DOI:10.13203/j.whugis20200212 文章编号:1671-8860(2020)06-0808-08

基于中文分词的加权地理编码在COVID-19疫情防控空间定位中的应用

彭明军[1] 李宗华[2] 刘 辉[1] 孟 成[1] 李 勇[3]

1 武汉市自然资源和规划信息中心,湖北 武汉,430014
2 武汉市政务服务和大数据管理局,湖北 武汉,430012
3 武汉大学测绘遥感信息工程国家重点实验室,湖北 武汉,430079

摘 要:地理编码是实现带有地址描述的信息空间定位的重要途径。比较研究了国内外地理编码方法,分析了中文地址的组成方式和定位方法。针对中文地址高度复杂性和多样性的特征,设计了一种顾及多种语义的地址匹配算法,并以武汉市新型冠状病毒肺炎(coronavirus disease 2019,COVID-19)病人入院时登记的地址描述信息为例,对匹配算法进行了实验验证,将匹配结果进行空间定位。结果表明,所提出的中文分词的加权地理编码方法匹配高效、定位准确、方法智能,能够实现基于语义的COVID-19病人入院时登记地址的快速定位,可为疫情防控提供准确的空间定位信息。

关键词:中文分词;新型冠状病毒肺炎;地理编码;空间定位;疫情防控

中图分类号:P208 **文献标志码**:A

在新型冠状病毒肺炎(coronavirus disease 2019,COVID-19)疫情防控中,对病人提供的居住地址进行精确的空间定位是开展流行病学调查的重要内容之一。通过统计患者的空间分布,可在防控阶段及时追踪人口往来动向,为地区疫情差异化防控提供技术支撑,对分析疾病传染途径和传播范围具有重要意义,是社会治理、应急管理和指挥决策工作的重要基础[1]。

地理编码(geocoding,又称地址编码)是指将自然语言描述的地址位置信息通过既定的地址模型或编码规则与空间位置相关联,从而确定其所代表的地理实体的位置[2-3]。由于社会经济信息(如人口、工商、民政、公安、社保、医保)一般都包含地址描述,因此可通过地理编码方法对其进行空间定位,从而实现地理空间数据和统计数据的有效集成,为研究和揭示各种社会经济现象的空间分布规律提供科学基础。利用地址匹配技术可以建立空间信息与非空间信息的联系[4-5],是实现社会经济信息与空间信息关联的有效途径之一。

1 国内外研究现状

1.1 地理编码

国外的地理编码技术发展得比较成熟,如美国建立了双重独立地图编码系统[2]、TIGER(topologically integrated geographic encoding and referencing)系统[3-4],英国基于British Standard 7666标准建立了全国地址数据库等,ArcGIS和MapInfo等商业化软件中包括了geocoding功能模块[4]。文献[5]提出了基于标准化地址的地理编码方法,通过对比非标准化地址的匹配结果,发现基于标准化地址的匹配结果精度有显著改善;文献[6]针对地理编码结果的匹配准确率问题,在综合考虑空间强度、聚类和聚集等因素后,利用统计方法进一步验证了地理编码结果的匹配准确率。与英语等语言不同,中文基本上没有形态变化,一个中文语句通常由一组前后连续的汉字组成,词与词之间没有明显的分界标志[7-8]。汉语的书面表达方式是以汉字为最小单位,因此,忽略中文地址的特殊性,直接沿用国外现有地理编码的思路无法达到理想的效果[9-10]。鉴于

收稿日期:2020-05-13

项目资助:国家重点研发计划战略性国际科技创新合作重点专项(2016YFE0202300);自然资源部研究项目(4201-240100123);中国工程院咨询研究项目(2020ZD16);国家自然科学基金(41771454);湖北省自然科学基金创新群体项目(2018CFA007)。

第一作者:彭明军,博士,教授,主要研究方向为智慧城市、GIS和遥感应用。pmj@zrzyhgh.wuhan.gov.cn

此，自20世纪80年代以来，中国学者开始了中文地址编码的研究工作。在中文地址分词中，对地址描述自动识别词边界，将汉字串切分为正确的词串的汉语分词问题是实现中文分词中的首要问题[9]。将中文分词方法[10-11]引入到地理编码中，并将中文地址切分成较小的地址单元，再在标准化的基础上进行匹配[12]，成为了目前地理编码的主要思路。在中文地址分词中，有基于ArcGIS软件geocoding功能的中文地址编码方法[13]、基于词典和规则切分的方法[11]、基于大规模语料库的统计方法、基于规则和统计相结合的方法等[12-14]，在实际应用中取得了一定的效果。人工智能技术的不断发展也为中文分词提供了新的途径和方法。这种分词方法又称为理解分词法，主要分为两种：一种是基于生理学的模拟方法[15]，如神经网络等；另一种是基于心理学的符号处理方法，如专家系统等；另外还包括决策树、随机森林算法等机器学习方法[16]。

