

基于单点地名匹配和局部地名筛选的推特用户定位方法

薛锦^{1,2}, 袁福祥², 刘毅敏², 张萌², 乔亚琼^{2,3}, 罗向阳²

(1. 郑州大学网络空间安全学院, 河南 郑州 450003; 2. 河南省网络空间态势感知重点实验室, 河南 郑州 450001;
3. 华北水利水电大学信息工程学院, 河南 郑州 450045)

摘要: 用户推文中的地名信息是定位推特用户的重要基础数据之一, 但现有推特用户定位方法提取的地名在数量和可靠性方面均存在欠缺, 影响了用户定位准确性。提出基于单点地名匹配和局部地名筛选的推特用户定位方法。设计一种基于地名位置聚集度的地名类型判别算法, 根据位置分布的聚集程度构建单点地名库, 获取推文中更多可靠地名; 提出一种基于用户位置聚集度的局部地名筛选算法, 分别以地名经纬度和用户平均经纬度为中心, 计算用户位置聚集度, 筛选更高聚集度、更可靠的局部地名; 基于用户社交关系、用户对地名的提及关系构建用户-地名异构图, 并利用图表示学习和神经网络定位用户。基于常用公开数据集 GEOTEXT 和 TW-US 进行大量用户定位实验, 并与 HGNN、ReLP、GCN 等 9 种现有推特用户位置推断典型方法进行了对比, 结果表明, 所提方法对推特用户的位置推断准确率具有明显优势, 相比 9 种现有典型方法, 在 GEOTEXT 数据集上, 平均误差降低了 7.3~342.8 km, 中位数误差降低了 2.4~354.4 km, 大地区级定位准确率提高了 1.3%~26.3%; 在 TW-US 数据集上, 平均误差降低了 8.6~246.6 km, 中位数误差降低了 5.7~149.7 km, 大地区级定位准确率提高了 1.5%~20.5%。

关键词: 用户定位; 用户生成内容; 地名; 社交媒体

中图分类号: TP391

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023053

Twitter user geolocation method based on single-point toponym matching and local toponym filtering

XUE Jin^{1,2}, YUAN Fuxiang², LIU Yimin², ZHANG Meng², QIAO Yaqiong^{2,3}, LUO Xiangyang²

1. School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450003, China

2. Henan Key Laboratory of Cyberspace Situation Awareness, Zhengzhou 450001, China

收稿日期: 2022-10-30; **修回日期:** 2023-04-18

通信作者: 袁福祥, rookiefx@163.com

基金项目: 国家自然科学基金 (U1804263, U2172435, 62272163); 国家重点研发计划 (2022YFB3102900); 中原科技创新领军人才项目 (214200510019); 河南省科技攻关项目 (222102210036); 河南省自然科学基金 (222300420230)

Foundation Items: The National Natural Science Foundation of China (U1804263, U2172435, 62272163), The National Key Research and Development Program of China (2022YFB3102900), Zhongyuan Science and Technology Innovation Leading Talent Project of China (214200510019), The Key Science and Technology Project of Henan Province (222102210036), The Henan Province Science Foundation for Youths (222300420230)

引用格式: 薛锦, 袁福祥, 刘毅敏, 等. 基于单点地名匹配和局部地名筛选的推特用户定位方法[J]. 网络与信息安全学报, 2023, 9(4): 53-63.

Citation Format: XUE J, YUAN F X, LIU Y M, et al. Twitter user geolocation method based on single-point toponym matching and local toponym filtering[J]. Chinese Journal of Network and Information Security, 2023, 9(4): 53-63.

3. School of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450045, China

Abstract: The availability of accurate toponyms in user tweets is crucial for geolocating Twitter users. However, existing methods for locating Twitter users often suffer from limited quantity and reliability of acquired toponyms, thus impacting the accuracy of user geolocation. To address this issue, a twitter user geolocation method based on single-point toponym matching and local toponym filtering was proposed. A toponym type discriminating algorithm based on the aggregation degree of locations of the toponym was designed. In the proposed algorithm, a single-point toponym database was generated to provide more reliable toponyms extracted from tweets. Then, according to a proposed local place name filtering algorithm based on the aggregation degree of user location, the aggregation degree of user location centered on the longitude and latitude of toponyms and the average longitude and latitude of users were calculated. This process helped in extracting local toponyms with a high aggregation degree, which enhances the reliability of toponyms used in geolocation. Finally, a user-toponym heterogeneous graph was constructed based on user social relationships and user mentions of toponyms, and users were located by graph representation learning and neural networks. A large number of user geolocation experiments were conducted based on two commonly used public datasets in this field, namely GEOTEXT and TW-US. Comparisons with nine existing typical methods for Twitter user geolocation, including HGNN, ReLP, and GCN, demonstrate that our proposed method achieves significantly higher geolocation accuracy. On the GEOTEXT dataset, the average error is reduced by 7.3~342.8 km, the median error is reduced by 2.4~354.4 km, and the accuracy of large area-level geolocation is improved by 1.3%~26.3%. On the TW-US dataset, the average error is reduced by 8.6~246.6 km, the median error is reduced by 5.7~149.7 km, and the accuracy of large area-level geolocation is improved by 1.5%~20.5%.

