基于多种提及关系的社交媒体用户位置推断

乔亚琼[1,2]，罗向阳[1,2]，马江涛[3]，李晨亮[4]，张萌[1,2]，李瑞祥[1,2]

（1.信息工程大学网络空间安全学院，河南 郑州 450001； 2. 数学工程与先进计算国家重点实验室，河南 郑州 450001；
3. 郑州轻工业大学计算机与通信工程学院，河南 郑州 450001；4. 武汉大学国家网络安全学院，湖北 武汉 430075）

摘 要：针对现有基于生成文本和社交关系的联合位置推断方法对社交媒体中异质数据间的位置关联性挖掘不充分的问题，提出了一种基于多种提及关系的社交媒体用户位置推断方法。首先，综合考虑社交媒体文本中用户之间的提及关系、用户对位置指示词的提及关系和用户对地理名词的提及关系，构建包含用户、位置指示词和地理名词 3 种节点的异质网络；其次，基于共同提及关系提出用户−词语−位置简化算法来构建用户−位置异质网络，将位置邻近的用户更为紧密地联系起来；再次，提出有偏的随机游走算法对图中节点采样以充分探索网络结构，缓解已知位置的稀疏性问题；最后，采用基于多层感知机的神经网络分类器对用户进行位置推断。在 GEOTEXT、TW-US 和 TW-WORLD 这 3 个代表性 Twitter 数据集上的实验结果表明，所提方法可显著提高用户位置推断准确率。

关键词：社交媒体；异质网络；用户位置推断；提及关系

中图分类号：TP391

文献标识码：A

doi: 10.11959/j.issn.1000−436x.2020229

Social media user geolocalization based on multiple mention relationships

QIAO Yaqiong[1,2], LUO Xiangyang[1,2], MA Jiangtao[3], LI Chenliang[4], ZHANG Meng[1,2], LI Ruixiang[1,2]

1. Information Engineering University, Zhengzhou 450001, China
2. State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China
3. School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450001, China
4. School of Cyber Science and Engineering, Wuhan University, Wuhan 430075, China

Abstract: Aiming at the problem that the existing joint user geolocalization methods based on social media text and social relationships do not sufficiently mine the location correlation between heterogeneous data in social media, a social media user geolocalization method based on multiple mention relationships was proposed. First, a heterogeneous network was constructed by comprehensively considering the mention relationship between users, the user's mention relationship with location indicative words, and the user's mention relationship with geographic nouns. Then, a network simplification strategy was proposed to construct a user-location heterogeneous network that connects users live nearby more closely based on the common mention relationship. After that, a biased random walk algorithm was proposed for the graph node sampling to fully explore the network structure and alleviate the sparsity problem of known locations. Finally, a neural network classifier based on a multilayer perceptron was used to infer the user's location. Experimental results on three representative Twitter data sets of GEOTEXT, TW-US and TW-WORLD show that the proposed method can significantly improve the user geolocalization accuracy.

Key words: social media, heterogeneous network, user geolocalization, mention relationship

---

收稿日期：2020−06−11；**修回日期**：2020−07−20

通信作者：罗向阳，luoxy_ieu@sina.com

基金项目：国家自然科学基金资助项目（No.U1804263, No.U1636219, No.61872287, No.U1736214）；国家重点研发计划基金资助项目（No.2016QY01W0105, No.2016YFB0801303）；中原英才计划−中原科技创新领军人才基金资助项目（No.1052020KJLJ0025）；河南省科技创新人才计划基金资助项目（No.184200510018）；河南省科技攻关基金资助项目（No.202102310237）

Foundation Items: The National Natural Science Foundation of China (No.U1804263, No.U1636219, No.61872287, No.U1736214), The National Key Research and Development Program of China (No.2016QY01W0105, No.2016YFB0801303), Zhongyuan Talents Program-Zhongyuan Science and Technology Innovation Leading Talent Project (No.1052020KJLJ0025), The Plan for Scientific Innovation Talent of Henan Province (No.184200510018), The Scientific and Technological Project of Henan Province (No.202102310237)

1　引言

社交媒体用户位置推断是从社交媒体数据中挖掘用户位置信息。社交媒体用户位置推断技术主要用于对社交媒体用户所在的地理位置进行分析和定位，可为基于位置的服务[1]、基于位置的事件分析[2]和基于位置的敏感人物分析[3]提供帮助。然而，出于对个人隐私保护的考虑[4-5]，社交媒体中的位置数据十分稀疏[6]。因此，有必要开展社交媒体用户位置推断问题研究，以应对位置数据的稀疏性问题。

常见的社交媒体用户位置推断方法通过提取社交媒体文本中与位置相关的话题、位置指示词或地理名词等特征推断用户位置。社交媒体上讨论的话题通常因地理区域而异，因此，Eisenstein 等[7]和 Ahmed 等[8]使用主题模型建模话题与位置的关系来推断用户位置。社交媒体文本使用的词语具有地理位置偏向性，Wing 等[9]通过基于词语的信息增益率提取位置指示词来推断用户位置。统计分析结果表明，如果用户经常提到某个地理名词，他很可能生活在该地理区域，因此可以使用文本中提及的地理名词来推断用户位置[10]。常用的地名词典有 GeoNames 和 DB-pedia。Rahimi 等[11-12]使用词袋模型提取文本特征，然后结合逻辑回归分类器或多层感知机分类器推断用户位置。

除了基于文本的用户位置推断，基于用户社交关系的位置推断也比较常见。基于用户社交关系的方法假设有关注关系或者有提及关系的用户地理位置接近[13]。此类方法通过使用用户的关注关系或者用户在文本中的提及关系构建图 1 所示的同质网络来推断用户位置。如 Rahimi 等[14]提出的 MADCEL-W 方法利用用户的提及频次构建加权的用户社交网络，并去除名人节点，基于改进的吸附传播算法推断用户位置。Rahimi 等[15]提出的 GCN-LP 方法将用户邻居节点的独热编码作为用户节点特征，使用用户的提及关系构建用户的社交网络，通过图卷积网络推断用户位置。

