基于自然标注的跨平台虚拟账号关联方法研究

季　铎[1]　敬少杰[1]　彭如香[2]　孔华锋[3]

[1](中国刑事警察学院　辽宁 沈阳 110854)

[2](公安部第三研究所　上海 201204)

[3](武汉商学院　湖北 武汉 430056)

摘　要　随着大数据时代的到来,跨平台虚拟账号的关联成为网络监管领域亟待解决的问题。该文以微博、微信等用户文本数据为研究对象,通过对数据的抽样和人工标注,开展开放式社交平台中跨平台账号自然标注行为的量化分析,并由此提出基于用户自然标注的跨平台虚拟账号的关联方法。该方法针对自然标注特点,构建基于上下字词特征的虚拟账号识别的模型,并利用二分类的深度学习模型进行昵称和用户的同一认证,最终实现对跨平台虚拟账号的识别,识别准确率达到 85% 以上。

关键词　虚拟账号　自然标注　账号关联

中图分类号　TP319　　**文献标志码**　A　　DOI:10.3969/j.issn.1000-386x.2024.09.027

CROSS-PLATFORM VIRTUAL ACCOUNT ASSOCIATION METHOD BASED ON NATURAL ANNOTATION

Ji Duo[1]　Jing Shaojie[1]　Peng Ruxiang[2]　Kong Huafeng[3]

[1](*Criminal Investigation Police University of China*, *Shenyang 110854*, *Liaoning*, *China*)

[2](*The Third Research Institute of Ministry of Public Security*, *Shanghai 201204*, *China*)

[3](*Wuhan Business University*, *Wuhan 430056*, *Hubei*, *China*)

Abstract　With the advent of the era of big data, the association of cross-platform virtual accounts has become an urgent problem in the field of network supervision. This paper took user text data such as Weibo and WeChat as the research object. Through sampling and manual labeling of the data, a quantitative analysis of the natural labeling behavior of cross-platform accounts in the open social platform was carried out. And from this, a cross-platform virtual account association method based on the user's natural annotation was proposed. Aiming at the characteristics of natural labeling, this method constructed a virtual account recognition model based on the characteristics of upper and lower words, used a two-category deep learning model for the same authentication of nicknames and users, and realized the recognition of cross-platform virtual accounts with the recognition accuracy rate reaching more than 85%.

Keywords　Virtual account　Natural annotation　Account association

0　引　言

随着互联网的发展,尤其是移动互联网的普及,社交网络平台成为最具影响力的互联网应用类型之一。目前,网络中的社交平台众多,而大多数社交网络账号的获取通常采取匿名机制,导致违法犯罪人员真实身份关联分析难度加大,而解决这些问题的关键是掌握相关匿名账号背后的真实用户的有效信息。而单一社交网络中用户数据有限,如果能够识别出用户的多个社交账号,就可以更加全面地掌握用户信息,从而对发布有害信息的用户进行规制和管控。因此需要从技术上解决跨平台虚拟账号的关联问题,提高管理效率。

虚拟账号的关联成为近年来自然语言处理、数据

收稿日期:2021 - 05 - 26。国家重点研发计划项目(2018YFC0830401);辽宁网络安全执法协同创新中心开放课题。**季铎**,副教授,主研领域:自然语言处理。**敬少杰**,硕士生。**彭如香**,副研究员。**孔华锋**,研究员。

分析领域的热点问题,特别是其所包含的巨大商业潜力和现实挑战,获得了国内外专家学者越来越多的关注[1]。虚拟账号关联的研究,就是为了实现一种能够将不同网络社交平台中属于同一现实用户的虚拟账号身份关联在一起的方法。