Keywords: user geolocation, user-generated text, toponym, social media

0 引言

随着智能移动终端设备的普及,以推特为代表的社交网络应用已成为人们分享身边趣闻、位置动态的主要平台。在巨量推特数据中,用户位置信息在许多领域具有非常重要的应用价值,如疾病传播监控^[1]、选举预测^[2]、基于位置的网络攻击技术分析^[3]、社交网络链路预测^[4-5]等。然而,公开可获取的位置信息(如推文中的位置标签、用户档案中的位置字段等)或极其稀疏,或可靠性难以保证^[6],这些问题使得推特用户准确定位非常困难。如何利用用户推文和社交关系等公开数据准确推断用户位置仍是一个亟待解决的问题。

近些年,学术界在推特用户定位方面展开了研究,产生了一些代表性成果。这些成果可按照使用的数据类型分为三类:一是基于用户生成文本的推特用户定位方法;二是基于社交关系的推特用户定位方法;三是多源数据融合的推特用户定位方法。

基于用户生成文本的推特用户定位方法利用词或短语和位置间对应关系定位推特用户。

Eisenstein等^[7]建立主题词和地区间的层次概率模型来定位推特用户。Han等^[8]首次定义位置指示词为与位置相关联的词语,并分别基于TF-IDF^[9]、信息增益率、最大熵提取位置指示词,定位推特用户。Cheng等^[6]基于大量已知城市级位置的用户文本,估计词语出现位置的焦点和离散度来识别位置指示词,再将位置指示词对应位置预测为推特用户位置。Wing等^[10]提出了HierLR模型,在利用 k 维树(一种用 k 维超平面分割 k 维空间中点集的二叉树数据结构, k -dimensional tree)将地球表面划分为层次地理网格后,基于信息增益率提取格子对应位置指示词,再利用逻辑回归模型预测每个节点的子节点概率,最终将路径上概率乘积最大的叶节点作为推特用户位置。Rahimi等^[11]提出了MLP4Geo模型,提取用户的词袋模型,再训练多层感知机定位推特用户。

基于社交关系的推特用户定位方法的主要思想是,推特用户及其朋友的位置具有相关性。Davis等^[12]认为只有相互关注才是朋友关系,将朋友最集中的位置作为用户的位置。Kong等^[13]利用两个用户朋友集合的余弦相似度推断用户位置。

Jurgens等^[14]在用户关系图上,利用标签传播方法(一种基于图的半监督学习方法,其基本思路是用已标记节点的标签信息去预测未标记节点的标签信息)定位目标用户。McGee等^[15]提取了多种与用户亲密度相关的特征,训练分类树来预测用户亲密度,再根据亲密度,利用最大似然估计预测用户间距离。

多源数据融合的方法利用用户生成文本、社交关系等多种数据联合推断用户位置。Rahimi等^[16]提出了GCN模型,将用户文本的词袋模型作为用户节点特征,在坍塌的用户图上,训练图卷积神经网络模型来推断用户位置,此外,通过将文本词袋模型替换为基于用户邻居的独热变量,提出了基准模型GCN-LP。Do等^[17]提出了MENET模型,利用doc2vec^[18]、Node2Vec^[19]、TF-IDF^[9]等技术获取多种用户特征,再结合神经网络定位用户。Tian等^[20]提出了ReLP模型,利用信息增益率提取位置指示词加入社交关系图中,再使用图嵌入模型获取用户节点特征来计算节点间转移概率,并利用标签传播算法推断用户位置。Rahimi等^[21]提出了MADCEL-W模型,先基于用户词袋模型训练逻辑回归模型来定位用户,再将定位结果连到社交关系图中,利用标签传播方法定位用户。王凯等^[22]参考Cheng等^[6]方法,估计词语出现位置的焦点和离散度,再结合社交关系构建“朋友”、位置指示词和位置的概率模型。此类方法融合了多种数据,用户定位准确性普遍高于其他两类方法,并解决了基于社交关系网络的方法无法定位孤立用户的问题。Zhou等^[23]提出了MetaGeo模型,利用元学习范式,从大量小任务中提取定位任务的先验分布,同时将概率图推断引入参数更新,以应对新位置用户的定位任务。Zhou等^[24]提出了HGNN模型,基于用户簇间位置关系衡量用户与用户簇间距离,聚合用户和其邻居、相关用户簇特征,进而推断用户位置。

基于用户生成文本和基于多源数据融合两类方法的定位效果均与位置指示词的数量和准确性有关,其中地名与位置具有紧密的联系,是最重要的位置指示词。然而,由于有标签用户地理分布不均匀和用户文本中每个地名的出现频数低,现有位置指示词提取方法得到的地名在数量和可

靠性上存在欠缺,影响了推特用户定位准确性。为此,本文基于地名库匹配推文中的地名,以解决地名提取数量较少的问题,但匹配所得地名中,无法利用对应多个分散经纬度点的多点地名指示用户位置,而被多个不同位置推特用户提及的全局地名则会影响用户定位准确率。

针对上述问题,本文提出基于单点地名匹配和局部地名筛选的推特用户定位方法(单点地名指对应一个经纬度点的地名,局部地名指被局部区域推特用户提及的地名)。该方法首先利用基于所提地名类型判别算法构建的单点地名库匹配用户推文中的地名;然后在匹配地名中,基于局部地名筛选算法,提取被局部区域用户提及的直接地名和间接地名;最后利用直接地名对应位置和提及间接地名的有标签用户主要位置,共同定位目标用户。

本文所提方法的主要贡献如下。

1) 提出一种地名类型判别算法。设计经纬度聚集度指标,衡量地名相关的多个经纬度坐标的聚集程度,以过滤多点地名,保留单点地名,从而在用户推文中识别出位置指示性更强的地名,提升定位准确率。

2) 提出一种局部地名筛选算法。设计地名中心度和用户中心度两个度量指标,提取被局部区域用户提及的直接地名和间接地名作为局部地名,与现有方法提取的位置指示词相比,所提局部地名的位置指示效果更优。