基于文本的方法忽略了用户朋友对位置的影响，可达到的精度有限；基于社交关系的方法无法对无朋友的孤立用户进行位置推断。为此，学者们尝试使用文本和社交关系 2 种视图联合推断用户位置[16]。如 Rahimi 等[12]提出的 MADCEL-W-MLP 方法，首先基于用户之间的提及关系构建用户的社交网络，然后将基于文本的推断结果作为附加节点与对应用户节点相连，使用标签传播算法推断用户位置。Rahimi 等[15]提出的 GCN 方法将用户文本的词袋特征作为用户特征，使用用户的提及关系构建用户的社交网络，通过图卷积网络联合文本视图和网络视图推断用户位置。Rahimi 等[15]提出的 MLP-TXT+NET 方法将基于词袋模型提取的文本特征和用独热编码表示的社交关系特征串联，利用多层感知机分类器推断用户位置。Zhong 等[17]提出基于注意力机制的图神经网络模型，联合文本内容和社交网络推断用户位置。

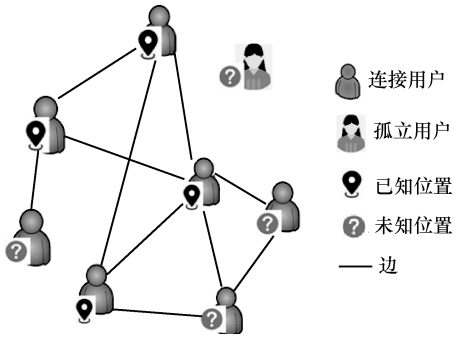


图 1　同质网络

尽管联合推断的方法在一定程度上降低了位置推断误差，却没有有效利用文本中的位置特征，且仅使用用户之间的提及关系构建网络，忽略了文本中位置指示词和地理名词对用户位置的指示性，导致位置推断误差仍然较大。为此，本文提出一种基于多种提及关系的社交媒体用户位置推断方法。该方法首先从用户文本中提取用户提及的朋友、位置指示词和地理名词；其次，根据用户之间的提及关系、用户对位置指示词的提及关系和用户对地理名词的提及关系，构建包含用户、词语（位置指示词和地理名词）和位置 3 种节点的异质网络；再次，基于用户对位置指示词和地理名词的共同提及关系提出一种异质网络简化方法，将地理位置邻近的用户更紧密地联系起来；为了充分探索网络结构，缓解已知位置的稀疏性问题，提出使用有偏的随机游走算法对网络中的节点采样以生成节点序列，用于用户特征向量的学习；最后，基于学习得到的用户特征向量，提出利用多层感知机分类器对用户进行位置推断。

本文的主要贡献如下。

1) 提出一种基于多种提及关系的社交媒体用户位置推断方法。与已有方法相比，该方法有效地

集成了文本中提取的位置特征，能够基于用户与位置指示词的提及关系、用户对地理名词的提及关系，以及用户之间的提及关系，将文本视图和用户关系视图更好地结合起来，联合推断用户位置，并将孤立用户连接到网络中，有效降低用户定位误差并提高可定位用户比例。

2) 提出一种新颖的用户表示学习方法。与现有的仅基于用户之间的提及关系构建社交网络并进行特征向量学习的方法不同，本文提出基于多种提及关系构建异质图，并根据用户对位置指示词和地理名词的共同提及关系对异质图进行简化，将位置邻近的用户更紧密地连接起来，并基于有偏的随机游走算法生成节点序列以学习用户特征向量，使地理位置邻近用户的特征向量距离更近。

3) 提出基于用户表示学习和神经网络分类器推断用户位置。与现有基于标签传播的位置推断算法相比，本文方法可以有效缓解已知位置数据的稀疏性问题，更好地利用网络结构推断用户位置，有效提高用户定位准确率。

2 问题描述

为了便于理解，本节给出本文要解决问题的定义和文中用到的主要符号及其含义。

给定社交媒体数据集 $D=(U,T_u)$，该数据集包含位置已知的用户集合 U^{L}、位置未知的用户集合 U^{N} 和用户发布的文本集合 T，T_u 表示用户 $u\in U$ 的推文集合。则用户集合 $U=U^{\mathrm{L}}\cup U^{\mathrm{N}}$。$U^{\mathrm{L}}$ 对应的位置集合为 Y^{L}。由于数据集中的双向提及十分稀疏，本文基于用户在文本中的单向提及构建用户社交网络，用户之间的社交关系集合用 F 表示。此外，用户的位置集合 L 已知。本文将用户位置推断问题视为分类问题，用户所在区域使用 k-d 树的划分方法进行区域划分[18]，每个网格代表一个位置类别，l_i^u 表示用户 u 所在的位置区域为 l_i。假设待推断的用户位置包含在已知的位置集合中，社交网络用户位置推断问题可以用式(1)描述，即通过对用户、用户文本和用户已知位置的分析，推断出 U^{N} 中用户的位置集合 Y^{N}。

$$f_1:(U,T,Y^{\mathrm{L}})\to Y^{\mathrm{N}} \tag{1}$$

下面给出本文用到的一些重要术语的定义。

定义 1 信息增益率。本文使用 Han 等[19]提出的方法基于信息增益率提取位置指示词。首先，对数据集进行数据预处理，去除停用词，得到词语集合 M；然后，计算数据预处理后的每个词语的信息增益率 IGR(m)，m 表示集合 M 中的词语，如式(2)所示。

