软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(11):2951−2963 [doi: 10.13328/j.cnki.jos.004907]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种大数据环境下的在线社交媒体位置推断方法*

王 凯[1], 余 伟[1], 杨 莎[2], 吴 敏[3], 胡亚慧[4], 李石君[1]

[1](武汉大学 计算机学院,湖北 武汉 430072)

[2](汉口学院 计算机科学与技术学院,湖北 武汉 430212)

[3](中船重工第七二二研究所,湖北 武汉 430079)

[4](空军预警学院,湖北 武汉 430000)

通讯作者: 李石君, E-mail: shjli@whu.edu.cn

摘 要: 随着在线社交媒体的快速发展和可定位设备的大量普及,地理位置作为社交媒体大数据中一种质量极高的信息资源,开始在疾病控制、人口流动性分析和广告精准投放等方面得到广泛应用.但是,由于大量用户没有指定或者不能准确指定位置,社交媒体上的地理位置数据十分稀疏.针对此数据稀疏性问题,提出一种基于用户生成内容的位置推断方法 UGC-LI(user generate content driven location inference method),实现对社交媒体用户和生成文本位置的推断,为基于位置的个性化信息服务提供数据支撑.通过抽取用户生成文本中的本地词语,构建一个基于词汇地理分布差异和用户社交图谱的概率模型,在多层次的地理范围内推断用户位置.同时,提出一个基于位置的参数化语言模型,计算用户生成文本发出的城市.在真实数据集上进行的评估实验表明:UGC-LI方法能够在15km偏移距离准确定位 64.2%的用户,对用户所在城市的推断准确率达到 81.3%;同时,可正确定位 32.7%的用户生成文本发出的城市,与现有方法相比有明显的提高.

关键词: 位置推断;用户生成内容;数据稀疏性;在线社交媒体;社交图谱

中图法分类号: TP311

中文引用格式: 王凯,余伟,杨莎,吴敏,胡亚慧,李石君.一种大数据环境下的在线社交媒体位置推断方法.软件学报, 2015,26(11):2951−2963. http://www.jos.org.cn/1000-9825/4907.htm

英文引用格式: Wang K, Yu W, Yang S, Wu M, Hu YH, Li SJ. Location inference method in online social media with big data. Ruan Jian Xue Bao/Journal of Software, 2015,26(11):2951−2963 (in Chinese). http://www.jos.org.cn/1000-9825/4907.htm

Location Inference Method in Online Social Media with Big Data

WANG Kai[1], YU Wei[1], YANG Sha[2], WU Min[3], HU Ya-Hui[4], LI Shi-Jun[1]

[1](Computer School, Wuhan University, Wuhan 430072, China)

[2](College of Computer Science and Technology, Hankou University, Wuhan 430212, China)

[3](The 722 Research Institute of China Shipbuilding Industry Corporation, Wuhan 430079, China)

[4](Air Force Early Warning Academy, Wuhan 430000, China)

Abstract: As a high-quality source in social media big data, the geographic location has been widely adopted in the fields of disease control, population mobility analysis and ad delivery positioning with the rapid development of online social media and the prevalence of localizable mobile devices. However, the location data are quite sparse because often the locations cannot be accurately specified by the users. To overcome this data sparsity problem, this paper proposes UGC-LI, a user generate content driven location inference method to infer the location where users and social texts are created. The method can provide supporting data for location-based personalized information services. A probability model is constructed by comprehensive considering the distribution of location words and social graph

* 基金项目: 国家自然科学基金(61272109, 61502350); 中央高校基本科研业务费专项资金(2042014kf0057); 湖北省自然科学基金(2014CFB289)

收稿时间: 2015-05-31; 修改时间: 2015-07-14, 2015-08-11; 定稿时间: 2015-08-26

of users via local words extracted from user generated texts to locate the users in multi-granularity. Further, a parameterized linguistic model based on location is presented to calculate the city where the tweet is published. The results of experiment on real-word dataset demonstrate that this new method outperforms existing algorithms. In the experiment, 64.2% of users are identified within 15km displacement distance, 81.3% of the living cities and 32.7% of the cities where the tweets were tweeted are correctly located.

Key words: location inference; user generate content; data sparsity; online social media; social graph

近年来,凭借着在内容生成方式、用户参与的广泛性与即时性、信息扩散模式与速度等方面的优势,社交网络、在线社交媒体和移动互联网发展迅猛.可定位设备,尤其是智能手机的迅速普及,使得地理位置数据逐渐成为一种质量极高的信息资源,开始在疾病传播控制[1,2]、广告精确投放[3,4]、公共安全监控[5]、城市人口流动性分析[6]等方面得到广泛应用.同时,社交媒体本身也出现了“本地地理关注(local focus)”的理念,如谷歌推出的应用“Google Now”(http://www.google.com/landing/now/)和“Google Flu Trend”(http://www.google.org/flutrends/).相对于传统的互联网社交模式,结合了位置数据的在线社交媒体平台将“虚拟现实”这一理念进行深化,提供了一种将知识从虚拟社交平台转换到现实世界的途径,让我们能够更好地理解用户思维和行为模式.

随着用户量的急速增长,社交媒体数据变得愈发庞大且更为复杂,逐渐呈现出典型的大数据 4V 特性.然而,与庞大的数据量形成对比,由于平台限制以及对隐私保护的考虑,大量用户没有指定或者不能准确指定位置,社交媒体中的位置数据普遍十分稀疏[7−10].以中国最大的在线社交媒体平台新浪微博为例,我们抽取了 200 万用户在 7 个月内发表的超过 1 亿条博文,仅有约 1%的博文带有位置标签.

位置数据的稀疏性,使得构建在地理信息上的社交媒体应用服务缺少足够的支撑数据,对基于位置的知识获取造成了极大的挑战.由于在数据生成及传播方式上的特殊性,当前广泛采用的对 IP 地址[11]和地理数据库[12,13]进行模式匹配从而获得位置的方法并不直接适用于在线社交媒体数据环境.例如,随着移动社交平台和网络代理(Web agent)的大量使用,社交媒体中的 IP 地址与地理位置不再直接关联;用户的主观性使得社交媒体生成文本中含有大量的噪声,如随处可见的缩写及网络用语、不同用户对位置的差异甚至针对性的描述、移动环境下位置相对描述的偏移等.此外,当前研究大多针对国外英文社交媒体平台(如 Twitter,Foursquare 等),而国内外社交媒体平台在用户分布、社交形式以及信息表现方式等方面均有很大的不同.例如,与 Twitter 相比,新浪微博的核心用户更加集中于大城市,并强制要求用户指定地级市的注册位置.同时,中英文的固有差异导致了中外社交媒体平台用户表达方式的部分对立,也使得相应的自然语言处理技术不可移植.