1 符号定义和问题描述

本节推特用户定位相关的概念进行介绍,列出文中使用的符号,以方便查找,并将本文要解决的问题进行形式化描述。

1.1 符号定义

定义1 地名词典。地名词典是汇总多个地理数据源(如官方地名数据、维基百科等)的全球地名库,其中包含地名、地名别名、所属国家、经纬度等信息。

定义2 单点地名和多点地名。单点地名是对应一个地点的地名,记作 TP^S 。多点地名是对应多个地点的地名,记作 TP^M 。

定义3 直接地名和间接地名。直接地名是

利用地名对应位置指示用户位置的地名，记作 TP^D 。间接地名是利用提及地名的有标签用户主要位置指示用户位置的地名，记作 TP^I 。

定义4 全局地名和局部地名。全局地名是被多个不同地区用户提及的地名。局部地名是主要被一个地区用户提及的地名，由直接地名和间接地名组成。

定义5 目标用户和有标签用户。目标用户是数据集中具有用户生成文本等信息的待定位用户，记作 V^U 。有标签用户是有用户生成文本等信息且位置已知的用户，记作 V^L 。

定义6 位置指示词。位置指示词是用户推文中能指示用户位置的词或词组，如Florida Agricultural and Mechanical University，位置指示词集合记作 LIW 。

定义7 社交关系图。社交关系图是基于用户推文中@提及关系构建的用户图，记作 $G=(V,E,W)$ ，其中 V 是用户集合， E 是边集合， W 是边权重集合。

定义8 用户-地名图。在社交关系图 G 基础上，加入目标用户-直接地名、目标用户-间接地名和有标签用户-间接地名关系，构成用户-地名图，记作 $G'=(V,E,W,LIW)$ 。

定义9 位置集合。本文参考Rahimi等^[16]的工作，将推特用户定位问题视为分类问题，利用 k 维树将有标签用户的经纬度聚集成多个地区，并使用户归属于最近的地区，从而将用户经纬度位置转换为位置标签，形成位置集合 P ，其中，用户的位置集合记为 P^V ，包括有标签用户位置集合 P^L 和目标用户位置集合 P^U ，地名的位置集合记作 P^{TP} 。

本文中使用的符号及含义如表1所示。

1.2 问题定义

定义10 推特用户定位问题。对于目标用户 V^U ，给定社交关系图 G 、地名的位置集合 P^{TP} 、有标签用户位置 P^L 和其用户文本，计算得到目标用户位置 P^U 。

2 本文方法

本节将详细介绍所提方法——基于单点地名匹配和局部地名筛选的推特用户定位方法

(GEOTOP, Twitter user geolocation method based on single-point toponym matching and local toponym filtering): 首先，利用地图服务爬取目标城市的细粒度地名信息，并基于地名类型判别算法，过滤细粒度地名（对应小地点的地名，如游乐场、餐馆）和地名词典中的多点地名，构成单点地名库；然后，在目标用户推文中，基于该库匹配地名，并利用局部地名筛选算法，来提取匹配所得地名中的直接地名和间接地名；最后，根据多种提及关系构建用户-地名图，并利用图表示学习方法和神经网络定位目标用户。基于单点地名匹配和局部地名筛选的推特用户定位方法如图1所示。

表1 符号及含义
Table 1 Notations and descriptions

符号	含义
TP	地名集合
TP^I	间接地名集合
TP^D	直接地名集合
LIW	位置指示词集合
k	基于 k 维树划分后每个位置标签包含的用户数
V^L	有标签用户集合
V^U	目标用户集合
G'	用户-地名图
X	用户推文集合
P^V	用户位置的集合
P^U	目标用户位置的集合
P^L	有标签用户位置的集合
P^{TP}	地名位置的集合
P	用户和地名的位置集合
f_1	提取直接地名的阈值
f_2	提取间接地名的阈值
s	全局名人过滤阈值

2.1 构建单点地名库

本节介绍单点地名库的构建，主要分为细粒度地名数据获取和多点地名过滤两部分。

2.1.1 细粒度地名数据获取

地名词典是一种公开地名数据库，包含大量地名信息。但是，地名词典缺少细粒度地名数据，在提取用户文本中地名时会存在提取不充分的问题。为此，本文利用地图服务爬取目标城市的细