1.2　疫情空间定位

对患者进行空间定位是开展人员管理、社区治理以及政府决策的重要基础，一方面可为疾控部门进行病人的定位和跟踪、开展疾病防控和医疗观察提供必要的条件；另一方面可为社区掌握本辖区实际疫情情况、患者分布等提供基础，并为开展疾控管理、人员救治和安全防护提供依据；同时，还能对政府开展应急管理、指挥调度、医疗物资投放、防控力量部署等提供参考[1]。

针对COVID-19疫情，中国的研究机构和学者通过不同途径和数据来源开展了疫情空间定位和趋势分析，主要包括两个方面：(1)在疫情暴发初期，针对人群移动轨迹开展了疫情地图，如百度迁徙(https://qianxi.baidu.com/?from=mapp)、智慧足迹、支付宝同乘查询等依据运营商手机定位、交通部门订票信息等大数据展开人员跟踪和定位，为疫情初期人员跟踪和防控提供了先进的手段和依据；(2)在疫情暴发过程中，根据患者空间位置(通常按照行政管理单元如区(县)级划分空间单元)进行空间分布统计并绘制疫情地图，用以反映疫情的时序变化和趋势特征。随着疫情在全球暴发，国外的研究机构和学者也开展了COVID-19疫情的研究工作，如美国约翰霍普金斯大学的董恩盛基于ArcGIS Dashboar搭建了全球疫情空间分布平台，利用全球疫情数据，结合有关地图和图表，对全球各个国家以及州市的疫情情况进行了展示(https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6)。

综上所述，目前国内外研究主要存在两方面问题：一是由于中文地址描述的特征以及数据标准化、规范化等诸多问题使得地址描述的形式多种多样，为中文地址编码带来了一定困难；二是目前疫情空间分布统计基本定位至区(县)级尺度空间单元，无法满足面向社区级分级防控、精细化治理、人员管理等方面的需要。因此，为满足不同来源的地址空间化需求，分析中国地名地址特有的组成形式和规律，本文将中文分词方法与地理编码相结合，提出了一种基于中文分词的加权地址层级模型。该模型首先根据不同地址在空间层级定位的差异性设定层级模型权重，然后计算待匹配地址与匹配地址之间的相似度，从而判断原地址与匹配地址的接近程度，并应用于武汉市COVID-19病例数据的空间定位。

2　地理编码方法

地理编码方法主要包括隐马尔可夫模型(hidden Markov model，HMM)、地理编码规则和层次编码模型3个部分。

2.1　HMM模型

HMM属于统计模型[17]，该模型包括2个序列和3个概率矩阵，即观测序列O、隐含状态序列Z、初始状态概率矩阵$\boldsymbol{\pi}$、状态转移概率矩阵$\boldsymbol{A}$及状态生成观测的概率矩阵$\boldsymbol{B}$。

HMM模型的建立满足两个假设：(1)齐次马尔可夫性假设，即任意时刻t的状态只受$t-1$时刻所处状态的影响，而与其他时刻的状态无关[18]；(2)所有的观测状态具有独立性假设，即任意时刻的观测状态都只受前一时刻的马尔可夫链状态影响，而与其他时刻无关[18]。

2.2　地理编码规则

在城市中进行地理编码涉及到建筑物、地块、道路等地理对象，对其进行描述的地理编码数据可分为基于建筑物的地理编码、基于地块的编码和基于道路的编码[12]。基于道路的编码以道路中心线为基础，记录道路的各种信息，包括名字、起始点标识、地址门牌号范围、街道中心线左右门牌号的起始点坐标以及属性数据和方向。由于基于建筑物或地块的地址采用点或多边形方式直接存储在数据库中，因此不需要通过道路或路段地址范围的中间环节过渡，就能实现建筑、宗地和点地物与地址的直接关联[11]。

