计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(12):94−103 [doi: 10.15888/j.cnki.csa.006101]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于推文与属性的社交网络用户重识别方法①

高　伟[1,2], 张　敏[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 软件研究所, 北京 100190)

摘　要: 大数据隐私安全正成为各界关注的热点. 攻击者通过识别用户不同网站的账户, 可以构建用户的完整画像, 对用户隐私形成威胁. 模拟评估攻击者的重识别能力是进行用户隐私保护的前提. 因此, 本文提出一种高相似同天同行为算法. 该算法通过检测账户在不同网站是否存在多次同天发表相近或相同内容的行为, 判断账户是否属于同一用户, 并通过为用户属性构建一种权重计算模型, 进一步提高用户重识别的准确率. 经过对两个国内主流社交网站的一万多用户进行实验, 本文算法表现出良好的效果. 实验表明, 即使不考虑用户社交关系, 用户的推文与属性依然提供了足够的信息使攻击者将用户不同网站的账户相关联, 从而导致更多的隐私被泄露.

关键词: 社交网络; 用户重识别; 推文; 属性; 相似度

引用格式: 高伟,张敏.基于推文与属性的社交网络用户重识别方法.计算机系统应用,2017,26(12):94–103. http://www.c-s-a.org.cn/1003-3254/6101.html

Method for Users Re-Identification across Social Networks Based on Tweets and Attributes

GAO Wei[1,2], ZHANG Min[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Big data Privacy security is becoming the hot spot in the various social industries, because attackers can build an integrate portrait to threaten privacy of users by identifying accounts in different sites. Simulation assessment of the attacker re-identification ability is the precondition of users' privacy protection. Therefore, this paper proposes a high similarity algorithm in same day with same behaviors. The core idea of the algorithm is as follows: if a couple account issues similar or identical content on the same day, which also appears many times in different websites, then these two accounts may belong to a person with a high possibility. In addition, this paper builds a new weighting model for the users' attributes to improve the accuracy of user re-identification. After the experiment on more than ten thousand users of the two major domestic social networking site, this algorithm proves to be effective. Experimental results show that even if attacker don't consider users' social relations, the users' tweets, attributes, still provide enough information to make the attacker correlate their different accounts, which will lead to leak of more privacy.

Key words: social network; users re-identification; tweets; attributes; similarity

1　引言

目前社交网络已广泛普及. 截止 2016 年, 全球最大社交网站 facebook 月活跃用户数已突破 16.5 亿, 新浪微博、QQ 月活跃用户分别突破了 3.9 亿、8.5 亿, 社交网络的迅猛发展为社会带来了巨大机遇. 在社交网络上, 每天有大规模数据产生, 如推文内容、签到信息、照片等. 随着"云计算"和"大数据"技术的不断深入, 众多研究机构、高校、互联网公司开始广泛搜集

① 基金项目: 国家自然科学基金重点项目 (61232005); 国家自然科学基金 (61402456)
收稿时间: 2017-03-16; 采用时间: 2017-04-07

这些碎片化信息, 通过对这些大规模数据的建模分析, 可以了解用户多维度的的画像, 如购物习惯、兴趣爱好等, 以此进行广告精准投放或者好友推荐等, 具有极大的商业价值和实用价值.

但这同时也带来了用户隐私泄露的威胁. 攻击者采集用户不同网站的信息, 通过对这些信息的链接并加以建模抽象, 可构建用户的完整画像. 当攻击者对这些信息进行非法利用时, 会严重破坏用户的隐私, 甚至直接威胁到用户的人身财产安全. 因此, 保护用户隐私就显得相当重要. 在此情形下为保护用户隐私, 需首先了解, 攻击者判定来源于不同社交网站的账户属于同一人所采用的技术手段, 即用户重识别技术.

用户重识别就是通过采集多个社交网站的数据, 通过对这些公开数据的比对, 来识别用户不同网站账户的一种技术. 目前, 用户重识别技术正受到国内外前所未有的关注. 2009 年 Jan[1]分别采用不同精度的匹配方法对 Facebook 和 StudiVZ 社交网站的用户姓名、生日、性别、高中、地址等进行了相似度计算, 并结合权重完成了用户的重识别工作. 2013 年 Goga[2]提出了基于多社交网站大规模账户的相关性识别算法. 分别采用 Jaro Distance、哈希感知算法等对用户姓名、用户头像及其他属性进行了相似度计算, 最后采用机器学习方法进行用户的匹配. 2013 年 Goga[3]根据用户推文的地理位置、发表时间和内容风格对 Twitter、Flickr、Yelp 的用户展开重识别研究. 2014 年 Cecaj[4]根据用户电话记录与推文记录的发生时间差Δt 和距离差Δs, 对某地区的用户进行重识别. 基于社交关系的用户重识别主要根据用户在不同的社交网络有着相似的朋友圈这一经验, 通过选定部分已知种子匹配用户, 根据网络拓扑关系、图结点的度数等进行用户的重识别研究. 如 2009 年 Narayanan[5]提出的基于种子匹配的重识别算法, 2016 年 Zhou[6]提出的基于好友相似度的重识别算法 FRUI. 2012 年 Bartunov[7]利用用户属性和社交关系综合计算了 Facebook 和 Twitter 的用户相似度. 2013 年 Kong[8]将用户的发推地理位置变化规律、时间变化规律、推文内容关键字出现频率相结合, 完成用户的匹配工作. Fu[9]利用用户属性、社交关系, 采用图结点算法进行了用户重识别研究.

以上方案基于不同特征对用户重识别进行了研究, 但在推文方面, 用户重识别的准确率还较低, 利用用户发布的大量推文进一步提高准确率仍有很大探索空间. 此外, 用户属性包含着一个人的重要信息, 对用户重识别具有一定的意义. 然而随着社交网络的不断发展, 用户的安全意识越来越强, 很多用户的关键属性信息被隐藏, 为这些仅存的属性信息构建权重计算模型, 面临着较大的困难.

