基于主题模型的社交网络匿名用户重识别

吕志泉[1] 李 昊[2] 张宗福[2] 张 敏[2]
(国家计算机网络应急技术处理协调中心 北京 100029)[1]
(中国科学院软件研究所可信计算与信息保障实验室 北京 100190)[2]

摘 要 近年来,社交网络已成为人们日常生活的一部分。社交网络在为人们的社交活动带来便利的同时,也对个人隐私造成了威胁。通常情况下,人们都希望对自身的部分私密社交活动信息进行保护,以阻止亲属、朋友、同事或其他特定群体的访问。较为常见的一种保护措施是以匿名方式进行社交。一些社交网络会为用户提供匿名机制,允许用户以匿名的形式进行部分社交活动,从而将这部分社交活动与主账号分隔开,以达到隐私保护的目的。此外,用户也可以创建额外的账号(小号),并将该账号的属性、朋友关系与主账号进行区别。针对这些保护措施,文中提出了一种基于主题模型的社交网络匿名用户重识别方法。该方法将用户匿名方式(或小号)和非匿名方式(主账号)发布的文本内容进行主题挖掘,并在主题模型的基础上引入时间因素和文本长度因素来构建用户画像,最后通过分析匿名(小号)和非匿名(主账号)用户画像之间的相似度来实现用户身份的重识别。在真实社交网络数据集上的实验表明,该方法能够有效地对社交网络匿名用户或"小号"用户实施身份重识别攻击。

关键词 大数据,社交网络,隐私保护,匿名,身份重识别
中图法分类号 TP309 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.06.021

Topic-based Re-identification for Anonymous Users in Social Network

LV Zhi-quan[1] LI Hao[2] ZHANG Zong-fu[2] ZHANG Min[2]
(National Computer Network Emergency Response Technical Team & Coordination Center of China, Beijing 100029, China)[1]
(Department of TCA, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)[2]

Abstract Social network has become part of people's daily life recently, and brings convenience to our social activities. However, it poses threats to our personal privacy at the same time. Usually, people want to protect part of their private social activity information to prevent relatives, friends, colleagues or other specific groups from visiting. One common protective method is to socialize anonymously. And some social networks provide anonymity mechanisms for users, allowing them to hide some private information about social activities, thus separating these social activities from the main account. In addition, users can create alternate accounts and set different attributes, friendships to achieve the same aim. This paper proposed a topic-based re-identification method for social network users to make an attack on these protection mechanisms. The text contents published by anonymous users (or alternate accounts) and non-anonymous users (main accounts) are analyzed based on topic model. And the time factor and text length factor are introduced to construct user profiles in order to improve the accuracy ofthe proposed method. Then the similarity between anonymous and non-anonymous user profiles is analyzed to match their identities. Finally, experiments on real social network dataset show that the proposed method can effectively improve the accuracy of re-identification for users in social networks.

Keywords Big data, Social networks, Privacy protection, Anonymity, Re-identification

1 引言

近年来,随着智能手机、平板电脑等移动终端的快速普及,社交网络得到了极大的发展。截至2016年,全球社交网络的用户数已高达23.4亿,占全球总人口的32%[1)]。人们在社交网络中的一举一动都是其政治倾向、宗教信仰、兴趣爱

1) www.qianzhan.com/analyst/180125-261e1b66.html

到稿日期:2019-02-21 返修日期:2019-04-15 本文受国家自然科学基金(61402456)资助。

吕志泉(1986—),男,博士,主要研究方向为网络与系统安全;李 昊(1983—),男,博士,副研究员,CCF会员,主要研究方向为数据隐私保护、可信计算、访问控制,E-mail:lihao@iscas.ac.cn(通信作者);张宗福(1991—),男,硕士,主要研究方向为数据隐私保护;张 敏(1975—),女,博士,研究员,主要研究方向为数据隐私保护、可信计算。

好、价值观等个性化信息的综合反映。然而,每个用户在满足自己社交需求的同时,并不希望上述所有信息都被人知晓[1]。针对这些用户隐私需求,部分社交网站给出了匿名机制,即用户可以发布匿名的文章或以匿名方式参与话题讨论等社交活动。这些社交活动信息仅能够被该帐号自身查看,其他社交用户无法获知这些匿名社交活动信息的属主,也无法获知它们是否属于同一个用户。还有一部分社交网络没有提供匿名机制,但用户可以选择注册小号来实现另一种形式的"匿名",即在需要保护个人隐私时使用小号参与这部分社交活动,从而实现与主账号社交活动的隔离。据日本广播公司 NHK 最近的调查结果[1),34%的日本年轻人在社交网络中有秘密的"小号"。他们把不愿意和亲朋好友透露的烦恼,或自己面对的社交压力通过"小号"抒发出来。