2.3 层次编码模型

中文地址一般可分为行政区地名、自然地名、街巷(道路、胡同)、道路门牌(门址)、住宅区、冠楼名等类型[12]。中文标准地址的形式由3部分组成,即〈标准地址〉::=〈行政辖区〉〈基本区域限定物〉〈局部点位置描述〉。其中,〈行政辖区〉为政区类地名;〈基本区域限定物〉可以是道路、小区等低于行政区地名的层次;〈局部点位置描述〉是具有确定的空间位置,可以用点表示的基本地址元素,包括门牌、楼号、冠名楼、村落等类型。由于空间实体之间存在包含、嵌套等关系,因此各类地址或地名之间也就存在上下级的层次关系。各类地址间的层次关系如图1所示。

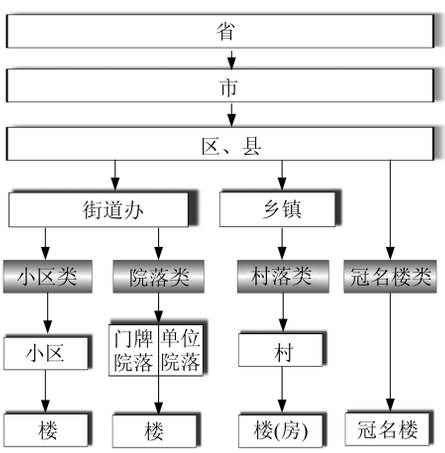

图1 面-点结构地址编码层次关系图

Fig.1 Hierarchy Diagram of Plane-Point Structure Address Coding

中文地址具有层次关系,从空间关系上,地址可分为有从属关系和跨从属关系两类,省、市、区(县)、街道、社区、小区、地片、标志物等按行政区划范围从大到小可以建立从属关系。道路、街、巷的从属关系不明显,存在道路跨越多个行政区、街的情况。根据编码规则及地址层次模型建立中文地址的概念模型,包括城市各级管理边界类、街区类、道路中心线类、点地址类、面地址类等,需要建立这些类之间的相关关系。

3 地址匹配流程

本文提出的基于中文分词的加权地理编码方法主要包括3个部分,即文本分词与权重计算、共有词选择和文本匹配结果的选取,具体流程如图2所示。

3.1 文本分词与权重计算

文本分词是通过对中文文本词语进行自动识别来达到对文本自动分割的目的。由于地名地址的笼统性、复杂性与特殊性,难以遵循一些固有规律[8,11],因此,本文首先使用维特比算法计算初始状态的概率值,然后逐步计算各时刻的转移概率,再利用动态规划求解HMM模型,从而完成对文本的分词[17]。

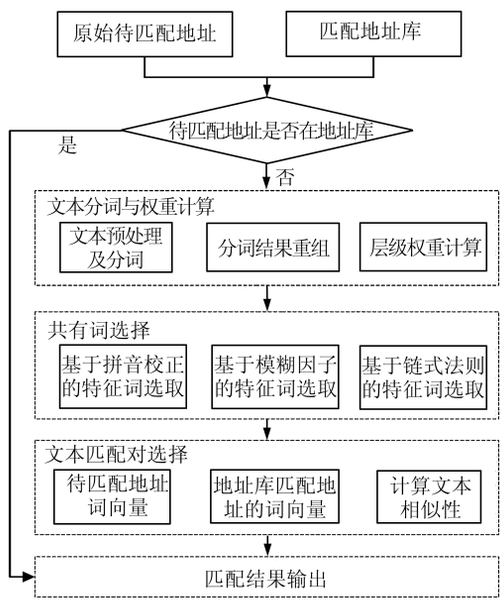

图2 基于中文分词的加权地理编码方法

Fig.2 Weighted Geocoding Method Based on Chinese Word Segmentation

由于地址层级描述的多样性,为提高分词的准确性,本文建立武汉市街道、社区、小区和路网等专有地名地址库,作为基于马尔可夫模型分词结果的补充。通过结合特征词库与文本词典方法[11-12],一定程度上解决了一些不常见的地址分割,如冯家畈等罕见村湾地名。