$$\mathrm{IGR}(m)=\frac{\mathrm{IG}(m)}{\mathrm{IV}(m)} \tag{2}$$

其中，IG(m)表示词语 m 的信息增益，IV(m)表示词语 m 的信息熵。

$$\begin{aligned}\mathrm{IG}(m)&=H(L)-H(L\mid m)\propto -HL\mid m\propto\\ &\mathrm{r}(m)\sum\nolimits_{l\in L}\Pr(l\mid m)\log\Pr(l\mid m)+\\ &\Pr(\bar{m})\sum\nolimits_{l\in L}\Pr(l\mid m)\log\Pr(l\mid\bar{m})\end{aligned} \tag{3}$$

$$\mathrm{IV}(m)=-\Pr(m)\log\Pr(m)-\Pr(\bar{m})\log\Pr(\bar{m}) \tag{4}$$

其中，Pr(m)和 Pr($\bar{m}$)分别表示文本中包含词语 m 和不包含词语 m 的概率；Pr (l|m)表示包含词语 m 的文本来自位置 l 的条件概率，Pr P($l|\bar{m}$)表示文本中不包含词语 m 时来自位置 l 的条件概率，$l\in L$。

定义 2 位置特征词。位置特征词包括位置指示词和地理名词。位置指示词有强烈的位置指示性[20]，具有紧凑的地理使用范围[21]，根据词语在不同位置被提及的统计特征筛选得到。例如，howdy 在美国德克萨斯州是一个典型的问候语，它提示用户在德克萨斯州或附近，而 august、peace 和 email 等词不具有位置指示性[21]。地理名词是表示地理位置的名词，如 Arizona。地理名词可以借助地名词典识别[22]，不需要借助词语的统计特征。

定义 3 用户−地理名词矩阵。$\boldsymbol{P}$ 是一个$|U|\times|M_p|$维矩阵，$\boldsymbol{P}[i]$是用户 u_i 的地理名词向量，$P[i][j]$表示用户 u_i 提及第 j 个地理名词 m_p 的次数。

定义 4 用户−位置指示词矩阵。$\boldsymbol{R}$ 是一个$|U|\times|M_l|$维矩阵，$\boldsymbol{R}[i]$是用户 u_i 位置指示词向量，$R[i][j]$表示用户 u_i 提及第 j 个位置指示词 m_l 的次数。

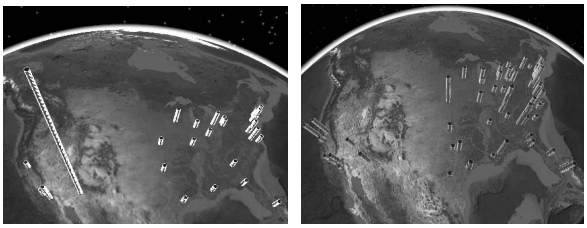
定义 5 用户−词语−位置异质网络。$G=(V, E, W)$，其中 $V=V_{\mathrm{U}}\cup V_{\mathrm{M}}\cup V_{\mathrm{L}}$ 表示顶点集合，$V_{\mathrm{U}}=U$，$V_{\mathrm{M}}=M$，$V_{\mathrm{L}}=L$；$M=M_l\cup M_p$ 表示位置特征词集合，M_l 表示位置指示词集合，M_p 表示地理名词集合。E 表示边的集合，包含根据用户之间的提及关系建立的用户−用户边(u, u)、根据用户对位置指示词的提及关系建立的用户−位置指示词边(u, m_l)、根据用户对地理名词的提及关系建立的用户−地理名词边(u, m_p)，以及根据位置指示词与其归属关系建立的位置指示词−位置边(m_l,l)、根据地理名词与其位置的归属关系建立的地理名词−位置边(m_p,l)。边的权重依次为用户之间的提及次数、用户对位置指示词的提

及次数和用户对地理名词的提及次数、位置指示词−位置边和地理名词−位置边的权重为 1。此外，由于名人用户的社交关系复杂，其关注者或者提及的用户的位置分散，为了避免名人用户带来的偏差，本文将用户朋友数量大于阈值 γ 的用户视为全局名人[14]，从异质网络中剔除。

定义 6　用户−位置异质网络。G'基于 G 简化得到。$G'=(V', E', W')$，其中，$V'=V'_U \cup V'_L$，为顶点集合，$V'_U=U$，$V'_L=L$；E'表示边的集合，包含用户−用户边(u, u)、用户−位置边(u, l)；W'为边的权重集合，用户−用户边及其权重根据用户之间的提及频次和用户对位置特征词的共同提及频次构建和计算；用户−位置边的权重根据用户对位置特征词的提及关系和位置特征词的位置归属关系构建，其权重根据用户对位置特征词的提及频次计算。G'的详细构建方法见 4.1 节。

3　数据分析

本节基于真实的 Twitter 数据集 GEOTEXT[7]进行数据分析，展示位置特征词的位置指示性。图 2 给出了 Arizona 和 email 在 GEOTEXT 数据集中被提及频次的空间分布。



(a) Arizona　　(b) email

图 2　GEOTEXT 中 Arizona 和 email 被提及频次的空间分布

图 2 中柱体表示该词在该位置被提及，柱体的高度为该词语被提及的频次。可以看出，email 分布范围广，在各个地区被提及的频次相差不大，不具有位置指示性。Arizona 则被生活在亚利桑那州及其附近的用户多次提及，具有明显的位置指示性。

表 1 给出了 GEOTEXT 数据集基于信息增益率和字典匹配提取的部分位置指示词和地理名词。其中，l_7、l_{20}、l_{23}、l_{29} 和 l_{55} 为按照文献[15,23]方法，基于 k-d 树对连续空间的进行划分后得到的位置标签；地理名词的位置根据其表示的地理位置的坐标确定，位置指示词的位置基于以下的方法来确定。

对于位置指示词 m_l，设该词在所有位置出现的总次数为 n，在位置 k 出现的次数为 n_k。则位置 k 出现该词的概率为 $\varepsilon=\frac{n_k}{n}$。当 ε 最大值唯一且满足 $\varepsilon>\frac{1}{N}$ 时，位置指示词 n_l 的位置是 k，其中 N 为该词出现次数不为 0 的位置数。