本文通过探索词汇、社交关系与地理位置三者之间的关系,提出一种基于用户生成内容的位置推断方法 UGC-LI(user generate content driven location inference method),实现中文社交媒体环境下的用户与生成文本的位置推断,为基于位置的个性化应用提供数据支撑.在线社交媒体文本内容中,本文关注那些能够指示位置的词语.当前研究结果表明:在特定地位范围内频繁出现的词语与位置存在联系[7−9,14],本文中将这种词语称为本地词语(local words).其中,最典型的本地词语即为直接指示位置的地点或地标名词,如城市、街道、地标建筑等.另一种词语则潜在地描述了特定位置,如词语“那达慕大会”常在内蒙古自治区内出现,“过早”则是湖北地区常见的方言词汇.图 1 展示了本文数据集中地名“重庆”在全国各省市出现的频率.由图 1 中可以看出,以重庆市为中心,随着距离的增加,该词语的出现频率越来越小.此外,用户间的实际距离与其网络节点的虚拟距离之间存在正相关关系,即,随着地理距离的增加,社交媒体用户间形成朋友关系的概率也随之降低[7,14].基于此,本文将使用用户社交图谱辅助位置推断.

总体来说,本文的主要贡献有以下几点:

1) 研究了中文社交媒体环境下,用户生成文本词语的位置指向性.与现有研究相比,本文使用社交媒体中海量、细粒度的文本附加坐标进行词语位置关联研究.实验结果表明,在特定位置频繁出现,且随着与该位置距离的增加出现概率下降较快的词语,对位置的指向性是存在的.
2) 提出了一个混合概率模型推断社交媒体用户位置.与已有研究相比,该模型综合考虑了词语的位置指向性以及虚拟-物理节点邻近关系.实验结果表明,本文提出的方法能够有效地提高用户位置推断正确率和精准度.

3) 构建了一个基于位置的参数化语言模型,进行生成文本位置推断.与已有研究相比,该模型基于词语自身的位置属性进行文本定位,降低了常用及偏僻词语对生成文本地理散布的影响.实验结果表明,本文方法能够正确找到32.7%的用户生成文本发出的城市,与现有方法相比有了明显的提高.

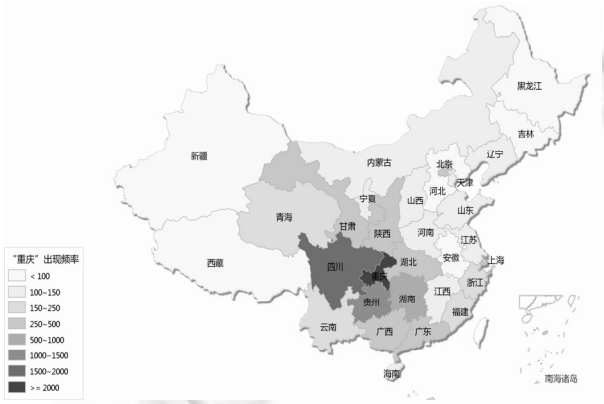


Fig.1 Occurrence frequency of the word "Chongqing" in different areas across the country

图 1 "重庆"在全国不同区域的出现频率

本文第1节介绍国内外研究现状,从位置推断理论方面阐述相关工作与UGC-LI方法的异同.第2节简要分析在线社交媒体中位置的稀疏性问题,并对问题做形式化定义.第3节介绍UGC-LI方法的推断理论和过程.第4节描述实验过程并对实验结果进行对比分析.最后,对本文工作做出总结和展望.

1 相关工作

当前,在基于社交媒体平台与社交网络服务的研究中,针对用户和生成文本的位置推断已经成为一个研究热点,本节分别从这两个方面介绍国内外的相关研究,并与UGC-LI方法进行理论对比.

在用户位置推断方面,一些研究通过分析用户历史轨迹数据,借助外部地理数据库推测社交媒体用户位置.如,Hecht等人[12]通过分析Twitter用户行为模式,使用基于词频的多项式贝叶斯模型探测用户位置;Eisenstein等人[13]通过语言一致性建模来预测作者和文本的位置.与 UGC-LI 方法完全基于用户生成内容不同,这些方法大多需要高精度的外部地理数据库支持,不同的地理数据库会导致位置的偏移.与之相比,基于词汇的地理特性从用户生成内容角度探索位置推断问题,是当前较常用的方法[15,16].最近,Backstorm 等人[17]提出一个语言生成模型,用于探测一条查询的位置,Cheng 等人[8]将其方法扩展到了社交媒体用户位置推断研究上.他们通过识别文本中的地标名词对词汇的地理属性建模,并基于社交关系等特征设计平滑算法修正定位误差,这与本文提出的方法类似.但是,Cheng 等人的方法需构建分类器来识别地标名词,在模型训练前要进行大量的人工标注工作.同时,该方法将各城市的中心点坐标作为参照点来计算词与位置的关系,这使得参照点过少且与实际位置间的偏差较严重,定位偏移距离较高.实验结果显示:Cheng 等人的方法能够正确推断出51%用户的位置,平均偏移距离则超过了100英里.Ryoo 等人[9]则进一步改进了 Cheng 的方法,使用 tweet 附加的 GPS 坐标点代替原来有限的城市坐标,在10km的误差范围内精确定位了超过50%的用户.但是该方法没有充分考虑社交关系对用户地理聚集的影响,计算过程完全依赖于 tweet 文本,当文本中的位置词汇较稀疏时(第 4.1 节中的数据集属性统计显示,这种情况在中文社交媒体平台中并不罕见),该算法的推断准确率会大幅度下降.Ren[18]和 Backstrom[19]等人的工作考虑了用户社交关系对位置推断的影响,并探索了在线虚拟朋友关系与实际地理距离之间的联系.但是,其测试数据量普遍较小(不到 1 000 位用户),缺乏对大型社交网络数据集的研究.除此之外,Chandra[20],Jurgens[21],Li[22],McGee[23]等人均从社交图谱着手对用户位置推断问题进行了探索.

在用户生成文本定位方面,Li 等人[24]尝试集成表达与位置以预测一条 tweet 所属的 POI(point of interest,兴

趣点).他们为每一条 tweet 选取了候选 POI,使用从 Web 上抽取的信息为每一个 POI 建立语言模型,计算查询 tweet 的 KL 差异并排序.但是,当属于一个 POI 的 tweet 数目发生变化时,该方法的推断准确率会发生剧烈的浮动. Kinsella 等人[10]利用 Twitter 中的 GPS 标签建立了一个位置计算框架,通过位置语言模型生成一条 tweet 的概率对地点进行排序,这与我们的工作类似.Kotzias 等人[7]基于 Kinsella 等人[10]的方法,探测处于同一个城市内的用户并对 tweet 进行坐标定位.但是,当前的研究大多未考虑词汇自身的位置属性对文本定位的影响,进行位置语言建模时,对所有词语赋予了统一的权限.而 UGC-LI 方法则在建模过程中基于词与位置的关联强度对词语分配权重,从而降低非本地词语对总体生成概率的影响,提高了定位准确率.