根据地址层次模型,不同等级的地址在描述地址时的重要程度不同。为进一步表征地址的层级特征,采用高斯模型度量各层级的权重。如“江岸区中山大道××号江花庭院”与“江岸区三阳路××号江花庭院”两条地址,江花庭院是小区名称,道路名分别是“中山大道××号”与“三阳路××号”,地址完全迥异,但表达的小区是一致的。这表明道路名和地址编号在地址描述中存在较大的差异性,而社区、小区层级却是固定不变的。基于此,本文基于标准地址及其层次模型建立以社区、小区为中心的归一化正态分布函数,衡量各层级地址的权重并进行归一化。由于编号式描述的地址差异性较大,如“中山大道××号3栋”与“三阳路××号3栋”,因此本文建立了部分街道交叉路口、高校、公司企业、酒店

等特殊指代地方的别名词典,以解决此类纯编号不一致的情形。

3.2　共有词选择

在本文地理编码方法中,共有词选择是计算文本相似度的重要步骤。通过对待匹配地址进行梳理,发现待匹配地址与地址库存在3类情形:(1)存在因口误、前后鼻音、平舌、卷舌以及多音字引起的登记错误,对此采用基于拼音校正的特征词选取方法基本解决了此类问题。(2)存在同一地名地址描述形式的多样性问题,本文提出了基于链式法则的特征词选择方法,该方法在很大程度上改善了此类问题。(3)存在某些地名地址未建立别名的情形,对此本文采用基于模糊因子的方法度量二者之间的相似性,在一定程度上改善了该类问题。由于情形(1)和(3)存在的情况较少,本文着重介绍情形(2)的解决方法。

基于链式法则的特征词选择是根据不同词组在地址中的重要程度并借鉴链式求导法则原理提出的方法。根据链式求导法则可知,参数求导依赖于函数映射关系,函数可微性是函数是否可导的必要条件。借鉴该原理,将分词后结果分为中文和非中文两种特征词,其中中文特征词为主特征词,非中文特征词为次特征词。文中约定,当次特征词不能与主特征词关联时,认定其为无效链式。当有多个主特征词时,会将原地址划分为多个主特征词-次特征词形式的子地址串,再进行共有词的选取。如图3所示,将待匹配地址划分为“中山大道975号”与“华清园2栋3单元”两个地址链,其中“中山大道”与“华清园”为主特征词,“975号”、“2栋”、“3单元”为次特征词,在地址库中,虽然同时出现了“2栋”和“3单元”等次特征词,由于没有“华清园”主特征词,“江花庭院2栋3单元”组成的链式无效,那么“2栋”和“3单元”也无效,因此二者共有词为“中山大道”。

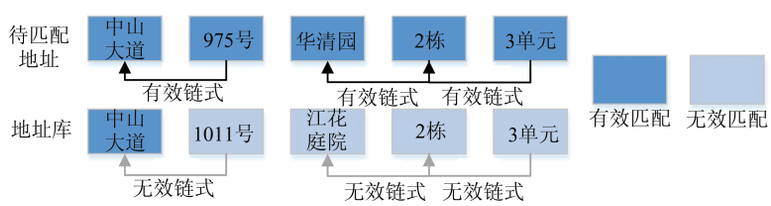

图3　基于链式法则的特征词选取

Fig.3　Feature Word Selection Based on Chain Rule

3.3　文本匹配对选择

为了计算文本之间的相似程度,采用如下方法计算匹配对之间的相似性,以克服人为主观因素对匹配结果的影响,实现匹配对的自动化选择。

1)待匹配地址词向量化。借鉴自然语言中词向量技术,将文本分词的词序结果进行词频统计,按照文本分词顺序构建待匹配地址的词向量。同理,对地址库中的候选地址作同样处理,得到其词向量。如图3中的待匹配地址词向量为[1,1,1,1,1],地址库中的地址词向量为[1,0,0,0,0]。

2)文本相似度计算。在乘上权重基础上,采用余弦相似度字符串与候选地址的文本相似度,选取最大相似度为文本匹配结果。

4　实验分析

4.1　实验数据

为了验证本文方法的可行性,采用2020年1月—3月COVID-19疫情期间医院记录的病人地址数据进行实验。其中,病人地址数据约4.9万条,匹配地址库包含约78万条2019年高德POIs(point of interests)和约140万条公安地址数据。由于病人描述地址时受地方口音、医院记录、手写笔误及时间紧迫等多方面因素的影响,普遍存在3种问题:(1)地址描述不完整或缺失行政区、街道或社区门牌等信息,如武汉市硚口区不详乡镇;(2)地址或门牌号描述不规范,如湖北省武汉市洪山区东湖高新区;(3)受人为客观原因的影响,如语言上平舌、卷舌、前后鼻音等原因,导致多音字、生僻字的记录问题,如“同馨花苑”记录为“同心花园”。