总的来说,当前在虚拟账号关联领域的研究中使用的方法可以分为两大类:基于文本内容的账号关联和基于用户关系的账号关联。基于文本内容的虚拟账号关联的方法主要思想是通过自然语言的处理抽取文本中大量的用户特征来构造用户画像进行分析,但由于用户特征的提取是一个复杂繁琐的过程,而且通过传统方法提取到的特征并不一定能够具有很强的表现力,从而关联分析得出的用户信息的准确性不能得到保证。虽然有学者引入深度学习的方法进行文本信息处理,挖掘用户的特征来构建用户画像,通过用户画像去计算用户间的相似性来实现虚拟身份关联。这种方式在一定程度上改进了传统算法的缺陷,但实现用户画像所要求的数据信息属性众多,且最后得出的实验结果也并不理想。例如,闫洲等[1]在中使用基于深度学习的神经网络的分布式处理方法进行社交网络中虚拟账号的关联,虽然初步得到了一定效果,但是模型的训练时间较长,这个时间复杂度在研究过程和实际应用中不太能够让人接受。

基于文本内容的虚拟账号关联的另一种方法是基于用户昵称进行关联,虽然有些用户不会在不同的网络社交平台上注册时使用相同的用户昵称,但是大多数人在不同平台上的昵称具有一定的相似性和关联度,且有些人会在每个社交平台上将自己其他社交平台的昵称或用户名展示出来以求得更多的关注。因此,基于用户昵称的虚拟账号关联的方法相对来说具有更高的可行性和研究价值。

近些年来有学者针对跨平台账号的身份识别,提出其本质就是找出多个虚拟账号背后的实体用户,进而提出利用复杂网络具有的网络特性可以链接不同社交网络中的节点,并分析节点之间存在的联系,结合相关的匹配算法可以有效地识别出用户在不同社交网络上的虚拟账号[2]。这种方法就是通过用户关系来开展账号关联研究,这些研究普遍基于例如“共同好友”等的简单逻辑,或者图论中的子图等匹配方法,但是这些方法忽略了利用多维度信息时间计算的复杂性,要求数据信息属性众多,且并不能够完全适用于跨平台的异构信息网络,难以达到良好的跨社交网络平台用户身份关联的效果,最终得出的账号关联准确度也不高。例如,齐林峰等[3]在中提出,针对两个不同平台的用户节点,计算其邻居节点中属于跨平台关联的节点对数量,然后与指定阈值进行对比,若超过阈值则判定这两个用户属于同一自然人,但是如果不同的社交媒体的两个账户具有相似的属性但没有关联节点,则无法通过此方法进行匹配,这就是这种方法的局限性所在。

根据手工标注取样分析,以微博用户简介和所发布内容中出现的微信号或微信名为关键字搜索相关用户,从而找到对应用户的微信或公众号,得出通过微博查找对应用户的微信具有可行性。本文基于自然标注的角度,摆脱传统思路的束缚,克服传统匹配算法的缺点,同时引入机器学习的思想方法,研究跨平台虚拟账号关联的可行性和优势,并通过相关实验进行论证和探讨,从而为公安机关对不同平台的侦查目标进行进一步收集信息和分析取证提供一种思路和理论支持。

1　基于自然标注的数据挖掘

自然标注,是指由互联网使用者根据其自身目的,通过语言文字的处理和文本结构的识别等方式,对各种互联网资源进行的一定程度的“不自觉”的手工标注,而同时用户本人并没有意识到这一点。例如,网页上的空格、标点符号、句子开头和结尾就是所谓的自然标注。自然标注最早提出并应用于自然语言的处理中,计算机语言学家们可以将这些标注自觉地、系统性地应用在自然语言处理和文本数据挖掘的各种研究中。

自然标注应用于文本数据挖掘,主要是用来界定时间、地点、人物及其属性等内容在文本中的位置和范围[4],也就是将这些信息标注并挖掘出来。网页中具有特殊含义的或是具有明确指示性作用的短语,如用户名、时间、正文、下一篇、评论等都可以作为自然标注,从而界定出关键重要信息。自然标注的含义和布局通常是人们理解文本、提取有效信息的基本出发点。