在国内学者中,赵荣娇等人[25]提出了使用替代可定位用户实现源用户定位的方法,郭迟等人[26]也提出了针对位置大数据的分析方法.总的来说,相对于位置服务方面研究成果的大量涌现,国内对于位置推断的研究比较匮乏,尚处于起步阶段.在本文中,UGC-LI 方法使用海量且细粒度的文本附加坐标计算词语的焦点位置及地理散布,基于一个改进的概率模型计算词与位置的关联强度,在中文社交媒体环境下验证社交邻近与地理邻近间的关系.与上述方法相比,UGC-LI 方法的特点主要体现在:(i) 能够对本地词语进行自动识别,不需要进行手工标注; (ii) 使用关联度阈值筛选本地词语,结合用户社交图谱构建混合概率,提升用户位置推断效果;(iii) 基于词语的本地性计算词权,从而有效降低本地词语稀疏性造成的生成文本定位误差.

2 初步研究

本文的研究目标是:(i) 定位社交媒体用户,即,推断用户所在城市及更细粒度的城区、县等位置级别信息; (ii) 探测单条用户生成文本的发出城市,在本文的研究场景下,即推断一条博文的发出城市.本节中,我们对社交媒体中位置数据的稀疏性进行初步研究,然后对研究问题和推断方法做形式化定义和描述.

2.1 在线社交媒体位置数据的稀疏性

为了研究中文社交媒体中位置数据的分布,我们以中国最大的在线社交媒体平台——新浪微博为数据源.采用雪球抽样方法,从 5 个约有 1 000 粉丝的初始用户开始,基于广度优先搜索方法,按照用户间的关注关系抽取中国大陆用户的个人信息和历史博文.该数据采集策略能够保证得到新浪微博完整社交图谱的一个子图,覆盖了包括非活跃用户在内的全部微博用户类型,从而最大程度地降低数据的采集偏置性.从 2014 年 1 月~7 月,我们共抽取了 2 010 331 条用户个人信息、113 972 086 条博文以及 70 468 437 条关注——粉丝记录.如表 1 所示,对数据集的统计显示,约 69%的用户填写了地级市位置.目前,新浪微博规定的用户当前所在地最详细即为地级市级别.而全部博文中仅有 1.14%,即 1 299 281 条博文包含地理位置标签.显然,新浪微博中的位置数据非常稀疏且粒度过粗.

Table 1 Statistics of users' location information

表 1 用户位置信息统计

位置级别	统计值	
	占比(%)	数量
地级市	68.8	1 383 107
省	20.8	420 159
无	10.3	207 065
–	150	总 2 010 331

2.2 问题定义

针对在线社交媒体中的地理位置稀疏性问题,我们的研究目标是推断用户细粒度的位置及单条生成文本发出的城市.为了给出问题的形式化定义,我们首先对用户及词汇本身的地理属性进行一些预定义.以 $T=\{t_1, t_2,\ldots\}$表示全部博文集,$T_{Location}\subseteq T$ 表示 T 中附加了 GPS 标签的博文子集,则有:

定义 1(博文词语 w). 与 Backstorm[17]针对 Twitter 英语单词的观点类似,本文假设每个博文词语都有一个焦点位置,与该焦点位置距离越远,该词的出现概率就越低.博文词语 w 可用一个三元组表示:$w=\{te,p_w,\alpha_w\}$.其中,te

为 w 的文本表示;p_w 为 w 的焦点位置的坐标;α_w 表示 w 的“地理散布程度”,即,w 远离 p_w 时,其出现频率的下降程度.在本文中,α_w 的值即为 w 与 p_w 之间的关联强度.全部博文词语构成词表 W.

定义 2(本地词典 W_{GL}). 当词的α值很小时,该词对位置的指向作用已非常低.因此,本文使用 W_{GL} 表示本地词典,即,W 中α值最大的 K 个词语的集合.可知,$W_{GL}\subseteq W$ 且$|W_{GL}|=K$.

定义 3(社交媒体用户 u). u 可表示为一个三元组:$u=\{pro,T_u,S_u\}$.其中,pro 表示用户的个人属性(如用户名、性别、当前居住地等用户公开的个人信息);$T_u\subseteq T$ 表示 u 发表过的历史博文;$S_u=\{U_A,U_O\}$代表了 u 的社交关系网,其中,U_A 表示关注 u 的用户集(即粉丝),U_O 表示 u 关注的用户(即关注).全部用户构成社交媒体用户集 U.

根据上述定义,本文的研究问题可以形式化为:

1) 用户位置推断:对于用户 u,给定其历史博文 T_u、社交图谱 S_u 及城市列表 L,计算 u 所在位置坐标点.在特定情况下,计算 u 位于城市 $l(l\in L)$的概率似然值 $S_{likehood}(l|u)$,选择使 $S_{likehood}(l|u)$最大的城市 l_u 作为 u 的所在城市.
2) 生成文本位置推断:对于一条生成文本 t,给定 W_{GL} 及城市列表 L,计算 t 由城市 $l(l\in L)$发出的概率似然值 $S_{likehood}(l|t)$,选择使 $S_{likehood}(l|t)$最大的城市 l_t 作为发出 t 的城市.

3 UGC-LI

3.1 本地词语度量与识别

本研究面对的第 1 个问题是如何从词表 W 中准确识别本地词语.即,给定词 w,我们首先需要判断 $w\in W_{GL}$ 是否成立.这要求我们给出一种计算方法,量化词语与对应位置的关联强度,也就是计算词语的本地强度.在线社交媒体平台对文本长度的限制以及用户的主观性,使得生成文本中含有大量的噪声信息.如微博中随处可见的网络用语、用户对位置的个性化描述、衍生程序生成的新位置等,这种高噪声的数据环境使得传统的地理数据库检索方法不再适用.本节中,我们基于一个改进的概率模型计算计算词语的本地强度.

根据定义,对于任意一个博文词语 w,它的两个参数 p_w 和α_w 分别表示 w 的焦点位置和“地理散布程度”.位置点距离 p_w 越近,w 的出现频率越高;当 w 远离 p_w 时,其出现频率下降越快、α_w 值越大.对本文研究问题来说,α_w 是决定 w 是否为本地词语的关键.因此,结合 Cheng[8]和 Ryoo[9]的方法,我们将社交媒体中海量且细粒度的博文附加坐标作为参照点,计算本地词语的焦点位置及地理散布.即,对于词 w,令 T_w 表示 $T_{Location}$ 中包含 w 的博文集,d_{tp} 表示博文 $t(t\in T_w)$附加的 GPS 坐标点(参照点)与 p_w 之间的距离,则针对 p_w 的对数化最大似然推理函数 f 可以定义为

$$f(C_w,\alpha_w)=\sum_{t\in T_w}\log(C_w\times d_{tw}^{-\alpha_w})+\sum_{t\notin T_w}\log(1-C_w\times d_{tw}^{-\alpha_w}) \tag{1}$$

上式中,C_w 和α_w 是确定 w 的焦点位置和本地强度的关键参数.其中,C_w 是一个常数,表示 w 在 p_w 出现的先验概率;指数α_w 则代表了 w 与 p_w 的联系强度,α_w 值越大,w 的地理聚集性就越高,本地性也就越强. $C_w\times d_{tw}^{-\alpha_w}$ 则代表了 w 由处于位置 t 的用户发出的概率.Backstorm[17]的研究已经表明,$f(C,\alpha)$在其定义域内仅有一个局部极大值.于是,给定焦点位置,可以选择 C 和α来使 $f(C,\alpha)$的值最大,同时也找到最优的 C 和α.