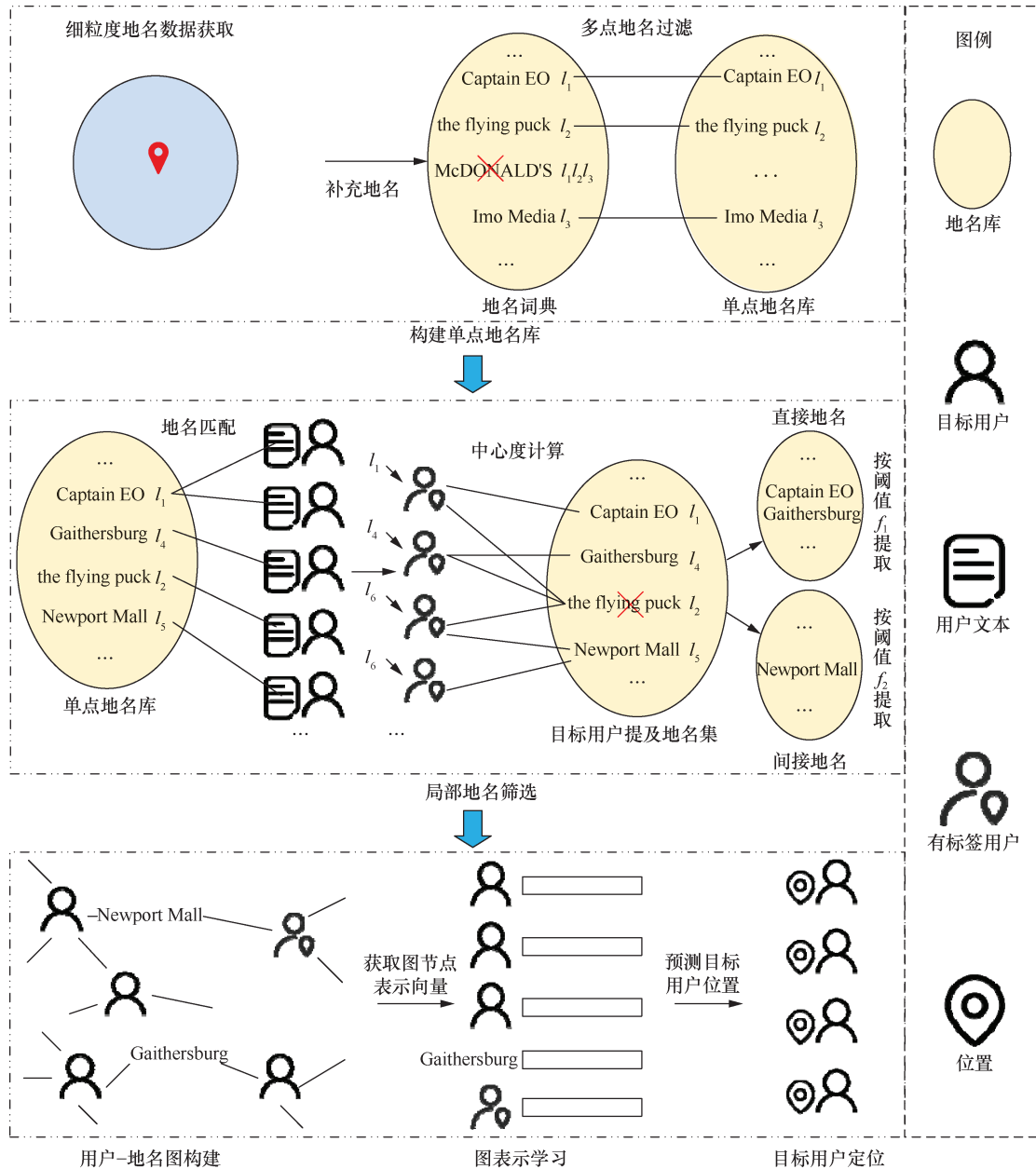


图1 基于单点地名匹配和局部地名筛选的推特用户定位方法

Figure 1 Twitter user geolocation based on single-point toponym matching and local toponym filtering

粒度地名信息，爬取流程如下。

1) 对每个目标城市，利用地理编码服务获取城市名对应的由4个边界经纬度构成的城市范围(NLat, ELon, SLat, WLon)。

2) 将(NLat, WLon)作为起始中心点，利用逆地理解析服务获取固定大小圆内的地名信息。

3) 按从左到右、从上到下的顺序，移动中心点，重复步骤2)，直到(SLat, ELon)为止。

通过解析获取的数据，可以获取细粒度地名

的所属国家、所属城市和经纬度等信息。

2.1.2 多点地名过滤

多点地名指向若干个不同位置，故难以推断提及该多点地名的用户所在位置，因此需要过滤多点地名。

多点地名在地名库中以多条记录的形式存在，但并非所有存在多条记录的地名都是多点地名。这是因为地名词典的数据存在冗余，且爬取细粒度地名时，为保证不遗漏，令多次爬取的范

围存在重叠,所以地名库中单点地名的记录可能存在多条。为了解决该问题,提出地名类型判别算法,如算法1所示。

算法1 地名类型判别算法

输入 地名集合 TP、地名位置集合 P^{TP}

输出 单点地名库 TP^S

1) 初始化直接地名集合 $TP^S(\emptyset)$;

2) for $tp_i \in TP$ do

3) 获取地名 tp_i 对应的位置集合 $P_{tp_i}^{TP}$;

4) 计算 $P_{tp_i}^{TP}$ 中经纬度位置的均值作为中心;

5) 计算 $P_{tp_i}^{TP}$ 中所有位置距中心的距离,提取最大距离 $dist$;

6) if $dist < d$

7) 将地名 tp_i 识别为单点地名,加入单点地名库 TP^S ;

8) end for

通过融合地名词典和细粒度地名信息,并用地名类型判别算法提取单点地名,一个单点地名库构建完成。

2.2 局部地名筛选

基于构建的单点地名库,本文利用正则表达式(1)来匹配目标用户生成文本中的地名,其中 wd_i 是组成地名的单词,以保证所提地名不会为用户文本中多个连续词的一部分。

$$(\backslash bwd_1 \backslash s + wd_2 \backslash s + \cdots wd_i \backslash s + \cdots wd_n \backslash s \backslash b) \quad (1)$$

在用户文本中提取的地名,全局地名被大地理范围用户所知,因此会被多个不同区域用户提及,不仅无法提供位置指示,甚至会降低用户定位精度。为此,本文提出局部地名筛选算法,来过滤全局地名,提取局部地名。

根据地名对目标用户位置的指示方式的不同,将局部地名分为两类:直接地名和间接地名。直接地名指向提及用户周边位置,故可以利用地名位置指示用户位置,而间接地名虽离用户较远,但更常被局部区域的用户提及,如Newport Mall,因此可以利用共同提及该地名的有标签用户位置指示目标用户位置。基于两类局部地名的特性,本文提出地名中心度(TC, toponym centrality)和用户中心度(UC, user centrality),分别用于提取直接地名和间接地名。地名中心度和用户中