为解决以上问题, 本文提出一种高相似同天同行为算法. 该算法通过检测账户在不同网站是否存在多次同天发表相近或相同内容的行为, 判断账户是否属于同一用户. 此外, 为利用用户属性进一步提高重识别准确率, 本文构建了一种属性权重计算模型. 为评估所提算法的性能, 本文以国内两个主流的社交网站 (以下简称“Q”、“R”) 作为实验对象, 分别采集了 Q 网站的 16173 个用户的 300 余万推文、R 网站的 10027 个用户的 70 余万推文, 经人工标注了 776 对真实匹配用户. 实验表明, 本文所提算法有着良好的效果, 明显优于其他模型.

2 用户重识别方法

本文提出的用户重识别方法主要基于用户发表的推文与用户属性, 并在此基础上, 将二者结合进行用户重识别. 首先, 定义以下基本符号.

2.1 符号定义

$UQ=\{\cdots UQ_i\cdots\}$, Q 网站用户集合.

$UR=\{\cdots UR_j\cdots\}$, R 网站用户集合.

$W=\{\cdots W_n\cdots\}$, 词集合.

$T=\{\cdots T_m\cdots\}$, 日期 (yyyy-mm-dd) 集合.

$TQ_i=\{\cdots T_c\cdots|T_c\subset T\}$, UQ_i 所有发推日期集合.

$TR_j=\{\cdots T_d\cdots|T_d\subset T\}$, UR_j 所有发推日期集合.

$tw=\{\cdots \mathrm{tw}_z\cdots|tw_z=W_1\times W_2\cdots W_x, x\subset N^+\}$ 推文内容集合, 推文内容由词组成.

$TW=\{UQ\cup UR\}\times T\times tw\}$, 推文向量空间, 其元素是三元组.

$TWQ^m=\{\cdots twq_h\cdots|twq_h\subset tw\}$, Q 网站所有用户在日期 T_m 发表的推文集合.

$TWR^m=\{\cdots twr_k\cdots|twr_k\subset tw\}$, R 网站所有用户在日期 T_m 发表的推文集合.

$TWQ_i^m=\{\cdots twq_f\cdots|(UQ_i,T_m,twq_f)\subset TW,twq_f\subset tw\}$, UQ_i 在日期 T_m 发表的推文集合.

$TWR_j^m=\{\cdots twr_e\cdots|(UR_j,T_m,twr_e)\subset TW,twr_e\subset tw\}$, UR_j 在日期 T_m 发表的推文集合.

2.2 基于推文的用户重识别方法

推文内容与用户联系紧密, 不同的用户其推文内

容也显示出较大差异, 因此在一定程度上, 推文内容可以反映用户特征、代表用户身份. 基于推文的用户重识别方法包括四部分. 1) 推文向量及相似度计算方法. 通过 word2vec[10,11]工具训练获得每个词的向量, 然后将推文中词的向量累加得到推文的向量, 推文相似度采用推文向量夹角的余弦值来表示. 2) 高相似同天同行为计算方法. 很多用户都有在同一日期于不同社交网站发表相似推文的经历, 因此在一定意义上, 高相似的推文对可以为用户重识别提供线索. 该方法正是基于此思想, 发现具有相同发推日期的高相似推文对. 3) 热点事件处理方法. 在一些特殊节日、热点事件时, 大量相似的推文会同时出现于不同的社交网站, 将严重影响用户的重识别结果, 因此该方法根据一定条件对与热点事件有关的推文记录进行删除, 以降低其负面影响. 4) 高相似多推文计算方法. 用户在两个社交网站的推文在整体上具有一致的情感、用词习惯等, 所以对于不同网站的账户, 可以根据它们推文的整体相似度, 展开用户重识别的研究工作.

2.2.1 推文向量及相似度计算方法

为计算推文相似度, 需要首先将推文内容使用数字向量表示. 因此, 推文相似度的计算就转化为数字向量的计算. 根据推文的特点, 本文将按照图 1 所示流程进行推文向量的计算.

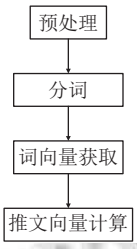

图 1 推文向量计算图

预处理: 去除推文中的乱码、符号. 如果推文只包含乱码或者符号等非文字性语言, 则不做处理.

分词: 采用 ICTCLAS 分词工具将推文分为单独的词.

词向量获取: 从已训练好的词向量表中获取对应词的数字向量. 该表中的词向量是利用 Google 的开源工具 word2vec, 经过对 20G 大规模维基百科语料训练而成, 其包含每个词的 50 维度语义向量.

推文向量计算: 推文由词组成, 其中每个词的向量为:

$$V(W_X)=\{R_1\times\cdots\times R_k\times\cdots\times R_{50}, R_k\in R, 1\le k\le 50\}$$

推文的向量由各个词的向量累加而成, 即:

$$V(tw_z)=\sum V(W_x) \tag{1}$$

本文曾将每条推文看作一个文本, 所有推文看作文档总数据集. 利用 TF-IDF 算法计算每个词在对应推文中的权重, 将权重乘以对应词的向量得到该词在推文中的语义向量, 然后将推文中所有的词向量累加, 得到该推文的向量. 但在后续实验中, 其效果与直接将词向量累加得到的推文向量所进行的实验效果几乎一致. 经分析, 其原因可能有以下两种: 1) 大部分推文属于短文本, 用户在不同网站同天发表相近含义的推文时, 用词基本一致, 极少部分用词不一致的相似推文也只是一些非关键词不同, 对用户重识别的影响很小. 2) 如果推文属于长文本, 用户往往在另一个网站发表的是该推文的拷贝版本, 二者完全相同. 所以, 出于算法与系统的整体效率考虑, 本文采用将词向量直接累加的结果来表示推文向量.