自然标注的概念早已被提出,并且这些年被广泛应用在数据挖掘、科学研究和实际工作中。例如,基于自然标注的微博文本对消费者的消费意图进行识别,既可用于电子商务公司挖掘用户当前需求,又有助于针对社会媒体富有价值的用户提供广告宣传,在产品策划、设计和营销过程中做到有的放矢。再如,基于自然标注进行文本类别分类,极大地优化了传统文本分类算法,提高了传统文本分类的效率。还如,基于自然标注进行网页信息抽取和挖掘,可以正确提取网页正文内容,方便普通网民阅读;可以正确抽取用户评论正

文数据和相关评论发表的元数据,以便进一步进行舆情分析;还可以抽取网页中的商品信息,以便进行竞争情报比较分析。

关于跨平台虚拟账号关联的自然标注,是通过对给定平台中某虚拟账号的简介和所发布内容等信息的分析,从中找到与该账号有关的其他平台的虚拟账号信息并标注出来。这种“自然标注”对于我们发现和分析目标账号关联线索具有很大的研究意义。

2 方法设计

2.1 问题的提出

从信息共享的方式上来看,网络上各大平台可分为信息开放式和封闭式两大类。信息开放式平台主要侧重于信息分享、传播以及获取,主要以微博、知乎等平台为代表。而信息封闭式平台主要侧重于用户关系,重在用户与用户间的互动,如微信等即时通讯类的平台。如图 1 所示,相比于封闭平台,开放平台中的用户更希望进行信息分享,包括跨平台相互间的信息分享。因此,在开放类平台中会存在大量有关同一用户跨平台的信息内容。

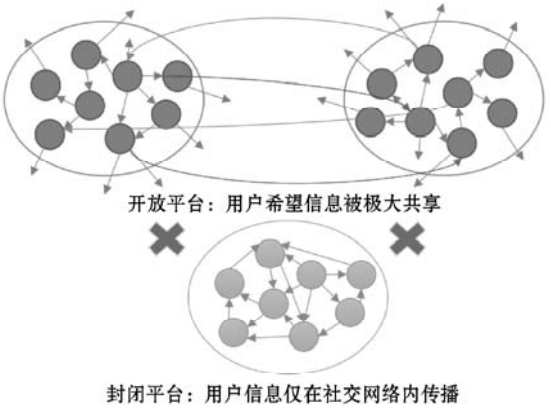

图 1 各类平台信息共享示意图

微博是一种典型的开放类信息分享平台,用户之间是点对面的非对等关系,每个微博用户都追求的是受众的广度。具有多个开放平台账号的用户,也更趋向于进行多账号的推广。如图 2 所示,微博中用户会推广微信公众号等信息。

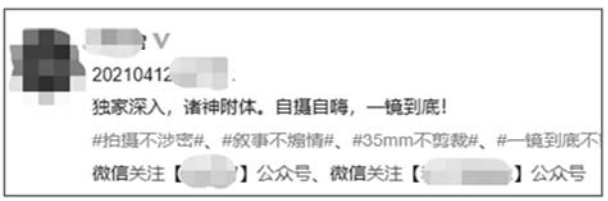

图 2 微博用户推广微信公众号信息示例

为了进一步分析开放平台中用户的推广其他跨平台账号现象,我们分三次随机从微博平台中获取了 1 000 人的微博信息,并在该数据中人工抽取其跨平台的账号,实验结果如表 1 所示。从该数据可以发现在开放平台中,发布跨平台账号信息的用户平均在 75%。由此,利用此类信息进行虚拟账号的关联是一种有效的思路。

表 1 微博文本人工标注实验结果

组号	取样用户数	关联到的跨平台用户数	关联率/%
1	387	295	76.2
2	356	263	73.9
3	257	192	74.7

通过分析数据,用户为了突出发布账号信息,会在发布时利用明显的标识进行标记,这类似于自然标注的信息。因此,本文采用基于自然标注的方法进行跨开放平台的虚拟账号关联。

2.2 虚拟账号的关联方法

虚拟账号的关联方法主要包括虚拟账号识别和识别账号与用户同一认证两个过程,如图 3 所示。其中,虚拟账号识别主要采用基于 CRF 的序列标注方法,通过对识别结果进行规则优化保证识别准确率;账号与用户同一认证采用分类方法,通过抽取识别账号的上下文构建特征,进而利用二分类模型进行同一身份的认证。