从表 1 可以看出，词语在社交媒体中的使用具有明显的地域特征。Austin 和 Dallas 被生活在位置 l_{20} 的用户较多地提及，l_{20} 的中心地理坐标为(−97.30,32.63)，Austin 的地理坐标为(−97.10953, 33.08234)，Dallas 的地理坐标为(−97.10953, 33.08234)。可以看出，这 2 个地理名词表示的地点奥斯汀和达拉斯在 l_{20} 表示的地理区域内。Chicago 的缩写 Chi 也较多地被该城市所属的位置区域内的用户所提及。

但是，本文也观察到，由于训练集中的数据偏差（例如 l_{92} 仅包含一个用户且只有少量推文），基于信息增益率获取位置指示词的方法无法提取某些位置的位置指示词。因此，可以得出结论，由于位置指示词基于词语在不同区域中使用的统计特征提取，受训练集的数据影响非常大，其对用户的位置指示性有限，这也是基于文本的位置推断方法准确率不高的原因之一。相比之下，地理名词只需要查询地理词典，不需要任何训练数据，且其本身具有明显的地域特征，因此，地理名词对用户的位置影响更为显著。

4　本文算法描述

如图 3 所示，本文提出的方法包括基于文本的位置特征提取、用户−词语−位置异质网络构建、用户−位置异质网络构建、基于有偏随机游走的用户表

表 1　词语在不同位置的分布（**GEOTEXT 数据集**）

位置标签	经度	纬度	地理名词						
l_7	−111.88°	33.39°	**Arizona**	softball	bubbles	josh	cracker	johnson	pissing
l_{20}	−97.30°	32.63°	**Austin**	**Dallas**	coconut	fallon	badu	chorus	jamming
l_{23}	−96.08°	39.7°	**kansas**	lambert	machines	insomnia	mayo	guts	slang
l_{29}	−87.63°	41.73°	**Chicago**	**Chi**	lbs	peppers	greg	julius	dez
l_{55}	−82.29°	39.32°	**Columbus**	**nashville**	sangria	gummy	graphic	ditto	lauryn

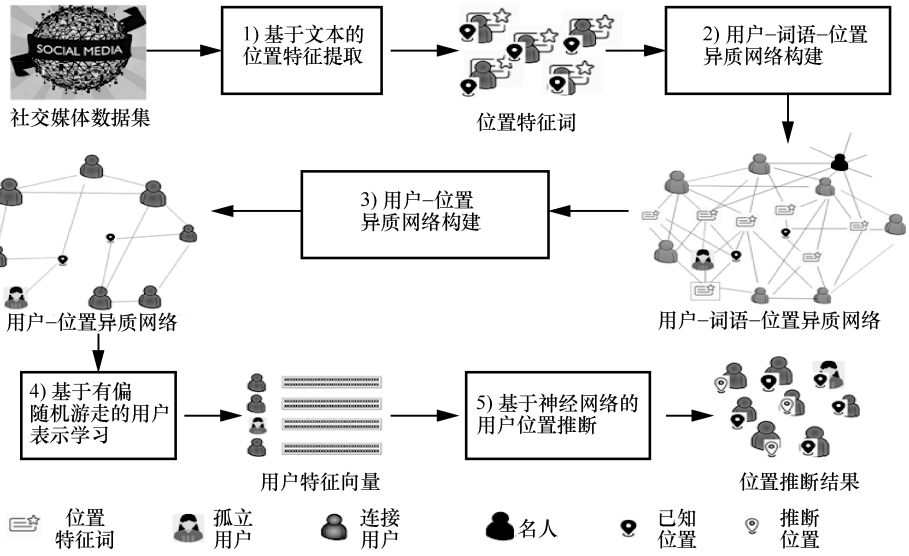


图 3　基于多种提及关系的社交媒体用户位置推断原理示意

示学习和基于神经网络的用户位置推断 5 个部分。

基于文本的位置特征提取包括基于信息增益率的位置指示词提取和基于地名词典的地理名词发现。由于用户文本中使用的词语中包含大量停用词和与用户位置无关的词语，使用全部的词语构建用户−词语−位置异质网络会导致网络结构复杂，增加计算开销。因此，本文基于词语信息增益率对词语进行初步筛选，过滤信息增益率较小的词语，以识别位置指示词，减少计算开销。此外，本文基于 GeoNames 来识别文本中的地理名词。

在提取文本中的位置特征词之后，根据定义 5 给出的方法构建图 3 所示的用户−词语−位置异质网络。为了将相同位置区域的用户更紧密地联系起来，使地理位置邻近的用户的特征向量距离更近，本文提出基于提及关系简化用户−词语−位置异质网络，以构建用户−位置异质网络，并提出基于有偏随机游走的用户表示学习算法学习用户特征向量。下面详细阐述这 2 个算法。

4.1　用户−位置异质网络构建

得到用户−词语−位置异质网络后，本文基于共同提及关系对其进行简化，通过去除词语节点将位置相近的用户更紧密地联系起来以构建用户−位置异质网络 $G'=(V',E',W')$，如算法 1 所示。

算法 1　用户−位置异质网络构建算法（GELP）

输入　用户−词语−位置异质网络 $G=(V,E,W)$，用户发布的推文集合 T

输出　用户−位置异质网络 $G'=(V',E',W')$

1) $G'\leftarrow G$ //初始化
2) for i in range($|V_U|$) do
3)　for j in range(i+1, $|V_U|$) do
4)　$m_p, s_p\leftarrow$FindMaxMLN($\boldsymbol{P}[i],\boldsymbol{P}[j]$)
5)　if $s_p>0$ then
6)　if G'.has_edge(u_i,u_j) then
7)　$G'.[u_i][u_j]$['weight'] += s_p
8)　else
9)　G'.add_edge(u_i,u_j,weight= s_p)
10)　G'.add_edge(u_i,l_p,weight= $P[i][k]$)
11)　G'.add_edge(u_j,l_p,weight= $P[j][k]$)
12)　end if
13)　end if
14)　$s_l\leftarrow$SumMLIW($\boldsymbol{R}[i],\boldsymbol{R}[j]$)
15)　if $s_l>0$ and G′.has_edge(u_i, u_j) then
16)　$G'.[u_i][u_j]$['weight'] += s_l
17)　end if
18)　end for
19) end for
20) RemoveNodes(V_M)
21) return $G'=(V',E',W')$