针对以上情况,首先对数据进行统一标准化处理,消除这种不明因素带来的问题;其次以武汉市高德POI数据和公安部门地址数据建立匹配地址库;再次根据街道、社区、小区、村镇等信息建立用户词典和同名词典,作为中文分词的依据;最后采用Python语言实现本文算法。

4.2　实验结果和讨论

根据本文提出的匹配方法对武汉市病例数据与高德POI数据、公安地址数据进行匹配实验。部

分匹配结果如表1～3所示。表1是基于拼音校正方法得到的匹配结果示例,其中待匹配地址为疫情医院记录的地址,匹配地址为采用本文方法匹配得到的结果地址。表1的匹配结果表明,基于汉语拼音校正的方法增加了正确匹配数,提高了匹配的准确率。由于没有考虑疫情期间汉语拼音、当地口音和人员记录等客观因素,文献[11,15]无法正确匹配到上述情况下的正确地址,而本文提出的方法综合考虑了这些客观因素,一定程度上解决了多音字、平舌卷舌、前后鼻音和边音等问题。

表1 基于拼音校正的匹配结果示例

Tab.1 Examples of Matching Results Based on Pinyin Correction

待匹配地址	匹配结果
江岸区四唯街麟趾社区四唯街道麟炙小区41号×××室	武汉市江岸区四唯街办事处麟趾小区9栋×层×××室易瑞丰地产
汉阳区龙阳街龙阳新村社区龙兴路龙阳欲园5-××-×××	武汉市汉阳区龙阳街龙阳新村社区龙阳御园8号楼4-×××号商铺
汉阳区五里墩街五春里社区玉龙湾2期×栋×××号	武汉市汉阳区五里墩街道钰龙湾临时营销中心
洪山区关山街枫林上城社区紫崧风铃上城9-×-××××	武汉市洪山区关山街道桃园路紫松枫林上城×-×号门面
汉阳区江堤街江城明珠社区南庭熙园7栋×××号	武汉市汉阳区江堤街道下马湖路广电兰亭熙园2栋商铺

表2是对描述层级多样化的地址采用本文方法匹配得到的结果。表3是通过建立别名,采用模糊匹配方法得到的数据匹配结果。它们都是针对地址描述多层级和别名系统的匹配,是利用本文方法正确匹配的情况。表2表明,针对层级描述的地址,如“武汉市硚口区荣华街道荣西社区57号×栋××”,经在地址库中查询,有“武汉市硚口区古田路57号”以及与“荣西社区”相关的若干条记录,采用基于共有词或概率统计的方法会将前者作为匹配结果,导致出现匹配错误的情况。根据层级模型,本文赋予街道或社区层级地址更高的权重,基于链式法则的方法在进行共有词选择时,因“荣西社区”不在该地址中,会认定“57号”、“3栋”等关键词都是无效链式,从而判定待匹配地址与该地址不存在共有词。本文选择“硚口区荣华街荣西社区居民委员会”作为社区级或小区级的匹配结果。

表2 描述层级多样化地址的匹配结果示例

Tab.2 Examples of Matching Results Based on Multiple Descriptions

待匹配地址	匹配结果
武汉市硚口区荣华街荣西社区57号×栋××号	武汉市硚口区武广商圈武胜路113号西70米荣华街荣西社区居民委员会
东湖高新区关东街关东社区民族大道99号健龙尚谷杰座1期3栋××号	武汉市东湖高新区民族大道99号3栋
汉南区沌阳街海滨城社区车城大道243号香格里居1栋×单元×楼×号	武汉市沌口开发区车城大道243号1栋
汉南区沌阳街江大园社区博学路1号江大园9栋-×单元-××号	武汉市沌口开发区博学路1号9栋

由于模糊匹配是最后无法匹配时考虑的情况,因此建立的别名系统不够完善,匹配的示例也相对较少。如表3所示,待匹配地址为一些简称,通过建立别名库和计算模糊相似度,能够找到最优的待匹配地址。如表3中的“东胜擎天”与匹配地址中的“东顺擎天”,由于“东胜擎天”不存在别名,因此通过其与“5栋”共同组建“东胜擎天5栋”地址,再采用模糊匹配法进行匹配以寻找最优的地址。由此可见,基于地址层级模型、拼音校正和链式法则的共有词方法和别名库、模糊相似度的度量方法能有效解决文本中行政地址多样化等匹配问题。