得到推文向量之后, 推文间的相似度就转化为两个向量的相似度, 本文采用向量夹角的余弦值来表示任意推文对(twq_f, twr_e)的相似度, 即:

$$Sim(twq_f, twr_e)=\frac{V(twq_f)\times V(twr_e)}{|V(twq_f)|\times|V(twr_e)|} \tag{2}$$

2.2.2 高相似同天同行为计算方法

为开展 Q 与 R 网站用户的重识别工作, 定义以下概念:

(1) 同天同行为: 如果推文 twq_f 与 twr_e 都发表于相同的日期 T_m, 则称这样的现象为推文对(twq_f, twr_e)的同天同行为.

(2) 高相似同天同行为: 如果推文对(twq_f, twr_e)具有同天同行为现象, 且其相似度$Sim(twq_f, twr_e)$大于某个相似系数 S, 则称该推文对是高相似同天同行为的.

(3) 高相似次数: 用户 UQ_i 与 UR_j 具有的高相似同天同行为推文对的总次数, 称为该用户对的高相似次数.

基于推文的用户重识别研究方法需要首先计算 Q 网站每个用户 UQ_i 与 R 网站所有用户的高相似同天同行为, 其原理如图 2 所示.

图 2 中白色部分与灰色部分对称, 分别表示 Q、R 网站的推文信息. 其中白色部分从左至右的每一列分别表示: Q 网站用户发推日期 T_x 列表、日期 T_x 对应的推文集合 TWQ^x 列表、推文集合 TWQ^x 包含的推文 twq_h 列表. 其中“×”表示对推文集合计算笛卡尔积, 然

后计算结果集中每个推文对元素的相似度. 完成以上计算后, 判断该推文对元素是否属于高相似同天同行为, 是则保存, 否则丢弃.

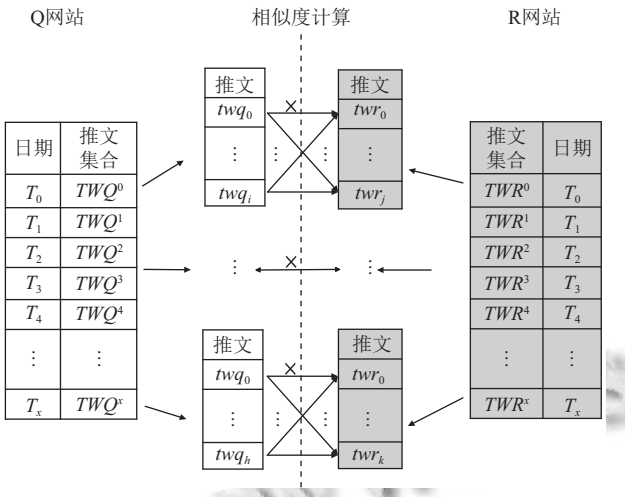

图 2 所有用户的高相似同天同行为计算图

其伪代码算法如下:

Algorithm 1. High Similarity

```
For c = 0 to n do
    TQ=TQ∪TQ_c #计算Q网站所有用户的发推日期并集
End For
For d = 0 to m do
    TR=TR∪TR_d # 计算R网站所有用户的发推日期并集
End For
TQR=TQ∩TR # 计算Q与R网站用户发推日期的交集
For each T_x∈TQR do
TWQR=TWQ^x × TWR^x# 作笛卡尔积得到同天同行为推文对
    For each pair of (twq_f, twr_e) ∈TWQR
        If Sim(twq_f, twr_e) >S # 判断是否高相似
            getUsers(twq_f, twr_e) #获取推文所属用户UQ_i, UR_j
            store(UQ_i, UR_j, T_x, twq_f, twr_e, Sim(twq_f, twr_e)) #存储信息
        End If
    End For
End For
```

经过以上计算, 即可得到 Q 网站的每一个用户 UQ_i 与 R 网站所有用户的高相似同天同行为.

2.2.3 热点事件处理方法

在特殊节日、热点事件时, 很多用户都会发表含义相近的推文于不同网站. 在这种情形下, 一个用户会与很多用户存在高相似同天同行为, 而大部分用户是非真实匹配用户. 如果以此为基础, 根据用之间的高相似次数决定匹配用户, 其结果很有可能是不准确的, 同时这些过多的干扰数据, 也会给系统的运行带来压力, 增大损耗. 如图 3 所示的真实案例.

Q网站用户	发推日期	Q网站推文
UQ_0	2011-03-31	希望是假的
	2011-02-02	新年快乐!

R网站推文	发推日期	R网站用户
希望是假的	2011-03-31	UR_0
新年快乐!	2011-02-02	UR_1
新年快乐!!	2011-02-02	UR_2
新年快乐	2011-02-02	UR_3
新年快乐!	2011-02-02	UR_4
新年快乐啊	2011-02-02	UR_5

图 3 用户 UQ_0 经高相似同天同行为方法计算后的结果表

由图可知, Q 网站用户 UQ_0 共有两条推文产生高相似同天同行为, 对应 R 网站的 6 个候选匹配用户, 且高相似次数都为 1. 按照常理判断, 其最可能的匹配用户应该是 UR_0. 因为 UR_0 与 UQ_0 对应的高相似推文相对其它推文, 出现率较低. 且其余候选匹配用户的发推日期均相同, 推文互相之间也都是高相似的, 所以对应的推文极有可能描述的是同一热点事件. 如果不对该类推文进行处理, 算法随机选择匹配用户, 能准确识别的概率仅为 1/6, 但如果将该类推文记录删除, 算法就能准确识别 UQ_0 的匹配用户.