根据第 3 节的分析结果，在去除词语节点时，对于地理名词，如果 2 个用户之间没有边，且他们共同提及同一地理名词的次数大于阈值 τ_1，则在这 2 个用户之间添加边，用户−用户边的权重为用户对所有地理名词的最大共同提及次数 s_p，对应的地理名词记为 m_p，是第 k 个地理名词。如果用户之间已

有边，则将用户对所有地理名词的最大共同提及次数与已有边的权重相加作为用户−用户边的权重。同时，将与共同提及次数最多的地理名词相连的用户节点和位置节点 l_p 直接相连，用户−位置边的权重为用户对地理名词的提及次数。对于位置指示词，如果用户对同一位置指示词的共同提及次数大于阈值 τ_2，且用户之间有边，则用户−用户边的权重为原有边的权重和用户对所有位置指示词的共同提及次数的累加 s_l。为了避免位置指示词带来的噪音，本文不根据位置指示词添加用户−用户边和用户−位置边。下面给出用户−位置异质网络构建算法。

4.2　有偏随机游走采样

用户−位置异质网络构建的目的是将位置邻近的用户更紧密地联系起来，且将用户节点与其邻近的位置节点紧密关联。为了更好地保留节点的邻域特征，本文提出有偏随机游走策略对用户−位置异质网络中的节点进行采样生成节点序列。

算法 2　有偏随机游走算法

输入　用户−位置异质网络 $G'=(V',E')$，单次游走长度 μ_1，采样长度 μ_2

输出　节点序列 $\boldsymbol{S}$

1) $\boldsymbol{S}\leftarrow\varnothing$//初始化
2) for iter1←1 to μ_2 do
3)　for $v_i\in\boldsymbol{V}'$ do
4)　$\boldsymbol{S}_i\leftarrow[v_i]$
5)　for iter2←1 to μ_1 do
6)　CurrNode=$\boldsymbol{S}_i$[−1]//获取当前节点
7)　NbrNode=GetNbrNode(G',CurrNode)
8)　NxtNode=GetNxtNode(NbrNode,$\boldsymbol{\psi}$)
9)　append NxtNode to $\boldsymbol{S}_i$
10)　end for
11)　end for
12)　append $\boldsymbol{S}_i$ to $\boldsymbol{S}$
13) end for
14) return $\boldsymbol{S}$

受 Grover 等[24]工作启发，在节点采样时本文使用回归参数 r 控制在随机游走中选择上一个节点作为下一个节点的可能性，使用进出参数 q 控制游走方向是“向内”或“向外”。r>1 时减少对已访问的节点进行采样的可能性，r<1 时随机游走徘徊在初始节点周围。q>1，则随机游走倾向于选择接近上一个节点的节点，这种游走类似于广度优先采样，采样得到的节点序列捕获初始节点附近的局部视图；q<1，则倾向于选择远离上一个节点的节点，这种游走向外采样，类似于深度优先采样。给定初始节点 v_i，则节点序列 $\boldsymbol{S}_i$ 根据式(5)定义的转移概率生成，其中 $\varphi(v_i,v_{i-1})$表示随机游走从节点 v_{i-1} 游走到节点 v_i 的概率，转移概率矩阵为 $\boldsymbol{\psi}$。假设节点 v_{i-2}，v_{i-1}，$v_i\in E'$。v_{i-2} 是 v_{i-1} 的前一个节点，d=0 表示随机游走从 v_{i-1} 回到 v_{i-2}；d=1 表示随机游走从 v_{i-1} 到与 v_{i-2} 直接相连的节点；d=2 表示随机游走从 v_{i-1} 到与 v_{i-2} 不直接相连的节点，Z 为归一化常数。下面给出有偏随机游走的算法步骤。

$$\varphi(v_i\mid v_{i-1})=\begin{cases}\dfrac{1}{Zr}e_{i-1,i}, d=0\\ \dfrac{1}{Z}e_{i-1,i}, d=1\\ \dfrac{1}{Zq}e_{i-1,i}, d=2\end{cases}\tag{5}$$

由于本文的目标是学习图中所有节点的特征向量，因此最终节点序列通过对图中每个节点进行 μ_2 次采样生成。最终的节点序列长度为$|V'|\mu_2\mu_1$。

得到节点序列后，将有偏随机游走得到的节点序列作为输入学习用户特征向量，本文用 skip-gram 模型[25]来解决用户特征向量学习的问题。

4.3　基于神经网络的用户位置推断

在得到用户的特征向量后，本文将其作为多层感知机的输入训练用户位置推断模型，模型的输出为基于 k-d 树的区域划分后的位置类别。

$\vec{\boldsymbol{x}}$ 是用户的特征向量，$\sigma()$为激活函数，本文取 ReLU 函数[26]为激活函数，k 为隐含层的数量，本文设置 k=1[12]，$\boldsymbol{b}_1$、$\boldsymbol{b}_k$、$\boldsymbol{b}_o$ 为偏差向量，$\vec{\boldsymbol{h}}^1$、$\vec{\boldsymbol{h}}^k$、$\vec{\boldsymbol{h}}^o$ 分别为第一、第 k 个隐含层和最后一层神经网络的输出。多层感知机的参数使用 Lasagne/Theano[27]基于 Adam 方法[28]进行优化。

$$\begin{cases}\vec{\boldsymbol{h}}^1=\sigma(W^1\vec{\boldsymbol{x}}+\boldsymbol{b}_1), & \text{第一个隐含层}\\ \vec{\boldsymbol{h}}^k=\sigma(W^k\vec{\boldsymbol{h}}^{k-1}+\boldsymbol{b}_k), & \text{第 } k \text{ 个隐含层}\\ \vec{\boldsymbol{h}}^o=\mathrm{SoftMax}(W^o\vec{\boldsymbol{h}}^k+\boldsymbol{b}_o), & \text{输出层}\end{cases}\tag{6}$$