4.3 方法评价

由于任务的特殊性,本文认定只要有返回结果都看作是地址的最优匹配结果,没有返回结果的作为本文方法漏匹配的数据。本文以匹配到小区、村为人工核查的匹配标准,采用人工检查的方法核对了疫情地址的所有匹配情况,并将实际匹配结果作为方法有效性评价的基础。表4统计了本文匹配方法得到的正确匹配数、错误匹配数、模糊匹配数、漏匹配数和匹配准确率。其中,模糊匹配数为没有明确社区、小区地址、没有道路编号以及村镇等地址,漏匹配数为没有识别出结果的数据。人工核查发现,原地址中存在大量拼音错误及记录错误的情况。本文方法因考虑了拼音问题、病人主观意愿(如不太愿意提供更详细地址)、人工记录(多音字、手写笔误、简化记录)等情况,减少了漏匹配和错误匹配的情况。这是因为本文方法通过建立地址层级模型,考虑了特征词的重要程度,并依据汉字拼音、别名等

原则,从而降低了错误匹配和漏匹配的情况,提高了匹配的准确率。

表 3　采用模糊匹配的结果示例

Tab.3　Examples of Matching Results Based on Fuzzy Matching

待匹配地址	匹配结果
东湖高新区关东街江南社区观澜园路万科城花璟苑4-×-××××号	武汉市东湖高新区关南园8号4栋
武昌区黄鹤楼街读书社区武大人民医院	武昌区黄鹤楼街武汉大学人民医院
东西湖区泾河街永丰大队东胜擎天5-×-××	东西湖区东顺擎天5栋
东西湖区泾河街永丰大队广景北区401-××	东西湖区海景北区401栋×号

表 4　本文方法匹配核查结果

Tab.4　Verification Results of the Proposed Matching Method in This Paper

匹配指标	数值项
正确匹配数/条	41 176
错误匹配数/条	5 552
模糊匹配数/条	2 284
漏匹配数/条	762
匹配准确率/%	82.72

为满足公共疫情防控需求,服务政府决策与评估,需要提供疫情在城市空间中的统计和分布情况,因此,在匹配完成后需要返回匹配地址的空间坐标,以满足城市疫情管控的定位需求。同时,以社区为单位,将匹配结果得到的COVID-19病例的定位坐标落到空间上,定位其所在社区的空间位置,并将定位结果按照社区患者数量进行分级可视化展示。图4是以社区为空间尺度,将武汉市中心城区患者的登记地址定位到社区的分布图。

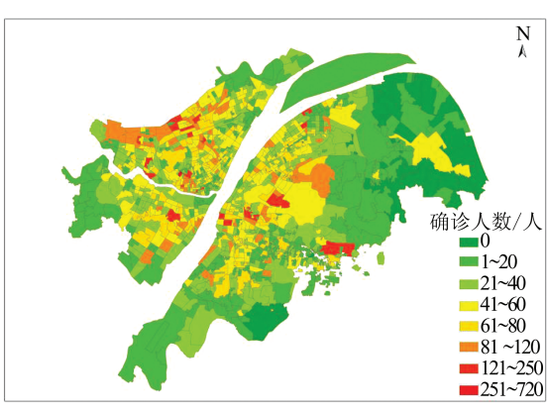

图4　COVID-19病例社区定位分布图

Fig.4　Location Map of COVID-19 Cases in the Community

5　结　语

本文主要探讨了基于中文分词的加权地理编码方法,研究了基于地址层级模型的地址模型构建方法,设计了基于汉语拼音校正、别名地址库和特征词重要度的地理编码方法,通过该方法实现了武汉市疫情记录的地址匹配和定位。实验结果表明:

1)该地址匹配方法能够较好地解决如“广电南亭”与“广电兰亭”、“风铃上城”与“风林上城”等地址描述中出现的汉语拼音前后鼻音混淆、多音字等匹配问题。

2)该方法能够在一定程度上根据规则自动识别和过滤由于地址描述不规范导致的地址匹配错误的问题,如地址描述中有超过两个“-”连接符的情况,如“武汉市江岸区百步亭社区管委会百步亭花园路158-××-×-×××”。