3.2 用户位置推断

得到词语的 C 和α值后,我们可以提出一个直观的用户位置推测方法:(i) 计算词表 W 中每个词的α值,挑选α值最大的 K 个词语形成 W_{GL};(ii) 对用户 u,统计 T_u 中本地词语的分布情况,计算这些词语焦点位置确定的地理几何中心作为 u 的位置.但是相对于少量的本地词语,在线社交媒体中非本地词语的数量要大很多[9,16,17],当 T_u 中词语的α值普遍偏小时,该直观方法的推断位置会向参照点严重偏移.此时,为了保证定位误差可控,我们基于用户社交图谱 S_u 计算 u 处于特定城市的概率.

现实社区中的成员仍然会以较高的聚集系数在社交网络上形成聚集,实际位置临近的用户会在社交媒体上对“本地事件”表现出相同的兴趣及偏好,也会加快其“本地朋友”与“非本地朋友”间新的社交链接的形成速

度[19,27].从另一方面看,根据 McGee 等人[28]的调查,在线社交媒体中,用户超过 40%的虚拟朋友居住在其 100 英里范围内,虚拟社交关系往往意味着实际位置的邻近,用户的虚拟社交邻居很可能是其"本地朋友".基于这种潜在关系,我们通过用户的社交图谱来改进上述直观的位置推测方法.即,给定用户 u、博文集 T_u、社交图谱 S_u、全国城市列表 L 及本地词典 W_{GL},计算 u 位于城市 $l_i(l_i \in L)$的似然值 $S_{likehood}(l_i|u)$.我们提出 3 种方法来计算用户处于特定城市的概率:

- 基于 T_u 中的本地词语

给定词语 $w(w \in W_{GL})$,则 w 由位于城市 l_i 的用户发出的近似概率为

$$S_{likehood}(l_i \mid w) = C_w \times d_{l_i w}^{-\alpha_w} \tag{2}$$

其中,C_w 和 α_w 为 w 的最优 C 和 α 值,$d_{l_i w}$ 为词 w 的焦点位置与城市 l_i 之间的距离.使用 T_u 中的本地词语由城市 l_i 发出的概率来表示 u 位于 l_i 的似然值,即

$$S_{likehood}(l_i \mid u_{GLocalWords}) = \prod_{w \in T_u} S_{likehood}(l_i \mid w) \tag{3}$$

- 基于 S_u

基于用户社交图谱计算 u 处于特定城市的概率时,统计 u 的直接社交邻居节点群(即 u 的关注和粉丝)中位于 l_i 的节点所占比例是最直观的办法.但是现有研究结果表明,与直接社交邻居相比,社交媒体中经历过两个关系层跳跃的节点对(2-hop)之间更易形成一种相关性较高且坚固的连接关系,如"朋友的朋友"关系[18,29].

Kwak 等人[30]通过对一个随机抽取的包含 4 千万条用户个人信息、10 亿条社交关系以及超过 1 亿条博文的 Twitter 数据集进行统计研究发现:用户社交邻居越少,用户与其邻居节点、用户邻居节点之间的社交关系越紧密,实际地理间距越短.此外,Ren 等人[18]的研究结果也表明,使用 2-hop 社交关系进行节点关联位置预测,比基于直接社交邻居方法的推断精确度高 6%~10%.本节中,我们综合考虑用户与其直接社交邻居以及 2-hop 社交关系节点间的地理邻近关系,则对于用户 u,若 u 关注了 u',则 u 位于 l_i 的概率似然值为

$$S_{likehood}(l_i \mid u_{socialNetwork}) = \frac{Count(U_O, l_i) + Count(U_A, l_i) + Count(U'_A, l_i)}{|U_A| + |U_O| + \sum |U'_A|} \tag{4}$$

其中,$|U_A|$和$|U_O|$分别表示 u 的粉丝集 U_A 和关注集 U_O 的大小,$|U'_A|$表示 u'的关注集 U'_A 的大小,$Count(U_A,l_i)$和 $Count(U_O,l_i)$分别代表了 U_A 和 U_O 中位于 l_i 的用户数量,$Count(U'_A,l_i)$ 则表示 U'_A 中位于 l_i 的用户数量.

- 混合模型

通过一个调节参数对以上两种方法进行集成,即

$$S_{likehood}(l_i|u) = \beta \cdot S_{likehood}(l_i|u_{GLocalWords}) + (1-\beta) S_{likehood}(l_i|u_{socialNetwork}) \tag{5}$$

其中,$\beta \in [0,1]$为调节参数,其最优值可通过实验获得.

根据上述定义,可以提出 UGC-LI 用户位置推断算法,如算法 1 所示.我们使用一个阈值 δ 控制定位方法的选择,当 T_u 中本地词语的平均 α 值 $Avg_u(\alpha)$低于 δ 时,基于公式(5)综合考虑词语的位置指向性以及用户的社交图谱进行用户定位.算法 1 中,$GeoCenter(W_{GL\text{-}u})$函数计算 $W_{GL\text{-}u}$ 中所有词语的地理几何中心,$Distribution(S_u,l_i)$函数基于公式(4)计算 u 的直接社交邻居及 2-hop 社交邻居节点位于 l_i 的概率,$Mix(S_{likehood}(l_i|u_{GLocalWords}),S_{likehood}(l_i|u_{socialNetwork}))$函数基于公式(5)计算 u 位于 l_i 的概率似然值.最后,$sort(S_{likehood}(l|u),L)$函数根据 $S_{likehood}(l|u)$值对城市进行排序,计算 $\arg\max_{l_i \in L}\{S_{likehood}(l_i \mid u)\}$ 作为推断结果城市.

算法 1. UGC-LI——用户位置推断.

输入:

- T_u:用户 u 发表的历史博文集;
- S_u:用户 u 的社交关系图谱;
- W_{GL}:本地词典;
- L:全国城市(市、区、县)列表;
- δ:α 阈值.

输出:u 的坐标 $Coord(u)$、u 所属的城市 l_u.

1. $W_{GL\text{-}u} \leftarrow T_u \cap W_{GL}$
2. $Avg_u(\alpha) = \sum_{w \in W_{GL-u}} \alpha_w \Big/ |W_{GL-u}|$
3. **if** $Avg_u(\alpha) > \delta$ **then**
4. $Coord(u) = GeoCenter(W_{GL\text{-}u})$
5. **else**
6. **for each** L 中的城市 l_i **do**
7. $S_{likehood}(l_i|u_{GLocalWords}) \leftarrow 0$
8. **for each** $W_{GL\text{-}u}$ 中的词语 w **do**
9. $S_{likehood}(l_i|u_{GLocalWords}) += S_{likehood}(l_i|w)$
10. **end for**
11. $S_{likehood}(l_i|u_{socialNetwork}) = Distribution(S_u, l_i)$
12. $S_{likehood}(l_i|u) = Mix(S_{likehood}(l_i|u_{GLocalWords}), S_{likehood}(l_i|u_{socialNetwork}))$
13. **end for**
14. $l_u = sort(S_{likehood}(l|u), L)$
15. **end if**

3.3 生成文本位置推断

相对于 T 庞大的文本量,本地词语在博文集中的分布是稀疏的.因在线社交媒体平台对单条博文长度的限制(新浪微博限制在140个字符以内),仅通过一条博文中本地词语的位置分布来实现博文定位,远比用户位置推断困难得多.因此,我们使用语言建模方法对城市建立基于位置的参数化概率语言模型,预测博文发出的城市.