5　性能测试与分析

为了验证本文提出的方法，本文使用 3 个真实 Twitter 数据集 GEOTEXT[7]、TW-US[18] 和 TW-WORLD[29]来验证算法的有效性。

5.1 实验设置

1) 实验数据

GEOTEXT 和 TW-US 数据集包含由来自美国的用户发布的推文。GEOTEXT 使用每个用户的第一条推文位置作为用户的基准位置[7]，TW-US 使用每个用户发布的带有位置标签的推文的中位数位置作为该用户的基准位置。TW-WORLD 包含来自全球的用户发布的推文，提取每个用户大部分推文位置附近的城市中心作为用户的基准位置。3 个数据集的统计数据如表 2 所示。

2) 评价标准

本文使用平均误差 mean、中位数误差 median、Acc@161 和覆盖率 coverage 来评估所提的位置推断方法的性能，其中，Acc@161 为推断位置与实际位置距离小于 161 km 的用户位置推断准确率。用户覆盖率为可定位的用户占所有用户的百分比。

3) 参数设置

对于本文用到的 3 个数据集，本文按照 Rahimi 等[12]的方法，基于 k-d 树对连续空间进行划分，以确保每个区域内有相似数量的用户。根据 Rahimi 等[12]的经验，本文将 GEOTEXT，TW-US 和 TW-WORLD 这 3 个数据集基于 k-d 树划分的参数依次设置为 50、2 400 和 2 400，分别生成了 129、256 和 930 个位置标签。与 Rahimi 等[12]的工作保持一致，本文每个区域内所有用户位置经度、纬度的中位数作为该区域位置标签的地理坐标。

此外，在构建用户−词语−位置异质网络时，本文将名人节点去除的阈值 γ 在 GEOTEXT、TW-US 和 TW-WORLD 上依次设置为 5、15 和 5。基于信息增益率选取候选位置指示词集的阈值设为 0.25[19]。在学习用户特征向量时，本文将有偏随机游走的参数设置为 r=4，q=0.25，μ_2=10，μ_1=80。

5.2 实验结果

本节将提出的用户位置推断方法与经典的方法进行比较，并对实验结果进行分析。

表 3 给出了所提方法与经典方法的性能对比。可以看出，在 3 个数据集上，所提方法在 Acc@161、mean、median 上的表现均优于所有经典方法。在 GEOTEXT 数据集上，Acc@161 比性能最好的 MAGNN 高出 2%，平均误差降低 25 km。表明本文提出的异质网络的社交媒体用户位置推断方法可以通过用户和位置的关系加强用户之间的联系，提高用户位置推断性能。

5.3 不同提及关系的影响分析

为了探索不同提及关系对用户位置推断的性

表 2 数据集的统计信息

数据集	推文数/个	用户提及数/个	用户总数/个	训练集用户数/个	测试集用户数/个	开发集用户数/个	每个用户推文数/条
GEOTEXT	378×10^3	109×10^3	9 475	5 685	1 895	1 895	39
TW-US	38×10^6	3.63×10^6	450×10^3	430×10^3	10×10^3	10×10^3	84
TW-WORLD	12×10^6	16.8×10^6	1.4×10^6	1.38×10^6	10×10^3	10×10^3	9

表 3 在 3 个 Twitter 数据集上的位置推断性能

数据来源	方法	GEOTEXT			TW-US			TW-WORLD		
		Acc@161	mean/km	median/km	Acc@161/km	mean/km	median/km	Acc@161	mean/km	median
Text	HierLR+k-d tree[9]	—	—	—	48%	686	191	31%	1 669	509
	(MLP + k-d tree)[12]	38%	844	389	54%	554	120	34%	1 456	415
	LR[11]	38%	880	397	50%	686	159	32%	1 724	450
Network	MADCEL-W[14]	58%	586	60	54%	705	116	45%	2 525	279
	GCN-LP[15]	58%	576	56	53%	563	126	45%	2 357	279
Text+Network	MLP-TXT+NET[15]	58%	554	58	66%	420	56	58%	1 030	53
	MADCEL-W-MLP[12]	59%	578	61	61%	515	77	53%	1 280	104
	GCN[15]	60%	546	45	62%	485	71	54%	1 130	108
	MAGNN[17]	62%	514	38	63%	452	67	55%	1 107	102
	GELP	64%	489	39	67%	430	44	59%	1 015	50

能的影响，本文提出以下 4 种用户–位置异质网络构建方法，并进行了对比测试。

1) GELP-MEW。通过将提及同一词语的节点直接相连去除用户–词语–位置异质网络中的词语节点。去除词语节点之后，用户–用户边的权重为共同提及位置指示词的最小次数加上用户之间的提及次数，用户–位置边的权重为用户对词语的提及次数。

2) GELP-MW。在去除词语节点时，仅将连接同一词语的用户节点和位置节点相连。去除词语节点之后，用户–用户边的权重为共同提及位置指示词的最小次数加上用户之间的提及次数，用户–位置边的权重为用户对词语的提及次数。

3) GELP-I。在去除词语节点时，仅将与该词相邻的孤立用户以及训练集中的用户和位置节点相连。去除词语节点之后，用户–用户边的权重为 1，用户–位置边的权重为 1。

4) GELP-UW。在去除词语节点时，仅将训练集中的用户节点和位置节点相连。去除词语节点之后，用户–用户边的权重为用户之间的提及次数之和，用户–位置边的权重为 1。