3)本文方法能够有效识别地址描述中层级缺失或重叠的情况,如“武汉市硚口区不详乡镇发展社区6-×-×××室”,并动态调整分词地址单元的权重,以提高匹配准确度。

综上所述,绝大多数地址匹配精度不高都可归结为中文地址描述的特征及其标准化、规范化等问题所导致,而目前主要通过大量的人工处理来解决此问题,尚未有高效智能的方法。因此,下一步将着重研究如何提高程序对这种特殊记录的对抗性、兼容性和匹配结果的准确度。

参　考　文　献

[1] Feng Mingxiang, Fang Zhixiang, Lu Xiongbo, et al. Traffic Analysis Zone-Based Epidemic Estimation Approach of COVID-19 Based on Mobile Phone Data: An Example of Wuhan[J]. *Geomatics and Information Science of Wuhan University*, 2020, 45(5):651-657,681(冯明翔, 方志祥, 路雄博, 等. 交通分析区尺度上的COVID-19时空扩散推估方法:以武汉市为例[J]. 武汉大学学报·信息科学版, 2020,45(5):651-657,681)

[2] Edgar H P. Introduction to the GBF/DIME: A Prime[J]. *Computers, Environment and Urban Systems*, 1983,8(3): 135-173

[3] Zandbergen P A. A Comparison of Address Point,

Parcel and Street Geocoding Techniques [J]. *Computers, Environment and Urban Systems*, 2008, 32(3): 214-232

[4] ESRI. ArcGIS Online Geocoding Service [EB/OL]. http://geocode.arcgis.com/arcgis/, 2020

[5] Matci D K, Avdan U. Address Standardization Using the Natural Language Process for Improving Geocoding Results [J]. *Computers, Environment and Urban Systems*, 2018,70: 1-8

[6] Briz-Redón Á, Martinez-Ruiz F, Montes F. Reestimating a Minimum Acceptable Geocoding Hit Rate for Conducting a Spatial Analysis [J/OL]. *International Journal of Geographical Information Science*, 2019, DOI: 10.1080/13658816.2019.1703994

[7] Computational Linguistics Lab, Institute of Applied Linguistice Ministry of Education. Standardized Set of Chinese POS Markers for Computational Uses [J]. *Applied Linguistics*,2001(3): 16-20(国家语委语言文字应用研究所计算语言学研究室. 信息处理用现代汉语词类标记集规范[J]. 语言文字应用, 2001(3): 16-20)

[8] Huang Changning, Zhao Hai. Chinese Word Segmentation: A Decade Review [J]. *Journal of Chinese Information Processing*, 2007,21(3): 8-19(黄昌宁, 赵海. 中文分词十年回顾[J]. 中文信息学报, 2007,21(3): 8-19)

[9] Long Shuquan, Zhao Zhengwen, Tang Hua. Overview on Chinese Segmentation Algorithm[J]. *Computer Knowledge and Technology*, 2009, 5(10): 2 605-2 607(龙树全, 赵正文, 唐华. 中文分词算法概述[J]. 电脑知识与技术, 2009, 5(10): 2 605-2 607)

[10] Liu Tao. The Improvement and Application of Marking in Chinese Address Automatic Segmentation [J]. *Computer Knowledge and Technology*, 2009, 5(11): 2 828-2 829(刘韬. 设立切分标志法在中文地址自动分词中的改进与应用[J]. 电脑知识与技术, 2009, 5(11): 2 828-2 829)

[11] Li L, Wang W, He B, et al. A Hybrid Method for Chinese Address Segmentation [J]. *International Journal of Geographical Information Science*, 2018, 32(1): 30-48

[12] Sun Cunqun, Zhou Shunping, Yang Lin. Chinese Geocoding Based on Classification Database of Geographical Names [J]. *Journal of Computer Applications*, 2010, 30(7): 1 953-1 955,1 958(孙存群, 周顺平, 杨林. 基于分级地名库的中文地理编码[J]. 计算机应用, 2010, 30(7): 1 953-1 955,1 958)

[13] Zhang Yifeng, Wu Jianping, Cheng Yi, et al. The Improvement of Geocoding in ArcGIS [J]. *Geomatics & Spatial Information Technology*, 2007, 30(3): 116-119(章意锋, 吴健平, 程怡, 等. ArcGIS中地理编码方法的改进[J]. 测绘与空间地理信息, 2007, 30(3): 116-119)