目前,已有的研究大多将词语的出现频率作为其位置权重,如位置相关 TF-IDF 算法[31].但是,这种方法没有考虑词语本身的位置属性,当词语的地理分布比较均匀,即,词语在若干个地理区域内均频繁出现时,该方法的推断准确率下降非常迅速.本文中,我们使用 w 由位于 t 位置的用户发出的概率来确定词语的权重.同时,为了强调词语与位置联系强度的差别,基于一个加权贝叶斯演化模型来预测单条博文的位置.

对于城市 l,若有 $w_i \in W_{GL}$,则使用 $S_{likehood}(l|w_i)$ 表示词语 w_i 相对于 l 的权重.若有 $w_j \notin W_{GL}$,则该词语地域分布较为松散.为了防止下溢且方便计算,使用 W_{GL} 中所有本地词语的最小 $S_{likehood}(l|w)$ 值来标识 w_j 相对于 l 的权重,则词语 w 相对于城市 l 的权重可表示为

$$Q_{w,l} = \begin{cases} S_{likehood}(l \mid w), & w \in W_{GL} \\ \min\limits_{w' \in W_{GL}} \{S_{likehood}(l \mid w')\}, & w \notin W_{GL} \end{cases} \tag{6}$$

对于一条博文 t,计算其由城市 l_i 发出的概率:

$$L(t) = \arg\max_{l_i \in L} p(l_i \mid \hat{\theta}_L) \prod_{w_m \in t} p(w_m \mid l_i; \hat{\theta}_L) \times Q_{m,i} \tag{7}$$

其中,$\hat{\theta}_L$ 即为基于位置的参数化模型,$l_i \in L$ 为城市列表中的城市,w_m 为博文 t 包含的词语,$Q_{m,i}$ 表示 w_m 相对于 l_i 的权重.

在线社交媒体位置数据的稀疏性,使得数据集中含有大量的“微数据(tiny data)”,例如,一些总体出现次数较少且仅在少数区域出现的词语,当某些区域人口较少时,这种现象更为严重[8].因此,为了防止出现零概率,使用 Dirchlet 方法[32]对词 w_m 的出现概率进行平滑:

$$p(w_m \mid l_i; \hat{\theta}_L) = \frac{\mu P(w_m \mid \theta_C) + Count(w_m, l_i)}{|W| + \mu} \tag{8}$$

其中,$\mu \in (0,1)$为平滑参数,可在实验中确定;$Count(w_m, l_i)$表示词 w_m 在 l_i 出现的次数;θ_C 表示 w_m 在 L 中所有位置的分布;$|W|$为词表的长度.

4 实验与分析

4.1 数据集与数据处理

目前,新浪微博对用户位置最细粒度的描述精确到了地级市,这对于本文的研究来说是不够的.为了构成本文实验的 ground-truth 数据,我们需要获得用户更细粒度的位置信息,在本文中,即为用户位置对应的坐标点.调查显示,新浪微博的核心用户是白领和学生(http://www.cnnic.net.cn/),所在地与工作或教育地点一致(工作地点优先)的用户位置可信度更高.因此,我们抽取了用户最新的工作信息与教育背景,使用高德地图(http://www.amap.com/)搜索与地理编码 API 获得工作和教育地点对应的详细坐标,作为用户的实际位置.提取实际位置在用户填写所在地范围内的用户群(约占全部用户量的 9.4%),构成本文实验的 ground-truth 数据集.同时,为了去除噪声,我们参照 Jurgens 等人[32]的做法,选择在其声明所在地 20km 半径范围内至少发表了 3 条附加 GPS 标签的博文,且总博文数不少于 10 条的用户群,从而过滤掉活动位置不确定的用户(如经常到处旅行的用户).最终的数据集由 136 945 条用户个人信息及 2 472 436 条博文组成,其中,724 540 条博文带有 GPS 标签.

对于得到的博文文本,使用语言技术平台(language technology platform,简称 LTP)[33]完成分词和命名实体识别操作.为了适应在线社交媒体文本数据的主观表达多、文本长度短等特性,我们定制了原有的 LTP 代码,在 NLPIR 微博语料库(http://www.nlpir.org/)上进行独立训练,从而更好地满足短文本分词和 NER 需求.从文本中筛选掉特殊符号和停用词之后,我们从 724 540 条带有 GPS 标签的博文中抽取了 255 203 个不同的词语.表 2 列出了对数据集各个属性的统计结果.

Table 2 Attributes of the data set

表 2 数据集属性

属性	值
带有 GPS 标签的博文数(%)	28.55
用户平均博文数(条)	17.3
单个用户平均带有 GPS 标签博文数(条)	4.95
词语平均长度(字符个数)	2.6±0.3
单条微博中的平均词语个数	8.5±1.4

4.2 位置计算方法

计算词语的焦点位置时,受地球曲率的影响,直接计算两个坐标点间的欧几里得距离会导致一定的位置偏移.因此,本文使用地图投影算法计算候选坐标点的间距.此外,我们使用 L_1 多元几何中心点[34]计算多坐标点的地理几何中心.即,对于候选坐标点集 Co,目标位置 m 为

$$m=\arg\min_{x\in Co}\sum_{y\in Co}dis(x,y) \tag{9}$$

具体来说,结合 Cheng 等人[8]的方法,对于词 w,计算 C_w 和α_w 的方法为:(1) 首先,将中国大陆地图划分为 10km×10km 的网格;(2) 对于任意一个网格 g_1,找到所有出现在 g_1 内的包含词 w 的博文,将这些博文附带的 GPS 坐标点确定的 L_1 多元几何中心点作为 w 的焦点位置,基于公式(1)计算此时的 f 值;(3) 合并所有与 g_1 直接临界的网格形成一个覆盖范围更大的网格 g_2,按照步骤(2)中的方法计算 f 值;(4) 重复步骤(3),直至全部地图只剩余一个网格;(5) 对比每一次扩大网格后生成的 f 值,选择最大 f 值对应的焦点位置、C 值和α值,即为 p_w,C_w 和α_w.

4.3 评估方法

为了衡量方法的推断正确率和推断精度,本文使用推断结果与真实位置间的平均间距(平均偏移距离)以及正确推断的用户及博文在实值数据集中的占比(推断准确率)作为评估指标,使用 5 重交叉验证方法对 UGC-LI 及所有对比方法进行实验验证.实验中,以 500m 误差距离作为可允许误差,以保证算法的可伸缩性.为了对 UGC-LI 方法实验结果有直观的认识,我们将 UGC-LI 与 Cheng,Kinsella 等人提出的方法进行对比.这些对比方法是目前广泛使用且运行效果最好的位置推断方法.同时,根据推断条件提出一些基本方法作为 Baseline,验证本文观点和方法的正确性和有效性.用户和博文定位的对比方法分别见表 3 和表 4.