因此, 对热点事件进行处理很有必要. 首先需要明确热点事件的判定条件. 一般从网络角度而言, 所谓热点事件, 就是有大量网民都在同一时间段内发表有关该事件的推文、评论等. 所以, 对于推文 tw_z, 判定其属于热点事件的条件为, 由其产生的候选匹配用户数 $|CandidateUsers_z|$ 大于某个系数时, 即可判定该推文属于热点事件. 即:

$$|CandidateUsers_z| > \alpha \tag{3}$$

α: 热点事件系数 (取值将在后续实验中确定). 含义: 如果发表于日期 T_m 的推文 tw_z 产生多于 α 个候选匹配用户, 则该推文描述的内容属于热点事件.

当判定一条推文属于热点事件后, 还需判定其是否满足相应处理条件, 才能决定是否将其删除, 否则可能会降低用户重识别结果的准确率. 例如: 假设图 3 中的用户 UR_0 没有在 2011-03-31 发表如上推文, 而是也在 2011-02-02 发表了类似“新年快乐”的推文. 此时, 用户 UQ_0 的候选匹配用户还是 6 个, 都由推文“新年快乐! ”产生, 高相似次数也都为 1, 如果此时该推文满足热点事件的判定条件, 将对应推文记录删除, 则用户 UQ_0 没有了匹配用户. 很明显, 相对于删除前, 算法的

准确率反而有所下降. 因此, 对于热点事件进行处理还需满足相应的处理条件.

假设 Q 网站用户 UQ_i 经过高相似同天同行为方法计算后, 他在日期 T_0 到 T_m 发表的推文产生了高相似同天同行为, 且由这些推文产生的候选匹配用户对应的高相似次数分别为:

$$|HighSimilartyTotal_i| = \{\cdots HST_y \cdots\}$$

于日期$T_x(0 \le x \le m)$发表的推文 tw_z 产生的候选匹配用户对应的高相似次数分别为:

$$|HighSimilartyTotal'_i| = \{\cdots HST'_f \cdots\}$$

当以上高相似次数满足如下条件时, 推文 tw_z 对应的高相似记录应该被删除. 即:

$$[(\sum HST_y \ne \sum HST'_f)\&(\forall HST'_f = HST'_{f+1})] \quad (4)$$

以上式子中左侧括号 ( ) 表示, 由推文 tw_z 产生的候选匹配用户的高相似次数之和不等于用户 UQ_i 所有候选匹配用户的高相似次数之和, 即除了推文 tw_z 产生候选匹配用户外, 还有其他推文也产生了候选匹配用户; 右侧括号 ( ) 表示, 由推文 tw_z 产生的候选匹配用户的高相似次数两两相等. 当以上两个条件都成立时, 才可对推文 tw_z 的相应记录执行删除操作. 因为此时, 由推文 tw_z 产生的候选匹配用户的高相似次数均相等, 且除了该推文外, 还有其他推文也产生了高相似同天同行为, 所以执行删除操作后, 不仅提高了算法的识别准确率, 还有效降低了因这些干扰数据带来的系统损耗, 提升了系统的整体运行效率.

经过高相似同天同行为计算方法、热点事件处理方法操作后, 即可得到 Q 网站每个用户 UQ_i 与其候选匹配用户的高相似次数.

2.2.4 高相似多推文计算方法

Q 与 R 网站属于类似的社交网站, 用户经常在上面发表与生活、工作相关的推文. 因此如果两个网站的某用户属于同一个人, 则他们于各自平台上的推文在整体上会展现出相似的情感、用词习惯、行文风格等特征.

根据调研, Q 与 R 网站用户的人均推文数都较少, 大部分用户的推文数小于 50 条, 而每条推文的字数也很少, 所以一个用户的推文总字数也相对不多, 经统计, 一般在 500-2000 之间. 推文的风格与用户的性格、发推时的心情、兴趣爱好有很大关系, 所以推文的格式也比较零散、随意, 且每条推文所表示的含义也不尽相同. 因此, 如果将每个用户的所有推文看作一个整体, 很难通过提取一个明确的主题来表示用户. 而 word2vec 工具在训练词向量时, 充分的考虑了上下文, 所以在一定意义上, 词向量综合了该词多方面的因素. 因此, 根据推文特点, 本文将对用户所有推文的词向量进行累加, 将所得结果称为用户的多推文向量. 用户的相似度采用多推文向量的相似度表示, 具体计算方式如下所示.

假设 Q 网站用户 UQ_i 与 R 网站用户 UR_j 的推文集合分别为:

$$TWQ_i = \{\cdots twq_a \cdots | twq_a \in tw\}$$
$$TWR_j = \{\cdots twr_b \cdots | twr_b \in tw\}$$

则 UQ_i 与 UR_j 的多推文相似度为:

$$STSim(UQ_i, UR_j) = \frac{\sum V(twq_a) \times \sum V(twr_b)}{|\sum V(twq_a)| \times |\sum V(twr_b)|} \quad (5)$$

根据高相似次数、多推文相似度在重识别用户时所占的重要程度, 分别赋予不同的权重, 通过综合计算得到用户之间的相似度. 然后选择与用户 UQ_i 相似度最大的用户, 作为其匹配用户.

2.3 基于属性的用户重识别方法

用户属性往往包含着一个人的重要信息, 它在用户重识别领域扮演着重要角色. 本文将利用同时存在于 Q 与 R 网站的属性: 昵称、性别、生日、情感状态、家乡、所在地展开用户重识别研究. 其主要步骤如图 4 所示.