心度的计算如式(2)和式(3)所示。

$$TL = \begin{cases} \frac{\left| \left\{ p_i \mid d_{p_i, P_{tp_i}} < d, p_i \in P_{tp_i}^L \right\} \right|}{\left| P_{tp_i}^L \right|}, & \left| P_{tp_i}^L \right| \neq 0 \\ 1, & \left| P_{tp_i}^L \right| = 0 \end{cases} \quad (2)$$

其中, $P_{tp_i}^L$ 是提及地名 tp_i 的有标签用户位置集合, $d_{p_i, P_{tp_i}}$ 是提及用户位置 p_i 和地名位置 p_{tp_i} 的半正弦距离, d 是设置的局部区域半径。

$$UL = \text{Max} \left(\frac{\left| \left\{ p_i^L \mid \text{dis}(p_i^L, c) < d, p_i^L \in P_{tp_i}^L \right\} \right|}{\left| P_{tp_i}^L \right|} \right), \quad (3)$$

$$c = \{Q_1, Q_2, Q_3\}$$

其中, Q_1 、 Q_2 和 Q_3 是提及用户位置集合 $P_{tp_i}^L$ 先后按纬度、经度排序后取得的第一分位数、中位数和第三分位数,其他符号同式(2)。

基于提出的地名中心度和用户中心度,设计直接地名和间接地名的提取算法,如算法2和算法3所示。这两个算法共同构成局部地名筛选算法。

算法2 直接地名的提取算法

输入 地名集合 TP、地名位置集合 P^{TP} 和有标签用户位置集合 P^L

输出 直接地名集合 TP^D

1) 初始化直接地名集合 $TP^D(\emptyset)$;

2) for $tp_i \in TP$ do

3) 获取提及地名 tp_i 的有标签用户位置集合 $P_{tp_i}^L$;

4) 根据式(2)计算 tp_i 的地名中心度 TL

5) if $TL > f_1$

6) 将地名 tp_i 加入直接地名集合 TP^D ;

7) end for

算法3 间接地名的提取算法

输入 $TP' = TP - TP^D$ 、地名位置集合 P^{TP} 和有标签用户位置集合 P^L

输出 间接地名集合 TP^I

1) 初始化间接地名集合 $TP^I(\emptyset)$;

2) for $tp_i \in TP'$ do

3) 获取提及地名 tp_i 的有标签用户位置集合 $P_{tp_i}^L$;

- 4) 先按纬度，再按经度对有标签用户位置集合 P_{tp}^L 排序；
- 5) 根据式(3)计算 tp_i 的用户中心度 UL ；
- 6) if $UL > f_2$
- 7) 将地名 tp_i 加入间接地名集合 TP^I ；
- 8) end for

2.3 基于图表示学习的用户定位

在提取局部地名后，本节首先构建用户-地名图，然后采用图表示学习方法获取节点特征，最后训练神经网络定位目标用户。

2.3.1 用户-地名图构建

首先，构建一个以用户为节点，用户单向提及关系为边的社交关系图 G 。现有方法通常将用户间的@提及视为朋友关系的体现，且因为推文中双向提及关系非常稀疏，所以根据单向提及关系构建社交关系图^[21]。

然后，因为被提及的用户中存在全局名人，与之相关的边不仅无法提供位置指示，反而可能干扰位置推断准确性，同时会使得社交关系图过于膨胀，所以，过滤掉无关用户中被 s 个不同用户提及的用户^[21]。

最终，将用户-直接地名、用户-间接地名关系加入社交关系图 G 中构成用户-地名图 G' 。

2.3.2 图表示学习

为了更好地捕获用户和位置间的映射关系，本文通过训练神经网络来拟合这种映射关系，从而定位用户。然而，这需要将用户-地名图中节点间的邻近性转化为用户和地名的特征向量间的邻近性，本文采用图表示学习方法Node2Vec^[19]解决这个问题。

该方法采用有偏随机游走算法提取目标用户周围的节点序列。首先，根据回归参数 p 、进出参数 q 和边权重计算在 $t \rightarrow v$ 条件下， $v \rightarrow x$ 的转移概率 $p(c_i = x | c_{i-1} = v)$ ，计算方法如式(4)~式(6)所示；然后，用于随机游走，获取多个固定长度的节点序列。最后，Word2Vec模型根据采样所得节点序列将节点映射到向量空间中。

$$p(c_i = x | c_{i-1} = v) = \begin{cases} \pi_{v,x}/Z, & (v,x) \in E \\ 0, & \text{其他} \end{cases} \quad (4)$$

$$\pi_{v,x} = \alpha_{p,q}(t,x) \cdot w_{v,x} \quad (5)$$

$$\alpha_{p,q}(t,x) = \begin{cases} 1/p, & d_{t,x} = 0 \\ 1, & d_{t,x} = 1 \\ 1/q, & d_{t,x} = 2 \end{cases} \quad (6)$$

其中， Z 是归一化系数， $d_{t,x}$ 是节点 t 和节点 x 间最短路径长度。

如图2所示，Node2Vec模型将图中节点映射到向量空间中，保证了连通节点向量间的相似性。

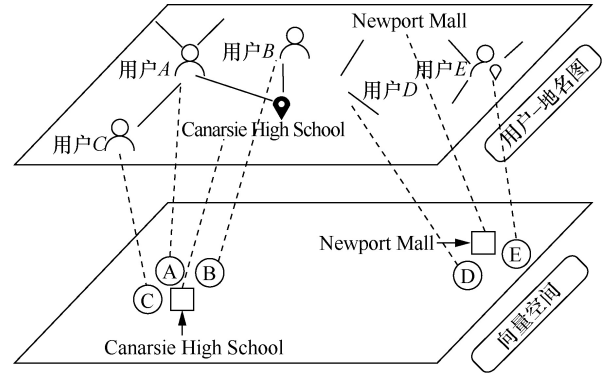


图2 Node2Vec 效果示意
Figure 2 Diagram of Node2Vec result

2.3.3 目标用户定位

在获得节点特征向量后，通过训练添加了随机失活层的两层神经网络来定位目标用户，如式(7)和式(8)所示。为了利用直接地名来指示目标用户位置，将有标签用户和直接地名的特征向量和位置标签用于训练分类模型。

$$h(\mathbf{x}) = \text{ReLU}(\mathbf{W}_1 \mathbf{x} + b_1) \quad (7)$$

$$\hat{\mathbf{y}} = \text{SoftMax}(\mathbf{W}_2 h(\mathbf{x}) + b_2) \quad (8)$$