Table 3 Comparison methods of locating user u

表 3 对比方法:定位用户 u

方法	描述
Baseline#1	计算 $W_{GL\text{-}u}$ 中所有词语焦点位置的几何中心,作为 u 的位置
Baseline#2	不考虑 S_u,基于公式(3),通过 $W_{GL\text{-}u}$ 中词语的位置属性推断 u 的位置
Cheng's method	基于文献[8],使用城市中心点坐标为参照点,计算词的焦点位置和地理离散程度,结合用户社交网络定位 u 所属的城市
Ryoo's method	基于文献[9],不考虑 S_u,计算词语与博文 GPS 标签间距来确定词的焦点位置和本地化程度

Table 4 Comparison methods of locating tweet t

表 4 对比方法:定位博文 t

方法	描述
Baseline#1	直接将 t 中各词语的中心点作为 t 的位置
Baseline#2	不考虑词语相对于位置的权重,计算位置语言模型生成博文的概率
Kinsella's method	基于文献[10],不考虑词语相对于位置的权重,使用 Kullback-Leibler 差异修正语言模型推断位置的地理偏移

4.4 实验结果与分析

实验的目的主要有 3 个:(i) 验证本地词语对位置的指向性;(ii) 验证基于本地性强度阈值和社交关系的方法能否改进用户定位效果;(iii) 验证基于词语位置属性的加权生成模型能否提高博文位置推断准确率.

4.4.1 本地词语分布

在进行本地词语识别时,为了防止数据集中 tiny data 造成的"伪本地词语"(如出现次数极少、范围集中但α值很大的词语),参照 Chneg 等人的方法[8],我们选择出现次数大于 50 次的词语,共有 18 747 个.本文数据集中,词语α值的取值范围为(0.01,1.43),约 80%的词语α值介于 0.2~0.6 之间.当词的α值小于 0.4 时,相邻两个词语之间α值差别已经很小.实验中,设置α阈值为 0.62,选择 1 500 个α值最高的词语构成 W_{GL}.经过人工识别,其中超过 1 000 个词语为建筑、城市名称或其他专有名词,能够明确地指示位置.

表 5 列出了数据集中出现次数最多的 10 个词语的焦点位置及α值,其中,对位置的描述通过高德地图(http://www.amap.com/)获得.由表 5 可知,此 10 个词语的焦点位置大部分集中在中部地区.其中,仅有一个地名词语"北京",该词共出现了 22 193 次,也是表中唯一α>0.3 的词语,其焦点位置指向河南省洛阳市,在物理位置上与北京市相距约 670km.将所有词语按α值降序排序,α值最高的 10 个词语位置见表 6.由表中可以看出,此 10 个词语均为城市、建筑或者机构名称,各自的焦点位置也均在相对应的地理区域内.如词语"国安"为机构名,其焦点位置距离北京国安足球队的主场工人体育场约 13km.

由此可见,本地词语对位置的指向性是存在的.

Table 5 Geographic distribution of the most used words

表 5 最频繁出现词语的地理分布

词语	α	经度	纬度	位置描述
一个人	0.127	33.43	116.41	安徽省亳州市
今天	0.094	31.11	119.97	湖北省宜昌市
大家	0.215	29.97	116.51	江西省九江市
我们	0.136	39.16	116.88	山东省泰安市
自己	0.113	31.77	119.50	江苏省常州市
感觉	0.242	29.27	107.20	重庆市酉阳土家族苗族自治县
开始	0.212	33.761	109.578	陕西省商洛市
第一	0.106	32.21	114.43	河南省信阳市
北京	0.377	34.91	111.88	河南省洛阳市
世界杯	0.188	26.23	114.00	湖南省郴州市

Table 6 Geographic distribution of the words with high α

表 6 α值最高的词语地理分布

词语	α	经度	纬度	位置
虹口	1.424	31.15	121.38	上海市闵行区
国安	1.420	39.91	116.30	北京市海淀区
滨海	1.273	39.047	117.35	天津市东丽区
华侨城	1.205	22.54	114.02	深圳市福田区
华南	1.114	23.32	113.28	广州市白云区
汉口	1.113	30.52	114.31	武汉市武昌区
复旦	1.102	31.27	121.50	上海市杨浦区
金陵	1.008	31.85	118.61	南京市江宁区
浦东	1.008	31.26	121.38	上海市普陀区
北京西站	1.007	39.84	116.31	北京市丰台区

4.4.2 用户位置推断

由算法 1 可知,UGC-LI 用户位置推断方法的时间复杂度为 $o(|L|\times|W_{GL}|)$.本文实验中,$|L|$为中国大陆最详细到县的城市数目,共有 6 275 个;$|W_{GL}|$为本地词库的大小,实验中取$|W_{GL}|$=1500.根据表 3 列出的对比评估方法,用户定位评估结果见表 7.由于单个用户平均带有 GPS 标签的博文不到 5 条,位置数据的稀疏性导致 Baseline#1 方法的平均偏移距离较大,超过了 120km.此外,Cheng 的方法无论是推断精度还是准确率,都无法与 Baseline#2, Ryo's Method 以及 UCC-LI 方法相比,这与前文的结论是一致的.当用户博文集中所有词语的平均α值 $Avg(\alpha)>\delta$(实验中设置δ=0.75)时,Baseline#2 与 Ryo's Method 推断思想基本类似;当所有词语的平均α值 $Avg(\alpha)\leqslant\delta$ 时,Baseline#2 通过计算词语由特定城市发出的概率来推测用户所属的城市.这使得在 15km 偏移距离内,Baseline#2 的定位准确率比 Ryoo 低约 20%,但在推测用户所属城市时(一般在 30km~45km 偏移距离内),Baseline#2 的推断效果与 Ryoo 的方法已经差别不大.此外,与其在英文数据集上的实验结果相比,Cheng 和 Ryoo 的方法在本文数据集上进行位置推断的平均偏移距离更小.这从另外一方面说明了中文社交媒体用户城市聚集度较高的现象.

Table 7 Result comparison of user location inference experiment

表 7 用户位置推断实验结果对比

方法	实验结果		
	平均偏移距离(km)	15km 偏移距离内(%)	所属城市(%)
Baseline#1	127.9	2.7	12.2
Baseline#2	53.5	42.8	68.4
UGC-LI	**29.6**	**64.2**	**81.3**
Cheng's method	110.4	29.8	57.6
Ryoo's method	37.0	61.5	72.1

图 2 展示了各对比方法的推断准确率随偏移距离变化的趋势图.