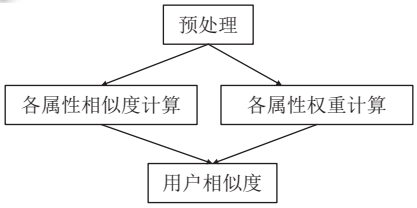

图 4 基于属性的用户相似度计算流程图

预处理: 将属性处理为统一格式, 如生日: yyyy-mm-dd、家乡: **省**市

各属性相似度计算:

(1) 昵称: 在 Q 与 R 网站中, 由于用户不同, 昵称的构成也不尽相同. 总体而言, 用户的昵称可根据所使用的语言划分为多类. 其中占比最大的是中文, 其次是英文, 只有极少部分用户使用其他语言. 为方便比较,

本文只考虑中文与英文昵称, 在具体情形下, 对昵称采用如图 5 的计算方式.

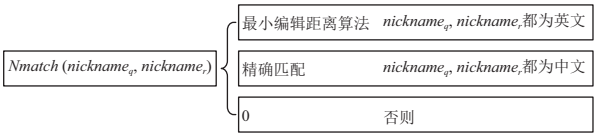

图 5 昵称相似度计算规则图

① 最小编辑距离算法: 最小编辑距离又称 Levenshtein[12]距离, 是指两个字符串之间, 由一个转换成另一个所需的最少编辑操作次数. 许可的编辑操作包括将一个字符替换成另一个字符, 插入一个字符, 删除一个字符. 一般来说, 编辑距离越小, 两个字符串的相似度越大.

② 精确匹配: 如果两个字符串完全相等, 即 $str_1 = str_2$, 则其相似度为 1, 否则为 0.

对于英文昵称, 由于其代表含义广泛, 且很多属于用户自创, 难以根据语义进行相似度度量. 因此本文采用流行的最小编辑距离算法计算其相似度. 对于中文昵称, 由于 Q 网站的用户昵称往往非真实姓名, 而 R 网站属于实名制社交平台, 用户昵称往往是真实姓名, 因此, 二者相似度可比较性较小, 本文采用精确匹配进行比较. 对于其它格式的昵称, 本文不进行比较, 将其相似度置为 0.

(2) 性别、生日、情感状态、家乡、所在地: 由于这 5 项属性均由用户通过下拉列表选择, 所以经过预处理后, 均具有统一的格式. 因此, 本文将采用精确匹配的方式计算其相似度.

各属性权重计算:

由于每个社交网站的定位不同, 所以其开放程度也不同. 而开放程度的不同将直接影响用户属性填写的完整程度. 如微博是开放的社交网站, 任何人均可访问其他用户的属性页面, 所以出于隐私保护的目的, 用户将生日、身份证号码等敏感属性选择空置或者隐藏, 因此这些属性的填写率很低; 而像如昵称、性别等与用户隐私程度关联不密切的属性, 填写率会相对较高. 据此猜想: 一个社交网站中, 如果用户的某项属性填写率越低, 则说明该属性的隐私程度越高, 在标识其身份的唯一性时, 所占权重应该越大. 根据此猜想, 用户各属性权重模型的构建, 可分为以下三个步骤:

① 选择社交网站用户数据集;

② 统计各属性填写率;

③ 计算各属性权重——填写率倒数和归一化.

表 1 以 R 网站的数据集 (共 10027 个用户) 为例, 计算各属性权重.

表 1 R 网站的属性权重计算表

属性	昵称	性别	生日	情感状态	家乡	所在地
用户填写数	6216	3919	3501	451	2529	1898
归一权重	0.0419	0.0664	0.0744	0.5772	0.1029	0.1372

表 1 中, 共包含 6 种属性, 第二行的数字代表填写了对应属性的用户数, 第三行表示经过计算后, 每种属性的归一权重. 其中, 各属性的归一权重计算方式如下:

① 填写率倒数和归一化, 即:

$$\left(\frac{10027}{6216}+\frac{10027}{3919}+\frac{10027}{3501}+\frac{10027}{451}+\frac{10027}{2529}+\frac{10027}{1898}\right)\times p=1$$

② 求得 p, 然后将属性填写率的倒数与 p 进行乘积计算, 结果即为该属性的归一权重.

如昵称: $\frac{10027}{6216}\times p$, 性别: $\frac{10027}{3919}\times p$

用户相似度:

根据以上内容, 可求得每对属性的相似度大小及各属性在标识用户身份唯一性时所占的权重. 则基于属性的用户相似度可表示为:

$$USim(UQ_i, UR_j)=\sum_1^6 (wt_q(p_{qx})\times wt_r(p_{rx})\times PSim(p_{qx}, p_{rx})) \quad (6)$$

其中, $USim(UQ_i, UR_j)$表示用户 UQ_i 与 UR_j 的相似度; p_{qx}、p_{rx} 分别表示 UQ_i 与 UR_j 的对应属性, $wt_q(p_{qx})$与 $wt_r(p_{rx})$分别表示属性 p_{qx}、p_{rx} 的权重; $PSim(p_{qx}, p_{rx})$表示属性 p_{qx}、p_{rx} 的相似度.

基于属性的用户重识别研究正是基于以上方法, 计算 Q 网站的每个用户 UQ_i 与 R 网站的所有用户之间的相似度, 然后将 UQ_i 的候选匹配用户按照相似度大小进行排序, 选择排序最高者作为 UQ_i 的匹配用户.

2.4 基于推文与属性相结合的用户重识别方法

前面章节分别叙述了基于推文、属性进行用户重识别的详细方法. 本节将推文与属性相结合, 共同进行跨社交网站的用户重识别研究.

在基于推文的用户重识别中, 如果两个来自不同社交网站的用户在同一日期发表的推文属于高相似同天同行为, 且推文内容与热点事件无关, 则他们很可能属于同一人, 而当这样的事件多次发生时, 则他们属于

同一人的概率几乎接近于 1. 而多推文相似度虽然在一定程度上可以代表用户相似度, 但也存在明显缺陷, 例如一个用户的 Q 网站推文很多, 而 R 网站推文很少, 则它们的多推文向量将相差很大, 所以较难取得准确匹配. 而在属性方面, 由于同时出现于两个网站的各属性在标志用户身份的唯一性时, 权重均很低, 且在这 6 项属性中, 很多用户会具有多项相同属性, 所以单独使用属性进行用户重识别研究也难以取得良好效果.