但是,采用匿名或小号的方式并不能完全割裂用户的兴趣爱好、宗教信仰、社交关系等个性化信息。例如,2017 年 NBA 总裁亚当-萧华的 Twitter 小号泄漏。由于其小号关注的篮球队信息与主账号存在相似,注册信息中的家乡是相同的,且小号与其弟弟的账号存在少量互动,因此被网友识别为同一用户。无独有偶,王菲在 2015 年弃用了使用多年的主账号,但在 2018 年被网友通过社交关系发现了其仍在使用的小号。这一系列事件都表明,人们在使用小号进行社交活动时,应该将小号的性别、年龄、爱好等属性,以及社交关系设置成与主账号完全不同的信息,以达到隐藏身份的目的。

然而,这些具有与主账号完全不同的社交关系和属性信息的小号仍然存在用户隐私泄漏问题。本文将给出一种基于主题模型的社交网络匿名用户重识别方法。该方法仅依赖于社交网络用户发布的文本信息,通过文本信息所蕴含的用户的兴趣爱好、关注主题等构建用户画像,将匿名身份或小号与主账号关联起来,实现对用户身份的重识别攻击。

本文的主要贡献如下:

1)本文提出的用户身份的重识别方法仅依赖于用户发布的文本内容,无需其他属性或社交关系等数据,扩大了重识别攻击的适用场景。

2)该算法能够自适应地提取文本主题,使其在重识别攻击时采用更恰当的主题分类,而不必局限于预先定义的主题分类,从而提高了算法的适用范围和用户画像之间的可区分性。

3)将文本长度和发布时间作为用户画像的重要影响因素,显著提高了重识别攻击的准确率。

4)在真实的社交网络数据集上进行了有效性验证,并进一步分析了算法准确性与参数选择之间的关系。

2 相关工作

目前,用户重识别的研究方法主要分为 4 类:基于用户属性的重识别[2-5]、基于用户位置和轨迹的重识别[6-9]、基于社交关系的重识别[10-16]和基于文本内容的重识别[17-20]。

基于用户属性的重识别方法是最为直观的一种方法。文献[2-3]提出了基于多个账号的昵称之间的相似性进行用户重识别的方法;除了用户昵称外,文献[4-5]更进一步对账号相关的生日、性别、学校、地址等属性信息进行相似性分析来提高重识别的准确性。

社交网络中用户所签到的地理位置也能够用来重识别用户的身份。文献[6]提出了基于用户地理位置来识别用户的方法。随后,文献[7]采用高斯分布改进了直接使用地理位置点进行匹配的方法。更进一步,文献[8-9]提出了基于隐马尔科夫模型的用户重识别方法,该方法能够对用户的地理空间活动轨迹建模,并识别用户身份。

在基于社交关系的用户身份重识别方面,文献[10-11]提出了基于社交关系的网络拓扑结构进行重识别的方法。在重识别时,较为常见的是利用社交用户节点的度数[12]以及子图拓扑结构[13]进行相似性比较。更进一步,文献[14]提出了基于条件随机场的 JLA(Joint Link-Attribute)算法,同时对社交关系和用户属性进行了分析。与此同时,也存在针对此类重识别攻击的匿名研究工作,如采用边扰动[15]、构建超级节点[16]等,通过这些方式来改变社交结构以实现对用户身份的隐藏。

然而,上述几类研究工作所基于的数据集是能够通过注册小号或匿名机制进行有效隐藏的。针对该问题,基于文本内容的重识别方法被提出。由于用户的文本信息是用户兴趣爱好、写作风格、用词习惯的真实体现,因此小号与主账号发布的文本内容是存在一定相似性的。文献[17]提出了基于 FOAF(Friend-Of-A-Friend)词汇表或权重兴趣(Weighted Interest)词汇表的用户画像构建算法,用于分析用户之间的相似性。文献[18]在构建用户画像时采用了另一种从维基百科分类获取主题,并与文本内容进行关联的方式来构建画像。然而上述采用预先定义主题分类的策略获得主题的方法将限制所分析的文本内容范围。鉴于此,文献[19-20]利用 LDA(Latent Dirichlet Allocation)获得主题分类,提高了方案的适用范围,但是在仅采用用户发布的文本数据进行重识别时其准确性仍然有待提高。