[14] Zhang Linman, Wu Sheng. Research on Place Names and Address Segmentation in Geocoding System[J]. *Science of Surveying and Mapping*, 2010, 35(2): 46-48(张林曼, 吴升. 地理编码系统中地名地址分词算法研究[J]. 测绘科学, 2010, 35(2): 46-48)

[15] Zhang Wenhao, Lu Shan, Cheng Guang. Design and Implementation of Chinese Address Segmentation Method Based on LSTM Networks [J]. *Application Research of Computers*, 2018, 35(12): 1-2(张文豪, 卢山, 程光. 基于LSTM网络的中文地址分词法的设计与实现[J]. 计算机应用研究, 2018, 35(12):1-2)

[16] Lin Y, Kang M, Wu Y, et al. A Deep Learning Architecture for Semantic Address Matching[J]. *International Journal of Geographical Information Science*, 2020, 34(3): 559-576

[17] Qian Zhiyong, Zhou Jianzhong, Tong Guoping, et al. Study on Automatic Word Segmentation of the Songs of Chu Based on HMM [J]. *Library and Information Service*, 2014, 58(4): 105-110(钱智勇, 周建忠, 童国平,等. 基于HMM的楚辞自动分词标注研究[J]. 图书情报工作, 2014, 58(4): 105-110)

[18] Gong Faming, Zhu Penghai. Word Segmentation Based on Adaptive Hidden Markov Model in Oilfield [J]. *Computer Science*, 2018, 45(6A): 97-100(宫法明, 朱朋海. 基于自适应隐马尔可夫模型的石油领域文档分词[J]. 计算机科学, 2018, 45(6A): 97-100)

Weighted Geocoding Method Based on Chinese Word Segmentation and Its Application to Spatial Positioning of COVID-19 Epidemic Prevention and Control

PENG Mingjun[1]　*LI Zonghua*[2]　*LIU Hui*[1]　*MENG Cheng*[1]　*LI Yong*[3]

1　Wuhan Natural Resources and Planning Information Center, Wuhan 430014, China

2　Wuhan Government Service and Big Data Administration Bureau, Wuhan 430012, China

3　State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

Abstract: Locating the coronavirus disease 2019 (COVID-19) cases in the accurate place is important in epidemic prevention and control. Geocoding is an effective method to achieve information space positioning with address description. The English based geocoding methodology is not suitable for Chinese address. Composition and positioning methods of Chinese address are discussed. According to the characteristics of high complexity and diversity of Chinese address, a Chinese word segmentation weighted address matching algorithm considering a variety of semantics is designed, including the same pronunciation but different Chinese word address, abbreviation and alias of Chinese address, different description of the same address. And the matching accuracy and efficiency of the algorithm are tested by using the COVID-19 cases' addresses in Wuhan. The result indicates the algorithm is efficient, accurate, and intelligent, which can realize the efficient location of the COVID-19 cases address, and provide accurate spatial location information for epidemic prevention and control by quickly positioning of the COVID-19 cases.

Key words: Chinese word segmentation; coronavirus disease 2019 (COVID-19); geocoding; spatial positioning; epidemic prevention and control

First author: PENG Mingjun, PhD, professor, specializes in smart cities, GIS and remote sensing applications. E-mail: pmj@zrzyhgh.wuhan.gov.cn

Foundation support: The National Key Research and Development Program of China on Strategic International Scientific and Technological Innovation Cooperation Special Project (2016YFE0202300); the Research Project from the Ministry of Natural Resources of China (4201-240100123); the Key Projects of Consultation and Research of the Chinese Academy of Engineering (2020ZD16); the National Natural Science Foundation of China (41771454); the Innovation Group Project of Natural Science Foundation of Hubei Province (2018CFA007).

引文格式:PENG Mingjun, LI Zonghua, LIU Hui, et al. Weighted Geocoding Method Based on Chinese Word Segmentation and Its Application to Spatial Positioning of COVID-19 Epidemic Prevention and Control[J]. Geomatics and Information Science of Wuhan University, 2020, 45(6):808-815. DOI:10.13203/j.whugis20200212(彭明军, 李宗华, 刘辉, 等. 基于中文分词的加权地理编码在COVID-19疫情防控空间定位中的应用[J]. 武汉大学学报·信息科学版, 2020, 45(6):808-815. DOI:10.13203/j.whugis20200212)