其中， $\mathbf{x}$ 是节点特征向量， $\mathbf{W}_1$ 、 b_1 和 $\mathbf{W}_2$ 、 b_2 分别是隐藏层和输出层的参数矩阵和偏置， $\hat{\mathbf{y}}$ 是预测的标签向量。随机失活层加在输入层和隐藏层后。

训练时损失函数采用 l_1 和 l_2 正则化后的交叉熵函数，如式(9)和式(10)所示，而神经网络模型参数的优化采用Adam方法。

$$\text{Loss} = - \sum_{v \in V^L} \sum_i y_{vi} \cdot \log(\hat{y}_{vi}) + \Phi(\mathbf{W}_1, \mathbf{W}_2) \quad (9)$$

$$\Phi(\mathbf{W}_1, \mathbf{W}_2) = \lambda_1 \left(\beta_1 \sum_{j=0}^{M_1} |\mathbf{W}_{1j}| + (1 - \beta_1) \sum_{j=0}^{M_1} \mathbf{W}_{1j}^2 \right) + \lambda_2 \left(\beta_2 \sum_{j=0}^{M_2} |\mathbf{W}_{2j}| + (1 - \beta_2) \sum_{j=0}^{M_2} \mathbf{W}_{2j}^2 \right) \quad (10)$$

其中, W_{1j} 和 W_{2j} 分别是参数矩阵 W_1 和 W_2 的元素, M_1 和 M_2 分别是参数矩阵 W_1 和 W_2 的元素个数, λ_1 和 λ_2 是隐藏层和输出层正则化项的权重因子, β_1 和 β_2 则分别是隐藏层和输出层中 l_1 和 l_2 正则化项的权重因子。

3 实验结果与分析

3.1 实验设置

3.1.1 细粒度地名与单点地名库

按照 2.1 节的方法, 本文利用 HERE MAPS 地图服务爬取了 3 035 437 个细粒度地名数据, 然后将这些数据与地名词典 Geonames 合并成一个单点地名库。表 2 展示了细粒度地名和单点地名库的统计数据。

表 2 细粒度地名和单点地名统计数据
Table 2 Statistics of fine-grained toponyms and single-pointed toponyms

项目	覆盖州数/个	覆盖城市数/个	地名数/个
细粒度地名	50	307	3 035 437
单点地名	50	311	2 043 527

3.1.2 参数设置

实验的参数设置如表 3 所示, 其中, k 是基于 k 维树的区域划分中每个位置标签中包含的用户数, λ_1 、 λ_2 、 β_1 和 β_2 是损失函数中的正则项的权重因子, d 是局部区域阈值, f_1 和 f_2 分别是识别直接地名和间接地名的阈值, s 是全局名人阈值, p 和 q 是 Node2Vec 模型的回归参数和进出参数。

表 3 参数设置
Table 3 Parameter settings

参数	GEOTEXT	TW-US
k	50 人	2 400 人
输入层维度	128	128
隐藏层维度	256	256
输出层维度	129	256
λ_1	0.5	0.5
λ_2	0.5	0.5
β_1	0.5	0.5
β_2	0.5	0.5
d	150 km	150 km
f_1	0.7	0.75
f_2	0.8	0.82
s	5 人	15 人
p	4	4
q	0.25	0.25

3.1.3 推特数据集

本文采用两个常用公开推特数据集, 数据集的划分和统计情况如表 4 和表 5 所示。

表 4 推特数据集划分信息
Table 4 Statistics of twitter dataset split

数据集	训练集 用户数/个	验证集 用户数/个	测试集 用户数/个
GEOTEXT ^[7]	5 685	1 895	1 895
TW-US ^[25]	1 366 766	10 000	10 000

表 5 推特数据集统计信息
Table 5 Statistics of twitter datasets

数据集	用户数/个	推文数/条	用户提及数/个
GEOTEXT ^[7]	9 475	377 504	232 598
TW-US ^[25]	449 200	38 036 187	8 128 105

GEOTEXT^[7] 数据集和 TW-US^[25] 数据集包含用户名、用户经纬度、用户文本等数据, 其中用户经纬度分别为用户推文中第一个位置标签和多个位置标签。

3.2 性能指标

本文沿用 3 个本方向广泛采纳 (如文献[16,24]) 的性能指标: 大地区级定位准确率 (ACC@161)、平均误差 (Mean) 和中位数误差 (Median)。其中, 大地区级定位准确率是预测经纬度和真实经纬度间地理距离在 161 km 以下的用户比例, 即定位准确的用户比例, 平均误差是所有用户的预测经纬度和真实经纬度间地理距离均值, 中位数误差是所有用户预测经纬度和真实经纬度间地理距离中位数, 计算方法如式(11)~式(13)所示。

$$\text{Acc@161} = \frac{\left| \left\{ v_i \mid d_{y_{v_i}, \hat{y}_{v_i}} < 161 \text{ km}, v_i \in V^U \right\} \right|}{|V^U|} \quad (11)$$

$$\text{Mean} = \frac{\sum_{v_i \in V^U} d_{y_{v_i}, \hat{y}_{v_i}}}{|V^U|} \quad (12)$$

$$\text{Median} = \text{Sorted}(\{d_{y_{v_i}, \hat{y}_{v_i}} \mid v_i \in V^U\})_{|V^U|/2} \quad (13)$$