3 基于主题模型的匿名用户重识别算法

基于主题模型的匿名用户重识别算法包括 3 个主要步骤:主题提取、画像构建和相似度计算。

3.1 主题提取

区别于已有的预先指定主题分类的重识别方法,本文采用 LDA 算法在原始的文本集中训练获得的主题。

LDA 是一种无监督机器学习算法,它能够将文档 D_i 转化为文档-主题向量,即概率分布 $P(\boldsymbol{T}|D_i)$,其中 $\boldsymbol{T}=(T_1,T_2,\cdots,T_k)$是主题向量。在 LDA 模型中,一篇文档包含多个主题,文档中的每个词都是由其中一个主题产生的。即存在两个多项式概率分布 $P(\boldsymbol{T}|\boldsymbol{D})$和 $P(\boldsymbol{W}|\boldsymbol{T})$,$P(\boldsymbol{T}|\boldsymbol{D})$是文档上的主题分布,$P(\boldsymbol{W}|\boldsymbol{T})$是主题上的单词出现的概率分布。因此,一个文档可以按照如下步骤产生:

1) http://www.sohu.com/a/227716187_485557

1）从文档 i 的主题分布 $P(\boldsymbol{T}|D_i)$ 中抽样生成文档 i 的第 j 个词的主题 $T_{i,j}$；

2）从主题 $T_{i,j}$ 的单词分布 $P(\boldsymbol{W}|T_{i,j})$ 中抽样产生单词 $W_{i,j}$。

按照上述步骤能够逐个产生单词，从而形成一篇文档。更进一步，LDA 模型认为 $P(\boldsymbol{T}|\boldsymbol{D})$ 和 $P(\boldsymbol{W}|\boldsymbol{T})$ 也应该满足一定的概率分布，而不是固定值，因此引入了 α 和 β 两个狄利克雷分布参数来完善文档的生成过程。经过 LDA 算法处理后，文档集中的每个文档都被转化为概率分布 $P(\boldsymbol{T}|D_i)$。

3.2　画像构建

本文画像构建方法的基础假设为：不同用户在发布文本内容时，会由于兴趣爱好的不同选择不同的主题；同一个用户在主账号上发布的文本内容与以匿名方式（小号）发布的文本内容仍然存在主题选择上的相似性。因此，在 LDA 模型的基础上，用户画像被表示为关于主题的概率分布 $P(\boldsymbol{T}|u)$，称作用户画像向量。$P(T_i|u)$ 的值越大，表示用户 u 对主题 T_i 的兴趣越大。具体地，用户画像向量的计算方法如下：

$$P(\boldsymbol{T}|u)=\sum_{D_j\in u}P(\boldsymbol{T}|D_j,u)*P(D_j|u)$$

其中，$P(\boldsymbol{T}|D_j,u)$ 表示用户 u 发表文档 D_j 时对主题选择的概率分布。由于社交网络中每个文档通常仅对应一个用户，因此 $P(\boldsymbol{T}|D_j,u)$ 就是 $P(\boldsymbol{T}|D_j)$，它能够在主题提取过程中得到。而 $P(D_j|u)$ 表示用户 u 发表文档 D_j 的概率，可以简单假定：

$$P(D_j|u)=\frac{1}{|D_j\in u|}$$

这是等权重的情况，即每个文档对于用户画像向量的贡献是相同的。然而实际上，用户在发布每篇文档时对于兴趣爱好的重视程度是存在差异的。本文将引入文档长度因子和发布时间因子来对每个文档在构建用户画像向量时的权重进行调整。

文档长度因子的计算方法如下：

$$f(l_{D_j})=\ln(1+l_{D_j})$$

其中，l_{D_j} 为文档 D_j 的长度。文档长度因子为函数 $f(l_{D_j})$，它反映了作者在该文档上所耗费的时间和精力。通常情况下，文档越长说明用户在写作中选择的主题越符合其兴趣爱好，因此应对用户画像向量具有更大贡献。

