3	Sina RenRen	1 120 3 554	242	101	0.452	0.228
4	Sina RenRen	987 2 948	202	74	0.513	0.371

3 结 语

本文提出一种基于 CLA 方法的跨社交网络的身份识别,并将其应用于真实社交网络的数据集上。通过在不同数据集上的实验,结果表明该方法可获得较高的准确率,效果优于传统的 JLA 和 NS 方法,而且单单基于网络链接结构的用户身份识别效果依然优于 NS 方法。然而,CLA 方法依然存在一定的不足。该方法可应用于海量社交网络群的用户匹配,但主要还是面向两两社交网络。若针对三个或三个以上的社交网络,可能存在两两社交网络之间的用户匹配结果不一致的情况。因此,在今后的工作中,本文拟针对上述问题进行改进,以适应多社交网络的用户身份匹配,同时在算法效率上拟作进一步提升。

参 考 文 献

[1] Balduzzi M, Platzer C, Holz T, et al. Abusing social networks for automated user profiling[C]//Proceedings of the 13th international conference on Recent advances in intrusion detection. Berlin: Springer-Verlag, 2010, 422-441.

[2] Narayanan A, Shmatikov V. De-anonymizing Social Networks[C]//Proceedings of the 2009 30th IEEE Symposium on Security and Privacy. IEEE Computer Society, 2009: 173-187.

[3] 孟波. 多用户社交网络识别算法研究[D]. 大连: 大连理工大学, 2015.

[4] Perito D, Castelluccia C, Kaafar M A, et al. How unique and traceable are usernames? [C]//Proceedings of the 11th international conference on Privacy enhancing technologies. Berlin: Springer-Verlag, 2011: 1-17.

[5] Motoyama M A, Varghese G. I seek you: searching and matching individuals in social networks. [C]//Proceedings of the eleventh international workshop on Web information and data management. New York: ACM, 2009: 67-75.

[6] Kong X, Zhang J, Yu P S. Inferring anchor links across multiple heterogeneous social networks[C]//Acm International Conference on Information & Knowledge Management. ACM, 2013: 179-188.

[7] Goga O, Lei H, Parthasarathi S H K, et al. Exploiting innocuous activity for correlating us-ers across sites[C]//Proceedings of the 22nd international conference on World Wide Web. New York: ACM, 2013: 447-458.

[8] Sergey B, Anton K, Seungtaek P, et al. Joint link-attribute user identity resolution in online social networks[C]//Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis. ACM, 2012.

[9] Narayanan A, Shmatikov V. De-anonymizing Social Networks[C]//2009 30th IEEE Symposium on Security and Privacy. IEEE Computer Society, 2009: 173-187.

[10] Shu K, Wang S, Tang J, et al. User Identity Linkage across Online Social Networks: A Review[J]. ACM SIGKDD Explorations Newsletter, 2017, 18(2): 5-17.

[11] Zhou X, Liang X, Zhang H, et al. Cross-Platform Identification of Anonymous Identical Users in Multiple Social Media Networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 28(2): 411-424.

[12] Buraya K, Farseev A, Filchenkov A, et al. Towards User Personality Profiling from Multiple Social Networks[C]// Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[13] Peled O, Fire M, Rokach L, et al. Matching Entities Across Online Social Networks[J]. Neurocomputing, 2016, 210: 91-106.

[14] Buccafurri F, Lax G, Nocera A, et al. Discovering links among social networks[J]. Lecture Notes in Computer Science, 2012, 7524: 467-482.