其中, $d_{y_{v_i}, \hat{y}_{v_i}}$ 是目标用户 v_i 的实际经纬度和预测经纬度间误差, V^U 是目标用户集合, $|V^U|$ 是目标用户总数, 经纬度间距离的计算采用半正矢公式 (haversine formula) 计算, $\text{Sorted}(\cdot)_{|V^U|/2}$ 函数取排序后数列中的中位数。

3.3 推特用户定位实验结果对比

本节在常用公开数据集GEOTEXT和TW-US上进行用户定位实验，所提方法和多个现存经典方法的性能对比结果如表6所示，表6中加粗部分是对比结果中的最高值，“—”表示该方法没有报告的数据。

如表6所示，所提方法在Acc@161、Mean和Median上均优于经典方法。与经典方法的最佳数据相比，在GEOTEXT数据集上，本文方法的Mean和Median分别降低7.3 km和2.4 km，Acc@161提高1.3%；在TW-US数据集上，本文方法的Mean和Median分别降低8.6 km和5.7 km，Acc@161提高1.5%。

用户定位准确性提升的主要原因在于本文方法提取的位置指示词更多，位置指示效果更好。现有经典方法无论是利用统计指标提交位置指示词，还是构建词和位置的概率模型或词袋模型，都会引入更多与位置指示词同地理分布的噪声词，且提取的位置指示词较少。而所提方法基于地名库匹配推文中地名，提取了更多地名，同时利用地名类型判别算法和局部地名筛选算法减少了噪声词的引入，提取了更多、更准确的位置指示词，提高了用户定位准确性。

3.4 局部地名对用户定位准确性的影响

本节分析社交关系网络和地名因素对结果的影响。在第一个实验中，本文利用图表示学习和神经网络在社交关系图上进行用户定位，验证社交关系图对用户定位结果的影响。第二个实验只基于局部地名定位，验证地名对用户定位结果的影响。该实验首先获取用户提及的直接地名位置以及与提及的间接地名相关的有标签用户位置，构成用户候选位置集合；然后在该位置集合上，计算UL，若UL大于阈值 f_2 ，则将UL对应的位置中心 c 作为目标用户预测位置，否则，定位错误。

消融实验结果对比表7所示，与GEOTOP相比，第一个实验的准确率降低了2.6%，Mean上升了17.1 km，Median上升了19.6 km，定位准确性下降；第二个实验的准确率提高了17.7%，平均误差下降了417.5 km，中位数误差下降了27.2 km，定位准确性显著提升，这表明所提局部地名对用户定位起到正面的影响。然而，由于局部地名所覆盖的用户比例低，仍需结合社交关系图来定位用户。

表8展示了UC或TC值最大的若干局部地名的位置指示效果，表中指示位置是直接地名经纬度位置或UC对应经纬度位置。结果表明，局

表6 推特用户定位性能对比
Table 6 Comparison of twitter user geolocation performance

模型	GEOTEXT 数据集			TW-US 数据集		
	大地区级定位准确率	平均误差/km	中位数误差/km	大地区级定位准确率	平均误差/km	中位数误差/km
基于用户文本方法	HierLR ^[10]	—	—	48%	656	191
	MLP4Geo ^[11]	38%	844	54%	554	120
基于社交关系方法	MADCEL-W ^[21]	58%	586	54%	705	116
	GCN-LP ^[16]	58%	576	53%	563	126
多源数据融合方法	MENET ^[17]	59.1%	570	50.5%	474	157
	GCN ^[16]	60%	546	65%	485	71
	ReLP ^[20]	62%	527	67%	418	47
	MetaGeo ^[23]	62%	533	63%	479	70
	HGNN ^[24]	63%	508.5	65%	423.2	59.5
	GEOTOP	64.3%	501.2	68.5%	409.4	41.3

表7 消融实验结果对比
Table 7 Comparison of results of ablation experiments

模型	大地区级定位准确率	平均误差/km	中位数误差/km	覆盖率
第一个实验	61.7%	518.3	54.2	98.2%
第二个实验	82%	83.7	7.4	3.3%
GEOTOP	64.3%	501.2	34.6	98.4%

表8 部分局部地名位置指示效果
Table 8 Geolocation performance of partial local toponyms

地名	指示位置	UL / TL	大地区级定位准确率
paerdegat	40.629, -73.906	1.00	100%
florin mall	38.497, -21.433	1.00	100%
Florida Agricultural and Mechanical University	30.426, -84.285	1.00	100%
suitland	38.848, -76.923	0.91	100%
anaheim stadium	33.8, -117.882	1.00	100%
kearny mesa	32.921, -117.117	0.87	100%
lansing mall	42.743, -84.625	0.86	100%
wheaton mall	38.803, -77.134	0.93	100%

部地名对目标用户的定位具有优良效果。

4 结束语

现有位置指示词提取方法提取地名不充分,遗漏了大量位置指示词。本文利用地名匹配提取用户文本中的地名,提取了更多位置指示词,并提出了多点地名识别算法和局部地名筛选算法,保证了所提地名只对应一个位置且被局部区域用户提及,提高了地名位置指示效果。实验结果表明,在两个公开数据集上,所提方法的所有指标优于当前最优方法。然而,本文方法没有利用用户关注关系和元数据,未来将获取这些数据,进一步探索多源数据的规律。