发布时间因子的计算方法如下：

$$\varepsilon(\Delta t)=(1-\gamma)*\exp(\lambda*\Delta t)+\gamma$$

其中，$\Delta t=t_{D_j}-t_{\text{mid}}$ 是文档 D_j 发布时间 t_{D_j} 与中间时间 t_{mid} 的时间差。t_{mid} 是构建用户画像向量时所采用的文档集合之间的中间时间点。由于用户的兴趣爱好会随时间变化而改变，因此在对比两个用户画像向量时，时间上越接近的文档对用户画像向量的贡献应该越大，以此来提高重识别的准确率。此外，γ 和 λ 是通过实验进行估计的，在不同数据集上可以根据识别效果进行调整。

最终，加入文档长度因子和发布时间因子后的用户 u 发布文档 D_j 的概率为：

$$P(D_j|u)=\frac{f(l_{D_j})*\varepsilon(\Delta t)}{|D_j\in u|}$$

3.3　相似度计算

用户重识别是通过计算两个用户画像向量的相似度实现的。目前已有很多计算两个向量之间相似度的方法，例如余弦相似度、Pearson 相似度、Jensen-Shannon 相似度等。余弦相似度是对两个归一化向量做点积，更侧重于衡量两个向量在矢量空间内的夹角；Pearson 相似度为去中心化后的余弦相似度；Jensen-Shannon 相似度是基于 K-L 距离构建的相似度，K-L 距离是用来衡量两个概率分布之间差异的量，Jensen-Shannon 相似度基于两个概率分布的差异，以对称的方式构建用户相似度。由于本文构建的用户画像是一种概率分布向量的形式，因此本文采用了 Jensen-Shannon 相似度来计算用户画像向量之间的距离。用户 p 与用户 q 的相似度计算如下：

$$Sim(p,q)=\frac{1}{\sqrt{\frac{D_{kl}(P(\boldsymbol{T}|p)\parallel M)+D_{kl}(P(\boldsymbol{T}|q)\parallel M)}{2}}}$$

其中，$M=\frac{P(\boldsymbol{T}|p)+P(\boldsymbol{T}|q)}{2}$，$D_{kl}(x\parallel y)=\sum_i\log\frac{x(i)}{y(i)}$。

4　实验结果与分析

4.1　实验概述

本文实验采用的数据集来自知乎 2039 个用户在 2015 年 11 月至 2016 年 11 月发表的文章。每个用户的文章数大于或等于 150 篇。每篇文章与该文章对应的问题描述被合并在一起作为一个文档。将每个用户最新发布的 50 个文档作为测试集，而将剩余部分作为训练集，并分别构建用户画像向量。对用户进行重识别是将 2039 个来自训练集的用户画像向量集 $\boldsymbol{T}$ 与 2039 个来自测试集的用户画像向量集 $\boldsymbol{R}$ 进行相似度计算。

本文方案的准确率的计算策略如下：为每个从 $\boldsymbol{R}$ 中选择 top-k 个相似度最高的画像向量构成集合 top-$k_i=\{r_{i,j}|r_{i,j}\in R,0<j<k\}$，然后验证 top-$k_i$ 中是否存在一个 $r_{i,j}$ 与 t_i 是同一个真实用户。准确率定义如下：

$$Accuracy=\frac{\sum_{i=1}^{n}\sum_{j=1}^{k}Matched(t_i,r_{i,j})}{n}$$

其中，n 为用户数。函数 $Matched(x,y)$ 的定义为：

$$Matched(x,y)=\begin{cases}1, & x\text{ 与 }y\text{ 为同一个真实用户}\\ 0, & x\text{ 与 }y\text{ 为不同的真实用户}\end{cases}$$

4.2　主题提取参数分析

在使用 LDA 进行主题提取时，数据集的大小和预先定义的主题个数共同决定了主题的范围。通常情况下，预先定义的主题个数较少时，产生的主题往往具有更加宽泛的含义，例如田径运动、球类运动。而主题个数较多时，产生的主题的含义往往更加具体，例如赛跑、标枪、足球、排球等。因此，主题的个数会影响重识别的准确率。然而，目前并没有一种通用的快速预测文档集主题个数的方法。幸运的是，随着主题个数的增加，准确率将很快趋于稳定。因此，本文针对当前数据集对主题个数、数据集大小与重识别准确率的关系进行实验分析，结果如表 1 所列。

表 1 用户数/文档数与最优主题数的关系

Table 1 Relationship between scales of user/document and number of topic

用户个数	文档数	最优主题数
26	4177	10
101	16236	20
245	39407	35
1218	195608	50
2039	327453	60