因此, 将推文与属性相结合进行用户重识别的研究, 高相似次数对匹配结果的准确率贡献很大, 而多推文相似度与基于属性的用户相似度贡献都较小. 因此, 当计算用户之间的相似度得分时, 本文将分别赋予它们 0.8、0.1、0.1 的权重, 以表示它们在衡量用户相似度时的贡献. 因此当高相似次数为 0 时, 用户之间的相似度理论上最大是 0.2, 这一值在衡量用户的相似度时, 说服力很小. 所以本文将只对用户之间的高相似次数大于 0 的用户对计算相似度得分, 然后将每个 UQ_i 用户的候选匹配用户按照分值大小进行排序, 选择排序最高者作为 UQ_i 的最终匹配用户.

3 实验结果与分析

3.1 数据集

为评估所提算法的效果, 本文以国内流行的社交网站 Q、R 作为实验对象. 首先对其进行数据采集, 并人工标注真实匹配用户对, 根据上述的用户重识别算法进行实验, 统计识别结果中真实匹配用户的对数, 以验证本文所提算法的可行性和有效性. 由于多推文相似度、基于属性的用户相似度在单独进行用户重识别时, 效果较差, 所以本文将它们与高相似同天同行为计算方法、热点事件处理方法相结合进行用户的重识别实验.

经调研, Q 网站的主要用户群是 18-28 岁的年轻人, R 网站则主要是大学生、研究生及部分白领等. 因此, 几乎每个 R 网站的用户均同时拥有 Q 网站的账户, 所以它们非常适合作为用户重识别实验的数据来源. 根据需求, 本文采集的数据主要包含两部分: (1) 推文信息; (2) 用户属性.

(1) 推文信息: 推文一般包含四类: 原创文字推文、原创多媒体推文、转发文字推文、转发多媒体推文. 一般而言, 原创推文在一定意义上唯一标识了用户身份, 因此在推文方面, 本文只采集原创文字推文与附带的发表日期.

(2) 用户属性: 在属性方面, 本文只采集同时出现于两网站的 6 种属性: 昵称、性别、生日、情感状态、家乡、所在地. 其中, 昵称由用户自创, 可包含图形、表情、文字、特殊符号等. 而其余 5 项均通过下拉列表选择填入, 具有固定的格式.

本文采用广度优先策略, 通过爬虫进行数据采集. 从种子用户开始, 首先抓取其自身数据, 再抓取其好友的数据, 然后再抓取其好友的好友数据, 通过该种子用户不断的向外延伸, 访问不同的用户. 抓取的数据总量如表 2 所示, Q 网站用户共抓取了 16173 个, 以及这些用户的 300 余万原创文字推文, R 网站用户共抓取了 10027 个, 推文约 75 万. R 网站推文数较少的原因可能有两点: (1) 近些年, R 网站业绩不断下滑, 导致用户使用率较低; (2) 用户只在上学时使用 R 网站, 一般年限为 3-7 年, 而之后便不再使用. 在 R 网站中, 大部分用户的主页是开放的, 任何人均可访问. 而在 Q 网站中, 大部分用户的空间设置了访问等级, 如只对自己开放、只对其好友开放等, 所以对于种子用户, 其好友的空间往往可以访问, 而其好友的好友空间, 只有部分可以被访问, 所以在采集的数据中, 真实匹配用户数量较少, 仅有人工标注的 776 对.

表 2 数据集总量统计表

属性	Q网站	R网站	真实匹配用户
用户数	16173	10027	776
推文总数(原创)	3019640	745341	-

Q 与 R 网站用户的单条推文长度与推文数一般都较小. 如由图 6 可知, 长度为 10-20 个字的推文占比最大, 且随着长度的增加, 相应推文逐渐减少, 当推文长度大于 70 时, 相应推文变多, 其原因是此统计数包含了长度大于 70 的所有推文. 因此可知, Q 与 R 网站推文大部分都是短文本, 不适合使用 LDA 等模型表示推文, 所以本文以词向量累加的方式计算推文向量. 此外, Q 网站推文的平均长度均大于 R 网站推文, 其原因可能是: 相较于 R 网站, Q 网站私密性更强, 用户更愿意将推文发表于 Q 网站. 由图 7 可知, 推文数小于 50 条的用户占比最大, 且随着推文数的增加, 相应用户逐渐减少. 经过对两图的综合统计可知, 一个用户的原创推文总字数一般在 500—2000 左右. 且这些推文包罗万象, 没有明确主题, 所以很难通过对这些文字的总体建

模, 实现良好的用户重识别效果. 因此本文选择对单条推文进行研究, 以克服这一困难.