参考文献:

- [1] LIANG L, MALBASA V, VUCETIC S. Spatial scan for disease mapping on a mobile population[C]//AAAI Conference on Artificial Intelligence. 2014: 431-437.
- [2] TUMASJAN A, SPRENGER T O, SANDNER P G, et al. Predicting elections with Twitter: what 140 characters reveal about political sentiment[C]//International Conference on Weblogs and Social Media. 2010: 178-185.
- [3] HALIMI A, AYDAY E. Profile matching across online social networks[C]//22nd International Conference on Information and Communications Security. 2020: 54-70.
- [4] 顾秋阳, 吴宝, 池仁勇. 基于高阶路径相似度的复杂网络链路预测方法[J]. 通信学报, 2021, 42(7): 61-69.
GU Q Y, WU B, CHI R Y. Link prediction method based on the similarity of high path[J]. Journal on Communications, 2021, 42(7): 61-69.
- [5] 顾秋阳, 琚春华, 吴功兴. 基于子图演化与改进蚁群优化算法的社交网络链路预测方法[J]. 通信学报, 2020, 41(12): 21-35.
GU Q Y, JU C H, WU G X. Social network link prediction method based on subgraph evolution and improved ant colony optimization algorithm[J]. Journal on Communications, 2020, 41(12): 21-35.
- [6] CHENG Z Y, CAVERLEE J, LEE K. A content-driven framework for geolocating microblog users[J]. ACM Transactions on Intelligent Systems and Technology, 2013, 4(1): 1-27.
- [7] EISENSTEIN J, CONNOR B O, SMITH N A, et al. A latent varia-

ble model for geographic lexical variation[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10). 2010: 1277-1287.

- [8] HAN B, COOK P, BALDWIN T. Geolocation prediction in social media data by finding location indicative words[C]//Proceedings of the 24th International Conference on Computational Linguistics(COLING). 2012: 1045-1062.
- [9] JONES K S. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of Documentation, 1988, 60(5): 493-502.
- [10] WING B, BALDRIDGE J. Hierarchical discriminative classification for text based geolocation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP). 2014: 336-348.
- [11] RAHIMI A, COHN T, BALDWIN T. A neural model for user geolocation and lexical dialectology[C]//Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL). 2017: 209-216.
- [12] DAVIS J R C, PAPPAS G L, OLIVEIRA D R R, et al. Inferring the location of twitter messages based on user relationships[J]. Transactions in GIS, 2011, 15(6): 1361-1682.
- [13] KONG L, LIU Z, HUANG Y. Spot: Locating social media users based on social network context[J]. Proceedings of the VLDB Endowment, 2014, 7(13): 1681-1684.
- [14] JURGENS D. That's what friends are for: Inferring location in online social media platforms based on social relationships[C]//Proceedings of the Seventh International Conference on Weblogs and Social Media. 2013: 273-282.
- [15] MC-GEE J, CAVERLEE J, CHENG Z Y. Location prediction in social media based on tie strength[C]//Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM '13). 2013: 459-468.
- [16] RAHIMI A, COHN T, BALDWIN T. Semi-supervised user geolocation via graph convolutional networks[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 2009-2019.
- [17] DO T H, NGUYEN D M, TSILIGIANI E, et al. Twitter user geolocation using deep multiview learning[C]//Proceeding of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018: 6304-6308.
- [18] LE Q, MIKOLOV T. Distributed representations of sentences and documents[C]//Proceedings of the 31st International Conference on

Machine Learning. 2014: 1188-1196.

- [19] GROVER A, LESKOVEC J. node2vec: scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 855-864.
- [20] TIAN H C, ZHANG M, LUO X Y, et al. Twitter user location inference based on representation learning and label propagation[C]//Proceedings of The Web Conference. 2020: 2648-2654.
- [21] RAHIMI A, COHN T, BALDWIN T. Twitter user geolocation using a unified text and network prediction model[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics-7th International Joint Conference on Natural Language Processing(NAAACL-HLT '15). 2015: 630-636.
- [22] 王凯, 余伟, 吴敏, 等. 一种大数据环境下的在线社交媒体位置推断方法[J]. 软件学报, 2015, 26(11): 2951-2963.
WANG K, YU W, WU M, et al. Location inference method in online social media with big data[J]. Journal of Software, 2015, 26(11): 2951-2963.
- [23] ZHOU F, QI X, ZHANG K, et al. MetaGeo: a general framework for social user geolocation identification with few-shot learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 1-15.
- [24] ZHOU F, WANG T, ZHONG T, et al. Identifying user geolocation with hierarchical graph neural networks and explainable fusion[J]. Information Fusion, 2022, 81: 1-13.
- [25] ROLLER S, SPERIOSU M, RALLAPALLI S, et al. Supervised text based geolocation using language models on an adaptive grid[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012: 1500-1510.

[作者简介]



薛锦（1996-），男，内蒙古乌兰察布人，郑州大学硕士生，主要研究方向为社交网络定位与数据挖掘。



袁福祥（1991-），男，山东济宁人，博士，信息工程大学讲师，主要研究方向为网络目标定位、网络拓扑分析与网络空间测绘。



刘毅敏（1995-），女，山东烟台人，信息工程大学博士生，主要研究方向为网络安全与社交网络数据分析。



张萌（1996-），女，河南偃师人，信息工程大学博士生，主要研究方向为数据挖掘和社交网络分析。



乔亚琼（1981-），女，河南开封人，博士，华北水利水电大学讲师，主要研究方向为数据挖掘与社交网络分析。



罗向阳（1978-），男，湖北荆门人，博士，信息工程大学教授、博士生导师，主要研究方向为网络与信息安全。