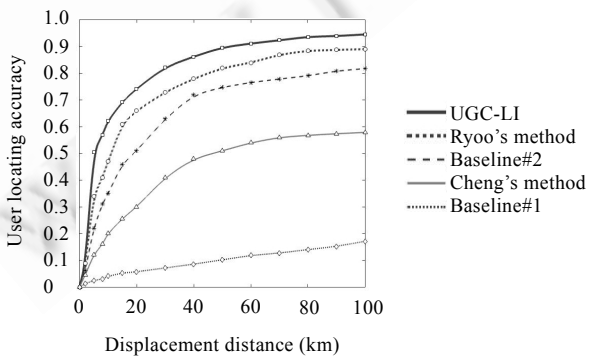


Fig.2 Experimental performance of user location inference methods

图 2 各用户位置推断方法实验性能

可以看到,UGC-LI 的平均偏移距离为 29.6km,在 15km 误差内能够正确定位 64.2%的用户,并能正确找到 81.3%的用户所属的城市.当允许误差在 70km 范围内时,UGC-LI 能够准确定位 90%以上的用户.无论从推断准确率还是推断精度来看,UGC-LI 都已超过了 Ryoo 和 Cheng 的方法.

由此可见,基于本地性强度阈值和社交关系的方法能够有效改进用户位置推断效果.

4.4.3 生成文本位置推断

数据集中带有位置标签的博文共有 724 540 条,约占全部博文的 29%.与用户位置推断不同,由于社交媒体平台对文本长度的限制,生成文本位置推断的偏移距离普遍较大.目前,Kinsella 等人[10]方法的推断准确率最高,能够正确识别 32%的博文发出的城市,但其平均偏移距离超过了 50km.考虑到中国城市的平均面积,我们的实验仍然在 30km~45km 允许偏移距离内推断博文所属的城市.由于用户位置推断方法中已进行了 $S_{likehood}(l_i|w)$的计算,UGC-LI 生成文本位置推断算法仅需对 L 与 W_{GL} 进行一遍交叉扫描,时间复杂度为 $o(|L|\times|W_{GL}|)$.表 8 列出了生成文本位置推断的实验结果.

Table 8 Result comparison of user generated text location inference experiment

表 8 生成文本位置推断结果对比

方法	所属城市(%)
Baseline#1	10.2
Baseline#2	21.4
UGC-LI	**32.7**
Kinsella's method	26.5

受限于文本长度,当博文中属于 W_{GL} 词语过少时,生成文本位置的推断准确率非常低,这导致 Baseline#1 方法的推断准确率仅有 10%.Baseline#2 方法基于位置语言模型计算博文从特定城市发出的概率,推断准确率提高 1 倍,达到 21%.Kinsella 的方法同样没有考虑词语相对于位置的权重,但其利用 KL 距离修正了定位误差,推断准确率达到26.5%.在利用词语本身的位置属性计算词语权重,并结合位置语言生成模型以后,UGC-LI 方法能够正确定位 32.7%的博文所属的城市.与其他方法相比,这一方法在目前来说推断效果是最好的..

由此可见,基于词语本身位置属性的位置语言概率模型能够有效改进生成文本位置推断效果.

5 结 论

得益于可定位设备的大量使用,地理位置数据已成为一种高质量的信息资源.然而,由于用户没有指定或不能准确指定位置,这种高质量的数据十分稀疏.本文提出一种基于用户生成内容的位置推断方法 UGC-LI,实现中文社交媒体环境下对用户与生成文本的位置推断.本文中,我们根据本地词语的分布差异,结合社交关系计算用户不同粒度的位置.同时,基于本地词语建立位置语言概率生成模型,探测发出博文的城市.在采集自新浪微博的真实数据集上,实验结果表明,UGC-LI 方法能够在 15km 偏移距离准确定位 64.2%的用户,对用户所在城市的推断准确率达到 81.3%;同时,可正确定位 32.7%的用户生成文本发出的城市.

目前,我们的研究针对中文在线社交媒体,但是本文提出的方法完全基于用户生成内容,这在理论上适用于任何 Web 社交平台.除了相应的自然语言处理技术以外,本文提出的方法也能应用在其他不同语言的在线社交媒体下,如 Twitter 平台.此外,本文研究基于用户的历史生成内容,消除了用户所在位置的动态变化带来的位置推断影响.我们将在下一步工作中加入时间维度,对用户的实时动态位置推断问题做进一步研究.

References:

[1] LaBute M, McMahon BH, Brown M, Manore C, Fair JM. A flexible spatial framework for modeling spread of pathogens in animals with biosurveillance and disease control applications. ISPRS Int'l Journal of Geo-Information, 2014,3(2):638−661. [doi: 10.3390/ijgi3020638]

[2] Lan L, Malbasa V, Vucetic S. Spatial scan for disease mapping on a mobile population. In: Proc. of the 28th AAAI Conf. on Artificial Intelligence. AAAI, 2014. 431−437.

[3] Tan ZX. Spatial advertisement competition: Based on game theory. Journal of Applied Mathematics, 2014,216193:1−5. [doi: 10.1155/2014/216193]

[4] Agarwal A, Hosanagar K, Smith MD. Location, location, location: An analysis of profitability of position in online advertising markets. Journal of Marketing Research, 2011,48(6):1057−1073. [doi: 10.1509/jmr.08.0468]

[5] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: Real-Time event detection by social sensors. In: Proc. of the 19th Int'l Conference on World Wide Web. Raleigh, 2010. 851−860. [doi: 10.1145/1772690.1772777]

[6] Nishi K, Tsubouchi K, Shimosaka M. Hourly pedestrian population trends estimation using location data from smartphones dealing with temporal and spatial sparsity. In: Proc. of the 22nd ACM Int'l Conf. on Advances in Geographic Information Systems. Dallas/Fort Worth: SIGSPATIAL, 2014. 281−290. [doi: 10.1145/2666310.2666391]

[7] Kotzias D, Lappas T, Gunopulos D. Addressing the sparsity of location information on Twitter. In: Proc. of the Workshops of the Joint Conf. of the 17th Int'l Conf. on Extending Database Technology and the 17th Int'l Conf. on Database Theory. Athens: EDBT/ICDT, 2014. 339−346.

[8] Cheng ZY, Caverlee J, Lee KM. A content-driven framework for geolocating microblog users. ACM Trans. on Intelligent Systems and Technology, 2013,4(1):Article 2. [doi: 10.1145/2414425.2414427]

[9] Ryoo KM, Moon S. Inferring Twitter user locations with 10km accuracy. In: Proc. of the 23rd Int'l World Wide Web Conf. Seoul, 2014. 643−648.

[10] Kinsella S, Murdock V, O'Hare N. "I'm eating a sandwich in Glasgow": Modeling locations with tweets. In: Proc. of the 3rd Int'l CIKM Workshop on Search and Mining User-Generated Contents. Glasgow, 2011. 61−68. [doi: 10.1145/2065023.2065039]

[11] Wang ZF, Feng J, Xing CY, Zhang GM, Xu B. Research on the IP geolocation technology. Ruan Jian Xue Bao/Journal of Software, 2014,25(7):1527−1540 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4621.htm [doi: 10.13328/j.cnki.jos.004621]

[12] Hecht B, Hong LC, Suh BW, Chi EH. Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In: Proc. of the Int'l Conf. on Human Factors in Computing Systems. Vancouver, 2011. 237−246. [doi: 10.1145/1978942.1978976]

[13] Eisenstein J, O'Connor B, Smith NA, Xing EP. A latent variable model for geographic lexical variation. In: Proc. of the 2010 Conf. on Empirical Methods in Natural Language Processing. MIT Stata Center, 2010. 1277−1287.