最优主题数是指固定用户个数和文档数后，使得重识别准确率最高的主题个数。由于数据集中每个用户的文档数量是接近的，因此主题个数 k 的影响因素可以直接采用用户数 n 计算。经过实验数据的拟合，最优主题数 k 的计算方法如下：

$$k \approx 11.2136 * \ln(n) - 28.7029$$

4.3 画像构建影响因子分析

本文在构建用户画像向量时，固定选择主题数为 60，然后进行了两组测试：等权重，以及增加文档长度因子和发布时间因子，实验结果如图 1 所示。

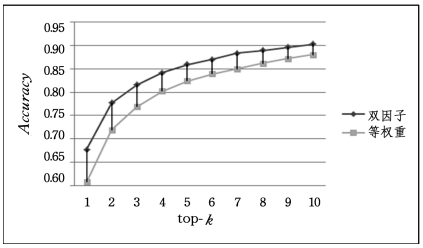

图 1 文档长度与发布时间对准确率的影响

Fig. 1 Effect of document lenght and time for accuracy

从图 1 可以看出，文档长度因子与发布时间因子的加入能够有效提高重识别的准确率。其原因在于文章越长，该用户在该文章上花费的时间与精力越多，该文章就越能真实地反映用户的兴趣偏好。同时，文档长度因子的构建形式应该是文章长度的一类凸函数。也就是说，虽然文章长度从十几个字增加到 100 个字，与从 200 个字增加到 300 个字时增加的字数相同，但是前者在表达用户兴趣爱好时所带来的增益应该是远大于后者的。因此，本方案所采用的文档长度因子 $f(l_{D_j}) = \ln(1 + l_{D_j})$ 能够更准确地为参与用户画像构建的每个文档赋予权重。另外，用户的兴趣偏好也会随着时间发生变化，时间间隔较大的文档集合在兴趣偏好上会有一定差异。本方案的发布时间因子增大了时间间隔较近的文章的权重，减小了时间间隔较远的文章的权重，因此能够提高重识别的准确性。

4.4 方案对比

为进一步表明所提方案的有效性，本文将其与文献[19-20]的方法进行了对比。文献[19-20]在计算相似度时不仅采用了用户发布的文档数据，还采用了用户之间的社交关系结构来提高相似度分析的准确性。而本文方案的重要贡献之一是仅依赖于用户发布的文本内容来实现高准确率的身份重识别攻击。因此，我们在仅有用户发布的文本数据集的情况下，将本文方法与文献[19-20]在处理文本内容时所采用的方法(记为 Simple-LDA)进行了对比。Simple-LDA 首先将同一用户的所有文档合并为一个文档，然后将该文档输入到 LDA 算法进行训练，从而获得该用户在主题向量上的概率分布，并以此为基础计算不同用户之间的相似度。

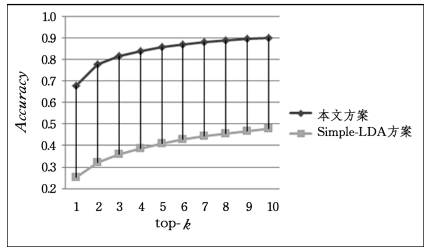

图 2 本文方案与 Simple-LDA 方案的对比结果

Fig. 2 Comparison of proposed method and Simple-LDA method

从实验结果可以看出，本文算法在仅采用文本数据集进行分析的情况下，具有更高的用户重识别准确率。主要原因包括两方面：1)本文在使用原始数据集时，没有将单个用户的所有文档进行合并，因此采用 LDA 算法提取的主题更加准确；2)增加文档长度因子和发布时间因子进一步提高了重识别的准确率。

结束语 本文针对社交网络中用户以匿名方式保护个人隐私所存在的问题展开研究，提出了一种基于主题模型的用户重识别方法。该方法通过主题向量的方式构建用户画像，并以度量用户画像之间的相似性实现用户身份的重识别。同时，在构建用户画像时，主题的选择范围是自适应的，且考虑了文档长度和发布时间对用户画像的影响，提高了重识别的准确率。最后，本文通过在真实社交网络数据集上进行重识别实验，进一步探讨了各种参数的选择问题，并将所提方法与其他方案进行了对比。实验表明，本文算法在社交网络用户重识别的准确率方面具有一定优势。下一步的研究工作将从两方面展开：1)本文方案在处理大规模数据集时，运行效率较低，难以应对较大规模的分析任务。因此，后续工作中将研究该方案的分布式实现来提高其效率。2)针对单数据源的用户重识别研究在实际应用中具有一定局限性。后续我们将针对同一用户在多个社交网络中采用不同用户名进行社交活动的行为展开用户重识别研究，以期深入探讨社交网络中的用户隐私问题。