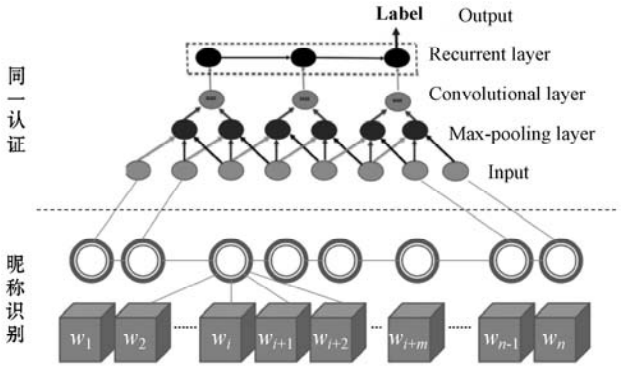

图 3 虚拟账号关联方法的两个过程

2.2.1 虚拟账号的识别

由于数据量相对较少,无法充分进行深度网络模型的训练。因此,考虑采用基于条件随机场(CRF)的序列标注模型,该模型是一种无向图模型,近年来在分词、词性标注和命名实体识别等序列标注任务中取得了很好的效果。

条件随机场模型的特征选择中,主要基于何径舟等的研究结果,采用基于窗口 L(待标注词语的左右 L 个词语)的特征进行标注特征的选择。特征的分类说明如表 2 所示。

表 2　特征的分类说明

词法级特征	
独立特征	$w_i(-L<i<L)$
	$p_i(-L<i<L)$
联合特征	$w_iw_{i+1}(-L<i<L-1)$
	$p_ip_{i+1}(-L<i<L-1)$
位置相关特征	
位置 + 独立特征	$i{:}w_i(-L<i<L)$
	$i{:}p_i(-L<i<L)$

其中 w_i 代表词或字特征,p_i 代表词或字的类别特征。

2.2.2　基于多策略的用户账号同一认定

在用户自然标注的文本中识别出用户账号后,还要进一步进行账号与用户同一身份的认证判别。论文采用基于多策略的账号同一认证,分别考虑分析账号所在的上下文信息和用户名的相似性和关联性,从而判断账号与用户的同一性。

其中,基于内容的虚拟账号同一身份认证问题可以转换为二分类问题,即通过文本中虚拟账号上下文特征建立起两种特征的分类模型(例如将同一性特征类别记为 1,非同一性特征类别记为 -1),让机器学习该分类模型,并进行测试,从而机器能够进行账号同一性自动判别;基于用户名的用户账号同一认定则是采用编辑距离的方法进行虚拟账号和用户名的相似度计算,最终系统将识别结果进行加权融合实现多策略的用户账号同一认定。用户账号同一认证的流程如图 4 所示。