[14] Kwak H, Lee CH, Park HS, Moon S. What is Twitter, a social network or a news media? In: Proc. of the 19th Int'l World Wide Web Conf. Raleigh, 2010. 591−600. [doi: 10.1145/1772690.1772751]

[15] Ahmed A, Hong LJ, Smola A. Hierarchical geographical modeling of user locations from social media posts. In: Proc. of the 22nd Int'l World Wide Web Conf. Rio de Janeiro: WWW, 2013. 25−36.

[16] Chang HW, Lee DW, Eltaher M, Lee JK. @Phillies tweeting from philly? Predicting twitter user locations with spatial word usage. In: Proc. of the Int'l Conf. on Advances in Social Networks Analysis and Mining. Istanbul: ASONAM, 2012. 111−118. [doi: 10.1109/ASONAM.2012.29]

[17] Backstrom L, Kleinberg JM, Kumar R, Novak J. Spatial variation in search engine queries. In: Proc. of the 17th Int'l Conf. on World Wide Web. Beijing, 2008. 357−366. [doi: 10.1145/1367497.1367546]

[18] Ren KJ, Zhang SW, Lin HF. Where are you settling down: Geo-locating Twitter users based on tweets and social networks. In: Proc. of the 8th Asia Information Retrieval Societies Conf. on Information Retrieval Technology. Tianjin, 2012. 150−161. [doi: 10.1007/978-3-642-35341-3_13]

[19] Backstrom L, Sun E, Marlow C. Find me if you can: Improving geographical prediction with social and spatial proximity. In: Proc. of the 19th Int'l Conf. on World Wide Web. Raleigh, 2010. 61−70. [doi: 10.1145/1772690.1772698]

[20] Chandra S, Khan L, Muhaya FB. Estimating Twitter user location using social interactions—A content based approach. In: Proc. of 2011 IEEE the 3rd Int'l Conf. on the Privacy, Security, Risk and Trust and 2011 IEEE the 3rd Int'l Conf. on Social Computing. Boston: SocialCom/PASSAT, 2011. 838−843. [doi: 10.1109/PASSAT/SocialCom.2011.120]

[21] Jurgens D. That's what friends are for: Inferring location in online social media platforms based on social relationships. In: Proc. of the 7th Int'l Conf. on Weblogs and Social Media. Cambridge: ICWSM, 2013

[22] Li R, Wang SJ, Chang KCC. Multiple location profiling for users and relationships from social network and content. PVLDB, 2012, 5(11):1603−1614. [doi: 10.14778/2350229.2350273]

[23] McGee J, Caverlee J, Cheng ZY. Location prediction in social media based on tie strength. In: Proc. of the 22nd ACM Int'l Conf. on Information and Knowledge Management. San Francisco: CIKM, 2013. 459−468. [doi: 10.1145/2505515.2505544]

[24] Li W, Serdyukov P, de Vries AP, Eickhoff C, Larson M. The where in the tweet. In: Proc. of the 20th ACM Conf. on Information and Knowledge Management. Glasgow, 2011. 2473−2476. [doi: 10.1145/2063576.2063995]

[25] Zhao RJ, Cao SX. A user relationship-based approach for location recommendation in microblog. In: Proc. of the 11th National Seminar on Internet and Audio/Video and Broadcasting Development. Wuhan, 2012. 165−169 (in Chinese with English abstract).

[26] Guo C, Liu JN, Fang Y, Luo M, Cui JS. Value extraction and collaborative mining methods for location big data. Ruan Jian Xue Bao/Journal of Software, 2014,25(4):713−730 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4570.htm [doi: 10.13328/j.cnki.jos.004570]

[27] Yardi S, Boyd D. Tweeting from the town square: measuring geographic local networks. In: Proc. of the 4th Int'l AAAI Conf. on Weblogs and Social Media. AAAI, 2010. 194−201.

[28] McGee J, Caverlee J, Cheng ZY. A geographic study of tie strength in social media. In: Proc. of the 20th ACM Conf. on Information and Knowledge Management. Glasgow: CIKM, 2011. 2333−2336. [doi: 10.1145/2063576.2063959]

[29] Lichtenwalter R, Lussier JT, Chawla NV. New perspectives and methods in link prediction. In: Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Washington, 2010. 243−252. [doi: 10.1145/1835804.1835837]

[30] Kwak H, Lee CH, Park H, Moon S. What is Twitter, a social network or a news media? In: Proc. of the 19th Int'l Conf. on World Wide Web. Raleigh: WWW, 2010. 591−600.

[31] Wang X, Xu M, Ren YZ, Xu J, Zhang HP, Zheng N. A location inferring model based on tweets and bilateral follow friends. Journal of Computer, 2014,9(2):315−321. [doi: 10.4304/jcp.9.2.315-321]

[32] Zhai CX, Lafferty JD. A study of smoothing methods for language models applied to information retrieval. ACM Trans. on Information Systems, 2004,22(2):179−214. [doi: 10.1145/984321.984322]

[33] Che WX, Li ZH, Liu T. LTP, A Chinese language technology platform. In: Proc. of the Coling 2010: Demonstrations. Beijing, 2010. 13−16.

[34] Vardi Y, Zhang CH. The multivariate L_1-median and associated data depth. Proc. of the National Academy of Sciences, 2000,97(4): 1423−1426. [doi: 10.1073/pnas.97.4.1423]

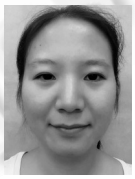
附中文参考文献:

[11] 王占丰,冯径,邢长友,张国敏,许博.IP 定位技术的研究.软件学报,2014,25(7):1527−1540. http://www.jos.org.cn/1000-9825/4621.htm [doi: 10.13328/j.cnki.jos.004621]

[25] 赵荣娇,曹三省.一种基于用户关系的微博位置推荐方法.见:第 11 届全国互联网与音视频广播发展研讨会.武汉,2012. 165−169.

[26] 郭迟,刘经南,方媛,罗梦,崔竞松.位置大数据的价值提取与协同挖掘方法.软件学报,2014,25(4):713−730. http://www.jos.org.cn/1000-9825/4570.htm [doi: 10.13328/j.cnki.jos.004570]

王凯(1988－),男,内蒙古巴彦淖尔人,博士生,主要研究领域为空间数据挖掘与推理,智慧城市.



吴敏(1988－),女,硕士,主要研究领域为数据挖掘,社交网络分析.

余伟(1987－),男,博士,讲师,CCF 会员,主要研究领域为数据质量评估,数据抽取,数据融合.

胡亚慧(1980－),女,博士生,主要研究领域为数据挖掘,城市计算,大数据处理.

杨莎(1980－),女,博士生,讲师,主要研究领域为电子商务数据挖掘与分析,电子商务服务质量评估,互联网经济行为数据挖掘.

李石君(1964－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为大数据,互联网搜索与挖掘,数据挖掘,数据库技术,移动数据挖掘与时空一致性研究.