不同方法在GEOTEXT数据集上的用户位置推断结果如表 4 所示。

表 4　不同异质网络构建方法的位置推断结果（GEOTEXT 数据集）

方法	Acc@161	mean/km	median/km	覆盖率
GCN[15]	60%	546	45	95.52%
GELP-MEW	38%	996	465	**99.86%**
GELP-MW	55%	683	79	**99.86%**
GELP-I	60%	553	48	**99.86%**
GELP-UW	63%	499	**38**	98.28%
GELP	**64%**	**494**	**38**	98.28%

从表 4 中可以看出，GELP 取得了最好的位置推断结果，并且具有较高的用户覆盖率。虽然其用户覆盖率不是最高，但仍然高出典型的用户位置推断算法（CGN）2.76%。GELP-UW 的位置推断结果比 GELP 稍差，与 GELP 的用户覆盖率相同。GELP-I、GELP-MW 和 GELP-MEW 具有最大的用户覆盖率，GELP-MEW 表现最差。

结合以上实验结果，本文得到以下结论。

1) 使用用户对词语的共同提及关系连接用户节点，会降低用户位置推断性能。原因是，尽管用户对位置指示词的提及反映了用户的位置，但是由于用户共同提及的位置指示词有多个，且这些位置指示词可能指示不同的位置，简单地根据用户对位置指示词的共同提及添加用户–用户边，引入了大量的噪声，从而降低了用户位置推断的性能。

2) 使用用户对位置指示词的提及添加开发集和测试集的用户–位置节点，并基于用户对位置指示词提及的次数对用户–用户边加权，可以提高可定位用户比例，但是不能提高用户位置推断准确率。原因是训练集上的数据偏差使基于信息增益率获取的位置指示词包含噪声词汇，基于位置指示词添加用户–位置边引入了噪声，降低了用户位置推断准确率。

3) 使用用户提及的位置特征词仅将孤立用户连接到网络中，可以增加可定位用户比例，并保持较高的用户位置推断准确率。

4) 使用用户之间的提及关系对用户–用户边加权，不能提高用户位置推断准确率，说明用户的提及次数并不代表用户位置的紧密度。

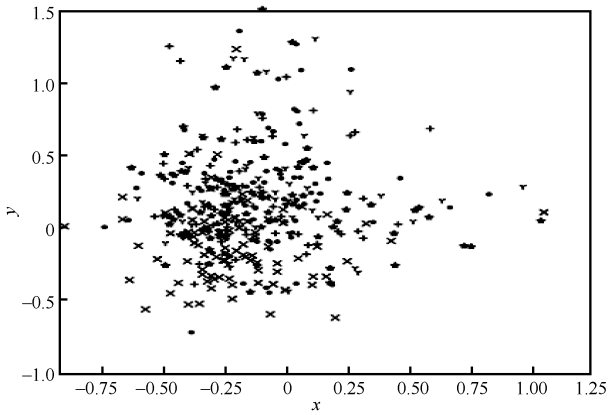
5.4　用户特征向量的可视化

图 4 给出了在 GEOTEXT 数据集中随机选择的 5 个区域内对用户特征向量进行主成分分析（PCA, principal component analysis）降维后的可视化结果。可以看出，与 GELP-MEW 相比，GCN 和 GELP 的用户特征向量在不同的位置的可分辨能力较强。与 GCN 和 GELP-MEW 相比，GELP 可以将相同区域的用户更好地聚集在一起。

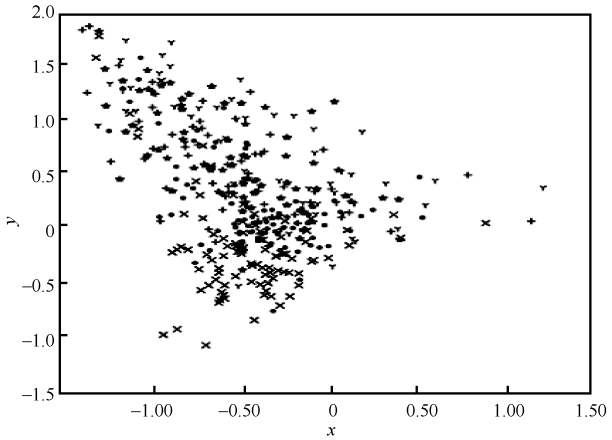
6　结束语

本文提出了一种多种提及关系的社交媒体用户位置推断方法。通过综合考虑用户之间的提及关系、对位置指示词和地理名词的提及关系构建异质网络，探索了异质社交网络中用户的位置推断方法。将异质网络用于用户位置推断是一个新的尝试，基于异质网络中丰富的异质信息，可以更全面地捕获用户的位置特征，在真实 Twitter 数据集上的大量实验表明，本文提出的方法有效提高了位置推断的准确率和覆盖率，降低了平均误差和中位数误差。

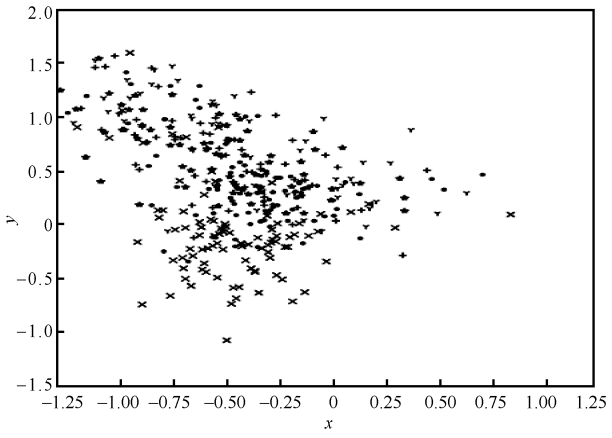
尽管本文方法获得了较好的效果，但如何使用更大规模的异质数据源构建异质网络仍然有待进一步研究。将来的工作中将对此进一步开展相关研究。



(a) GELP-MEW



(b) GCN



(c) GELP

图 4　来自 GEOTEXT 数据集的 5 个随机选择区域中的用户嵌入的 PCA 可视化效果比较

参考文献：

[1] KAWANAKA S, MORIWAKI D. Uplift modeling for location-based online advertising[C]//Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising. New York: ACM Press, 2019: 1-4.