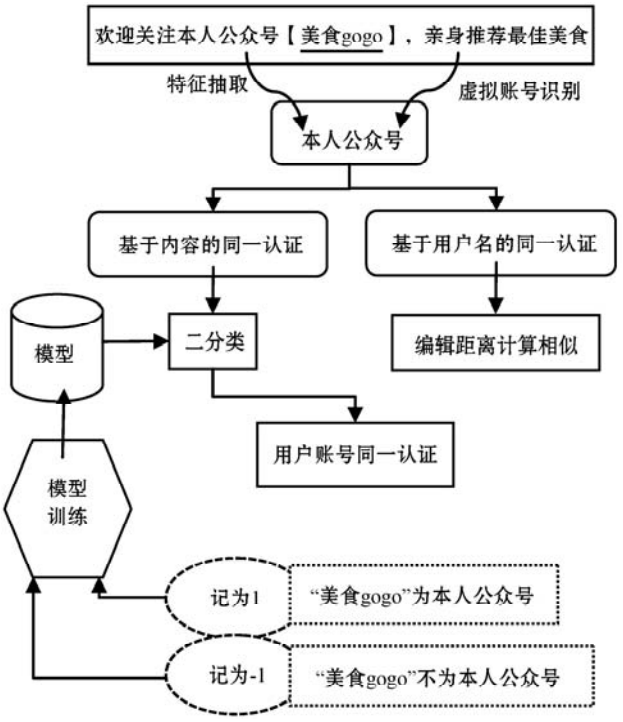

图 4　用户账号同一认定流程

3　实验及讨论

3.1　实验数据

实验数据通过微博平台的数据接口获取用户的基本数据,通过数据去重和过滤后共计 11 081 条。采用人工标注方式对昵称、手机号、微信号、QQ 号等该用户虚拟账号信息进行标注,共标记用户账号信息 XXX 条。数据集采用传统的 BIOE(B-begin、E-end、I-insert、O-other,分别代表开始、结束、中间、其他字符)五类标签对文本进行标注。

3.2　实验方法

3.2.1　基于 CRF 的昵称账号识别

实验采用随机 8∶2的数据划分方法,即 80% 的语料作为模型训练,20% 作为系统测试语料。为验证条件随机场对昵称识别的效果,本文随机构建了 4 组测试样本集,并采用准确率作为实验的评价指标,实验结果如表 3 所示。

表 3　基于 CRF 的昵称账号识别实验结果(%)

方法	实验 1	实验 2	实验 3	实验 4	平均
CRF	71.59	73.31	72.14	74.62	73.41
CRF + RUL	72.68	73.98	72.01	75.96	73.66

其中 CRF 代表直接用 CRF 进行识别,RUL 方法代表基于规则的方法,CRF + RUL 代表基于 CRF 和规则后处理的方法。从实验结果看,本文所描述利用 CRF 对命名实体识别的准确率可达到 73% 以上,每次实验结果都相对稳定。

3.2.2　用户同一认证

根据实验结果进行人工取样分析,发现通过 CRF 识别出的昵称账号不一定是对应用户的同一账号,因此下一步的实验任务是进行用户同一认证。

实验首先依然采用人工标注的方式对批量数据进行用户同一认证标记,包括微博用户的简介和微博内容,然后利用基于 fasttext 的机器学习方法处理数据,对文本数据进行用户同一认证的训练和测试。使用 fasttext 进行机器学习时,通过对三个参数 wordNgramss(1 -5)、epochs(5 -55)、lrs(0.1 -1.0)进行不同赋值,分别测试准确率和召回率,最后发现实验结果指标最高的参数为 wordNgrams = 1,epoch = 10,lr = 0.1。实验结果的平均准确率稳定在 79.5% 左右。

从上述实验结果可以看出用户同一认证的准确率并不是很高,因此考虑从以下几个方面对实验过程进行优化:首先,将每一条待处理的微博文本信息按照人类一般理解注明标签为“备注”或“内容”,并将标注为“备注”的用户简介信息中出现的昵称账号(如图 5 所示)全部标注为用户本人账号;其次,将文本中识别出的昵称账号与用户的昵称账号进行相似度对比,相似度高于 50% 的(如图 6 所示)都标注为用户本人账号;最后,在数据预处理时,将停用词(Stop Words)进行排除操作,从而减小或消除 Stop Words 对实验结果的影响。

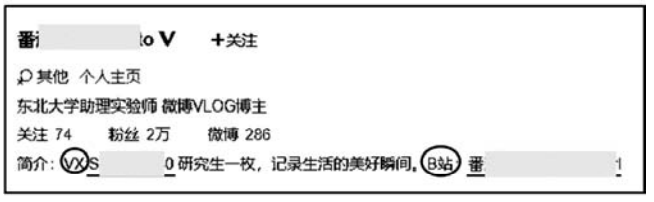

图 5 用户简介信息中出现的昵称账号

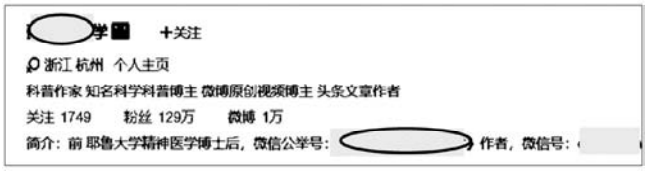

图 6 与用户昵称账号相似度高于 50% 的昵称账号

通过以上的优化操作,再次进行用户同一认证实验。我们发现实验结果指标最高值对应的参数为 wordNgrams = 2, epoch = 10, lr = 0.1。同时,通过实验优化,用户同一认证的准确率得到了较大的提升,平均值能够达到 86.3%。