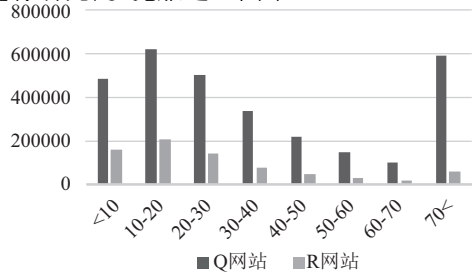

图 6 每条推文字数统计图

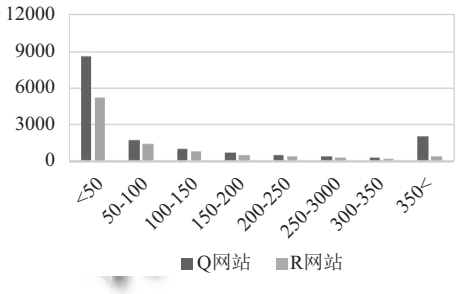

图 7 每个用户推文数统计图

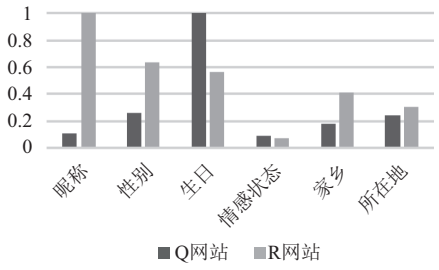

图 8 用户属性填写率

由于涉及隐私, 导致每个用户对待属性的态度不同, 因此填写情况也不同. 图 8 是用户属性填写率的统计图. 由于网站原因, 爬虫只获取到 Q 网站的部分用户昵称, 所以导致其填写率很低, 而实际每个用户均有昵称, 其填写率本该为 1. 对于 Q 网站的用户生日, 网站默认将未填写的用户生日置为 1970-01-01, 所以其填写率在统计时为 1. 由于基于属性的用户重识别方法核心是计算每对属性的相似度, 其可计算性由该对属性中填写率最低的那一项决定, 而由图可知, 各项属性的最低填写率都很低, 且这 6 项属性都难以标识用户身份的唯一性. 因此只通过属性进行用户重识别的研究很难取得良好效果.

3.2 实验结果

根据热点事件的原理可知, 热点事件系数是独立的, 当数据集越大, 得到的值越准确. 因此, 本文使用全部数据集进行实验. 经实验发现, 当热点事件系数取值为 4 时, 准确识别数取得最大值, 因此, 本文将热点事件系数取值为 4.

为确定相似系数的取值, 本文分别选取包含 100、500 对真实匹配用户的数据集, 使相似系数 S 从 1 以 0.01 的幅度依次递减至 0.90 进行实验, 观察经多种处理后的准确识别数的变化趋势, 实验结果如图 9、10 所示. 在图中, 高相似表示只经过高相似同天同行为方法计算后的结果; 热点表示经过高相似、热点事件处理后的结果; 多推文表示经过高相似、热点、高相似多推文计算后的结果; 属性表示经过高相似、热点、多推文、基于属性的相似度计算后的结果.

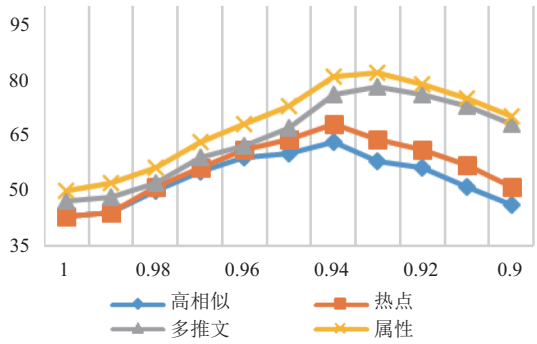

图 9 多种处理后准确识别数变化图 (100 对用户)

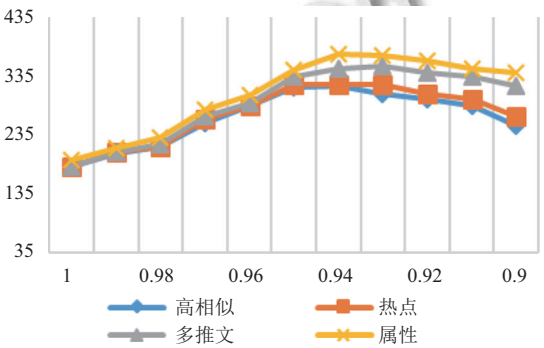

图 10 多种处理后准确识别数变化图 (500 对用户)

由图 9、10 可知, 当数据集分别包含 100、500 对真实匹配用户, 相似系数取值 0.93、{0.94, 0.93}时, 准确识别数都达到了最大值, 说明相似系数的取值不会随着数据规模的扩大而发生显著变化, 均能在其取值为 0.93 时, 使的准确识别数达到最大值. 因此, 本文的相似系数 S 取值为 0.93.

由上图还可得知, 当相似系数取值不低于 0.96 时, 热点事件处理方法、高相似多推文计算方法对准确识

别数的提升效果较小, 但当相似系数小于 0.96 时, 它们对准确识别数的提升效果明显, 尤其是多推文处理. 说明它们对重识别的效果有着良好的影响. 属性计算方法在相似系数变化的整个过程中, 一直对重识别的结果有着积极作用.

在确定了各项系数的取值后, 本文与 Goga[3]的方法进行了对比. Goga 根据推文的地理位置、发表时间、内容风格分别进行了用户重识别研究, 由于本文数据不包含推文的地理位置, 因此本文与 Goga 均只综合其余两项特征完成实验. 实验结果如图 11 至图 14 所示. 图中本文方法简记为“HS”, Goga 方法简记为“Goga”.

图 11、12 是设定了相似系数 S=0.93, 准确率和召回率随高相似次数 HST 的变化图.

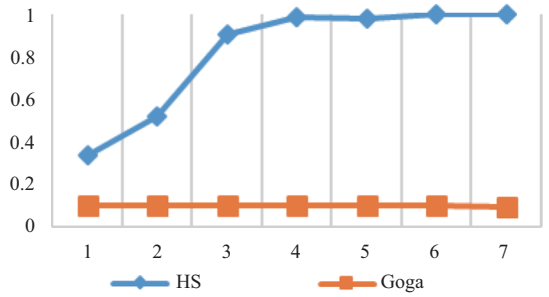

图 11 准确率随高相似次数 HST 的变化图 (S=0.93)