参 考 文 献

[1] FENG D G, ZHANG M, LI H. Big Data Security and Privacy Protection[J]. Chinese Journal of Computers, 2014, 37(1): 246-258. (in Chinese)
冯登国，张敏，李昊. 大数据安全与隐私保护[J]. 计算机学报，2014，37(1)：246-258.

[2] 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 techno-

logies. 2011：1-17.

[3] LIU J，ZHANG F，SONG X，et al. What's in a name?：an unsupervised approach to link users across communities[C]// ACM International Conference on Web Search and Data Mining. ACM，2013：495-504.

[4] MALHOTRA A，TOTTI L，MEIRA W，et al. Studying User Footprints in Different Online Social Networks[C]// IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE，2012：1065-1070.

[5] VOSECKY J，HONG D，SHEN V Y. User identification across multiple social networks[C]// 2009 First International Conference on Networked Digital Technologies. IEEE，2009：360-365.

[6] ZANG H，BOLOT J. Anonymization of location data does not work：A large-scale measurement study[C]// Proceedings of the 17th Annual International Conference on Mobile Computing and Networking. New York：ACM，2011：145-156.

[7] WANG H，GAO C，LI Y，et al. De-anonymization of mobility trajectories：Dissecting the gaps between theory and practice[C]// Proceedings of The 25th Annual Network & Distributed System Security Symposium (NDSS'18). 2018.

[8] WANG R，ZHANG M，FENG D，et al. A de-anonymization attack on geo-located data considering spatio-temporal influences[C]// Proceedings of the 2015 International Conference on Information and Communications Security. Springer，Cham，2015：478-484.

[9] CHEN Z，FU Y，ZHANG M，et al. The De-anonymization Method Based on User Spatio-Temporal Mobility Trace[C]// Proceedings of the 2017 International Conference on Information and Communications Security. Cham：Springer，2017：459-471.

[10] NARAYANAN A，SHMATIKOV V. De-anonymizing social networks[C]// 30th IEEE Symposium on Security and Privacy. IEEE，2009：173-187.

[11] FU H，ZHANG A，XIE X. De-anonymizing social graphs via node similarity[C]// International Conference on World Wide Web. 2014：263-264.

[12] LIN S H，LIAO M H. Towards publishing social network data with graph anonymization[J]. Journal of Intelligent & Fuzzy Systems，2016，30(1)：333-345.

[13] YUAN Y，WANG G，XU J Y，et al. Efficient distributed subgraph similarity matching[J]. The VLDB Journal，2015，24(3)：369-394.

[14] SERGEY B，ANTON K，SEUNGTAEK P，et al. Joint link-attribute user identity resolution in online social networks[C]// The 6th SNA-KDD Workshop. 2012：1-9.

[15] ZHANG L，ZHANG W. Edge anonymity in social network graphs[C]// Proceedings of the 2009 International Conference on Computational Science and Engineering，Piscataway，NJ：IEEE. 2009(4)：1-8.

[16] TASSA T，COHEN D J. Anonymization of Centralized and Distributed Social Networks by Sequential Clustering[J]. IEEE Transactions on Knowledge and Data Engineering，2013，25(2)：311-324.

[17] ZHENG R，LI J，CHEN H，et al. A framework for authorship identification of online messages：Writing-style features and classification techniques[J]. Journal of the Association for Information Science and Technology，2006，57(3)：378-393.

[18] KONG X，ZHANG J，YU P S. Inferring anchor links across multiple heterogeneous social networks[C]// Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. ACM，2013：179-188.

[19] ZHANG Y，WU Y，YANG Q. Community Discovery in Twitter Based on User Interests[J]. Journal of Computational Information Systems，2012，8(3)：991-1000.

[20] YAN G H，SHU X，MA Z C，et al. Community discovery for microblog based on topic and link analysis[J]. Application Research of Computers，2013，30(7)：1953-1957.(in Chinese)
闫光辉，舒昕，马志程，等. 基于主题和链接分析的微博社区发现算法[J]. 计算机应用研究，2013，30(7)：1953-1957.