优化前后选择不同参数进行实验的部分结果如图 7 所示。

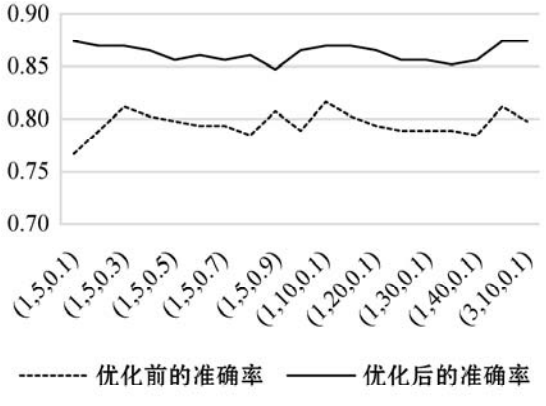

图 7 优化前后选择不同参数的部分实验结果

选择好优化后实验的最佳参数后,对数据进行多次随机的 8∶2 划分(随机选择 80% 的数据作为训练集,剩下的 20% 数据作为测试集),利用 fasttext 进行机器学习。从实验结果来看,对用户进行同一认证的平均准确率趋向于 85.8%,每次实验都相对稳定。用户同一认证的多次实验结果如表 4 所示。

表 4 基于 fasttext 的用户同一认证实验结果(%)

实验	实验 1	实验 2	实验 3	实验 4	平均
准确率	86.30	85.95	82.71	88.33	85.82

4 结 语

本文对命名实体识别工作进行了探究,并以微博用户的文本数据为研究对象,建立了 CRF 特征模板。然后采用了 CRF 对微博命名实体进行识别,从实验结果来看,随机选取不同的数据,识别出昵称或账号的结果都稳定趋于 73.7%。在识别出命名实体后,又对用户的同一性进行了初步验证,准确率平均在 79.5% 左右。对实验过程进行多重优化后,准确率能够达到 85.8%。但对于一些没有特征的命名实体,很难有效地识别。因此,未来的研究工作需要着重于研究特征不明显的命名实体,可以从语义的方面入手,对文本数据进行语义的分析,以期获得更好的识别效果。

参 考 文 献

[1] 闫洲. 基于深度学习的多社交网络中虚拟身份关联技术研究[D]. 长沙:国防科学技术大学.

[2] 邢玲,邓凯凯,吴红海,等. 复杂网络视角下跨社交网络用户身份识别研究综述[J]. 电子科技大学学报,2020,49(6):108 - 120.

[3] 齐林峰. 利用实体解析的跨社交媒体同一用户识别[J]. 图书情报工作,2017(6):107 - 114.

[4] 李志义,沈之锐. 基于自然标注的网页信息抽取研究[J]. 情报学报,2013,32(8):853 - 859.

[5] 罗梁,王文贤,钟杰,等. 跨社交网络的实体用户关联技术研究[J]. 信息网络安全,2017(2):51 - 58.

[6] 赵东生. 跨社交网络用户身份识别算法研究[D]. 杭州:杭州电子科技大学.

[7] 孙波,张伟,司成祥. 社交网络用户身份关联及其分析[J]. 北京邮电大学学报,2020,43(1):126 - 132.

[8] 付博,陈毅恒,邵艳秋,等. 基于用户自然标注的微博文本的消费意图识别[J]. 中文信息学报,2017,31(4):208 - 215.

[9] 张亮. 基于自然标注的文本分类[D]. 哈尔滨:哈尔滨工业大学,2015.