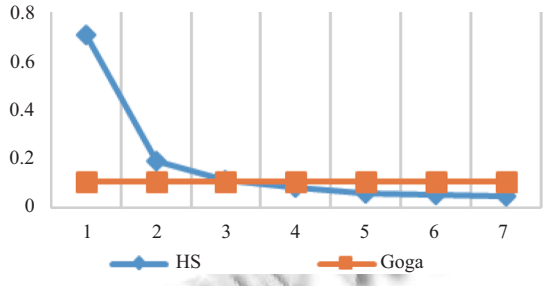

图 12 召回率随高相似次数 HST 的变化图 (S=0.93)

由图 11、12 可知, 当相似系数取值 0.93 时, 本文方法的准确率迅速上升, 当高相似次数 HST=3 时, 准确率达 90%, 当高相似次数 HST≥6 时, 准确率达到了 100%, 明显优于 Goga. 但随着高相似次数 HST 的不断增加, 本文方法的召回率也明显下降, 当高相似次数 HST=1 时, 召回率最高, 达到了 70.12%, 此后下降趋势逐渐缓和. 当高相似次数 HST≥4 后, 召回率低于 Goga.

图 13、14 是设定了高相似次数 HST≥3, 准确率与召回率随相似系数 S 的变化图.

由图 13、14 可知, 当高相似次数 HST≥3, 本文方法的准确率保持较高, 均达到了 90% 以上, 当相似系数 S≥0.98 时, 准确率达到了 100%, 此后随着相似系数的减小, 准确率也缓慢下降, 整个变化过程明显优于 Goga. 本文方法的召回率随着相似系数的增大一直趋于上升趋势, 当相似系数 S≥0.94 时, Goga 的召回率优于本文方法, 随着相似系数的减小, 当 S≤0.93 后, 本文方法的召回率优于 Goga.

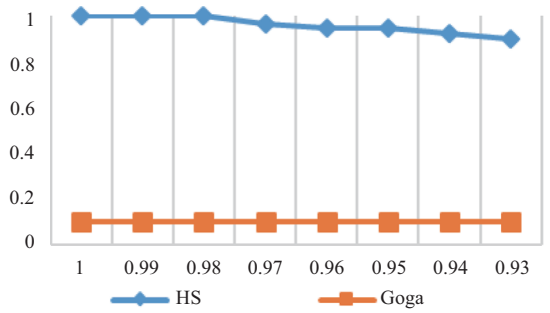

图 13 准确率随相似系数 S 的变化图 (HST≥3)

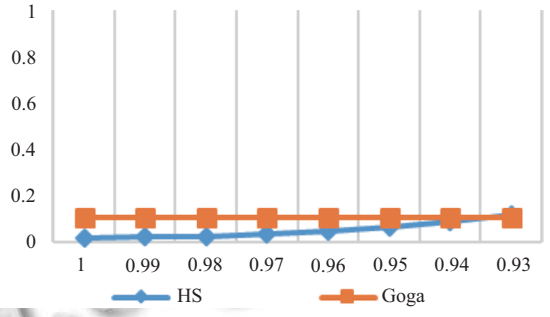

图 14 召回率随相似系数 S 的变化图 (HST≥3)

4 结语

针对社交网络领域的个人隐私保护问题, 本文提出了一个基于推文与属性的高相似同天同行为用户重识别算法, 其提出的多种方法可有效提高算法的准确率. 经过 Q 与 R 网站的 1 万多用户、300 多万推文进行实验评估, 该算法的准确率为 33.84%, 召回率为 70.12%, 明显优于 Goga 的方法. 且该算法可实现用户的精确重识别. 实验还揭示了当用户在不同网站发表相近或相同的内容达到 3 次及以上时, 可以为攻击者提供足够的信息, 将其不同网站的账户相关联, 从而导致更多的隐私被泄露. 本文的研究方法有多个应用领域: (1) 隐私安全研究; (2) 广告精准投放; (3) 社交网站好友推荐等.

参考文献

1 Vosecky J, Hong D, Shen VY. User identification across multiple social networks. Proc. of the 1st International Conference on Networked Digital Technologies. Ostrava, Czech Republic. 2009. 360–365.

2 Goga O, Perito D, Lei H, *et al*. Large-scale correlation of accounts across social networks [Technical Report]. TR-13-002. Berkeley, California, USA: International Computer Science Institute, 2013.

3 Goga O, Lei H, Krishnan SH, *et al*. Exploiting innocuous activity for correlating users across sites. Proc. of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil. 2013. 447–458.

4 Cecaj A, Mamei M, Bicocchi N. Re-identification of anonymized CDR datasets using social network data. Proc. of the 2014 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops). Budapest, Hungary. 2014. 237–242.

5 Narayanan A, Shmatikov V. De-anonymizing social networks. Proc. of the 30th IEEE Symposium on Security and Privacy. Washington, DC, USA. 2009. 173–187.

6 Zhou XP, Liang X, Zhang HY, *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. [doi: 10.1109/TKDE.2015.2485222]

7 Bartunov S, Korshunov A, Park ST, *et al*. Joint link-attribute user identity resolution in online social networks. Proc. of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis. Beijing, China. 2012.

8 Kong XN, Zhang JW, Yu PS. Inferring anchor links across multiple heterogeneous social networks. Proc. of the 22nd ACM International Conference on Information & Knowledge Management. San Francisco, California, USA. 2013. 179–188.

9 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)-Regular Papers and Special Section on Intelligent Healthcare Informatics, 2015, 6(4): 49.

10 Mihalcea R, Corley C, Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity. Proc. of the 21st National Conference on Artificial Intelligence. Boston, Massachusetts, USA. 2006, 1. 775–780.

11 Islam A, Inkpen D. Semantic text similarity using corpus-based word similarity and string similarity. ACM Trans. on Knowledge Discovery from Data, 2008, 2(2): 10.

12 Levenshtein VI. Methods for obtaining bounds in metric problems of coding theory. Proc. of the IEEE-USSR Joint Workshop on Information Theory (Moscow, 1975). New York, USA. 1976. 126–143.