[2] SHAHRAKI Z K, FATEMI A, MALAZI H T. Evidential fine-grained event localization using Twitter[J]. Information Processing & Management. 2019, 55(6):102045.1-102045.20.

[3] ZHOU N, ZHAO W X, ZHANG X, et al. A general multi-context embedding model for mining human trajectory data[J]. IEEE Transactions on Knowledge & Data Engineering, 2016, 28(8): 1945-1958.

[4] 张文静, 刘樵, 朱辉, 等. 基于信息论方法的多等级位置隐私度量与保护[J]. 通信学报, 2019, 40 (12): 51-59.
ZHANG W J, LIU Q, ZHU H, et al. Evaluation and protection of multi-level location privacy based on an information theoretic approach[J]. Journal on Communications, 2019, 40(12): 51−59.

[5] 李维皓, 曹进, 李晖, 等. 基于位置服务隐私自关联的隐私保护方案[J]. 通信学报, 2019, 40(5):57−66.
LI W H, CAO J, LI H, et al. Privacy self-correlation privacy-preserving scheme in LBS[J]. Journal on Communications, 2019, 40(5):57−66.

[6] POULSTON A, STEVENSON M, BONTCHEVA K. Hyperlocal home location identification of Twitter profiles[C]// Proceedings of the 28th ACM Conference on Hypertext and Social Media. New York: ACM Press, 2017:45-54.

[7] EISENSTEIN J, O'CONNOR B, SMITH N A, et al. A latent variable model for geographic lexical variation[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2010: 1277-1287.

[8] AHMED A, HONG L, SMOLA A J. Hierarchical geographical modeling of user locations from social media posts[C]//Proceedings of the 22nd International World Wide Web Conference. New York: ACM Press, 2013: 25-36.

[9] WING B, BALDRIDGE J. Hierarchical discriminative classification for text-based geolocation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 336-348.

[10] SCHULZ A, HADJAKOS A, PAULHEIM H, et al. A multi-indicator approach for geolocalization of tweets[C]//Proceedings of the Seventh International Conference on Weblogs and Social Media. Palo Alto: AAAI Press, 2013: 573-582.

[11] RAHIMI A, VU D, COHN T, et al. Exploiting text and network context for geolocation of social media users[C]//The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2015: 1362-1367.

[12] RAHIMI A, COHN T, BALDWIN T. A neural model for user geolocation and lexical dialectology[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2017, 2: 209-216.

[13] 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. Palo Alto: AAAI Press, 2013: 273-282.

[14] 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 and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 630-636.

[15] 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. Stroudsburg: Association for Computational Linguistics, 2018, 1(1): 2009-2019.

[16] DO T H, NGUYEN D M, TSILIGIANNI E, et al. Twitter user geolocation using deep multiview learning[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2018: 6304-6308.

[17] ZHONG T, WANG T, WANG J, et al. Multiple-aspect attentional graph neural networks for online social network user localization[J]. IEEE Access, 2020, 8: 95223-95234.

[18] 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. Stroudsburg: Association for Computational Linguistics, 2012: 1500-1510.

[19] HAN B, COOK P, BALDWIN T. Text-based twitter user geolocation prediction[J]. Journal of Artificial Intelligence Research, 2014, 49(1): 451-500.

[20] KRISHNAMURTHY R. Knowledge enabled location prediction of Twitter users[D]. Dayton: Wright State University, 2015.

[21] CHENG Z, CAVERLEE J, LEE K. You are where you tweet: a content-based approach to geo-locating twitter users[C]//Proceedings of the 19th ACM Conference on Information and Knowledge Management. New York: ACM Press, 2010: 759-768.

[22] ZHENG X, HAN J, SUN A. A survey of location prediction on Twitter[J]. IEEE Transactions on Knowledge & Data Engineering, 2018, 30(9): 1652-1671.

[23] WING B, BALDRIDGE J. Simple supervised document geolocation with geodesic grids[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2011: 955-964.

[24] 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. New York: ACM Press, 2016: 855-864.

[25] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26(2): 3111-3119.

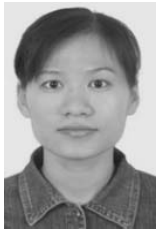
[26] NAIR V, HINTON G E. Rectified linear units improve restricted boltzmann machines vinod nair[C]//Proceedings of the 27th International Conference on Machine Learning. Madison: Omni Press, 2010, 807-814.

[27] AL-RFOU R, ALAIN G, ALMAHAIRI A, et al. Theano: a Python framework for fast computation of mathematical expressions[J]. arXiv e-prints, 1605.02688, 2016: 1-19.

[28] KINGMA D P, BA J. Adam: a method for stochastic optimization[J]. arXiv Preprint, arXiv: 1412.6980, 2014.

[29] 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. Mumbai: The COLING 2012 Organizing Committee, 2012: 1045-1062.

[作者简介]



乔亚琼（1981− ），女，河南开封人，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。

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。

马江涛（1981− ），男，河南开封人，博士，郑州轻工业大学讲师，主要研究方向为数据挖掘和社交网络分析。

李晨亮（1983− ），男，湖北武汉人，博士，武汉大学副教授，主要研究方向为数据挖掘、机器学习、社交网络分析、信息安全、信息检索、文本/网络挖掘和自然语言处理。

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。

李瑞祥（1994− ），男，湖南衡阳人，信息工程大学博士生，主要研究方向为网络实体定位、数据分析和信息安全。