

引文格式:王勇,刘纪平,郭庆胜,等.顾及位置关系的网络 POI 地址信息标准化处理方法[J].测绘学报,2016,45(5):623-630. DOI:10.11947/j.AGCS.2016.20150618.  
WANG Yong, LIU JiPing, GUO QingSheng, et al. The Standardization Method of Address Information for POIs from Internet Based on Positional Relation[J]. Acta Geodaetica et Cartographica Sinica, 2016, 45(5): 623-630. DOI: 10.11947/j.AGCS.2016.20150618.

## 顾及位置关系的网络 POI 地址信息标准化处理方法

王 勇<sup>1,2</sup>, 刘纪平<sup>2</sup>, 郭庆胜<sup>1</sup>, 罗 安<sup>2</sup>

1. 武汉大学资源与环境科学学院, 湖北 武汉 430079; 2. 中国测绘科学研究院, 北京 100830

### The Standardization Method of Address Information for POIs from Internet Based on Positional Relation

WANG Yong<sup>1,2</sup>, LIU Jiping<sup>2</sup>, GUO Qingsheng<sup>1</sup>, LUO An<sup>2</sup>

1. School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China; 2. Chinese Academy of Surveying and Mapping, Beijing 100830, China

**Abstract:** As points of interest (POI) on the internet, exists widely incomplete addresses and inconsistent literal expressions, a fast standardization processing method of network POIs address information based on spatial constraints was proposed. Based on the model of the extensible address expression, first of all, address information of POI was segmented and extracted. Address elements are updated by means of matching with the address tree layer by layer. Then, by defining four types of positional relations, corresponding set are selected from standard POI library as candidate for enrichment and amendment of non-standard address. At last, the fast standardized processing of POI address information was achieved with the help of backtracking address elements with minimum granularity. Experiments in this paper proved that the standardization processing of an address can be realized by means of this method with higher accuracy in order to build the address database.

**Key words:** POIs from internet; addresses tree; positional relation; standalization of address

**Foundation support:** The National High-tech Research and Development Program of China (863 Program) (Nos. 2012AA12A402; 2013AA12A403); The National Natural Science Foundation of China (No. 41471384); Research Projects of Public Welfare for Surveying and Mapping Industry (Nos. 201512021; 201512032)

**摘 要:** 针对互联网 POI(兴趣点)地址信息中广泛存在的地址要素不完整、文字表达不一致等不规范现象,提出一种顾及位置关系的网络 POI 地址信息标准化处理方法,首先对 POI 信息进行切分提取并逐层匹配地址树模型;然后基于 4 种位置关系从标准 POI 库中选出相应集合,作为丰富和修正非标准 POI 地址要素的候选;最后通过最小粒度地址要素的回溯,实现 POI 地址信息的快速标准化处理。试验表明该方法可以获得较高的准确率,尤其适用于在互联网数据环境中的 POI 地址信息标准化。

**关键词:** 网络 POI; 地址树; 位置关系; 地址标准化

中图分类号: P208

文献标识码: A

文章编号: 1001-1595(2016)05-0623-08

基金项目: 国家 863 计划(2012AA12A402; 2013AA12A403); 国家自然科学基金(41471384); 国家测绘地理信息局公益科研专项(201512021; 201512032)

近年来,随着互联网地理信息服务的蓬勃发展,网络 POI 已经成为大数据时代一种重要的空间信息资源。在我国,网络 POI 主要来源于地图服务商和用户标注,不同地图数据提供者对于同一个地址的文字表达不尽相同,而用户标注中的

地址信息也经常以口述和简化表达的方式来描述,使得同一个地址可能出现多种不同的文字表达,导致来源不同的 POI 数据融合困难,难以发挥多源信息的聚合作用。

地址标准化处理是网络 POI 数据清洗、融合与

分析的重要内容,是实现地址编码(geocoding)等网络服务的重要基础<sup>[1-8]</sup>,其核心是将不规范、不完整的“非标准”地址信息以符合常见地址表达模型的方式进行“规范化”处理和表达。现有的商业化地址标准化处理工具如 ArcGIS 的 Address Geocoding、MapInfo 的 MapMarker、Oracle 的 Spatial Geocoder 等,均基于内嵌判别规则来实现地址标准化<sup>[9-10]</sup>;文献<sup>[11]</sup>通过构建专家系统实现中文地址的标准化;文献<sup>[12-13]</sup>通过构建多层地址规则实现地名地址向标准化表达模型的转化;文献<sup>[14]</sup>采用决策树模型实现地址模式匹配。以上方法均需要构建大量领域规则或基于规则形成专家系统,这些方法能较好地满足英文地址信息的标准化与位置匹配要求,但对于中文地址信息处理效果较差,且规则构建过程需要大量人工参与。相比而言,机器学习方法可以基于大量标准化地址样本自动构建出地址要素间的组合规则,从而支持非标准化地址信息的标准化处理<sup>[15-22]</sup>,因而可移植性更强。文献<sup>[17]</sup>利用机器训练后获得的地址语料库及相关规则,通过局部模糊匹配实现地名地址解析与标准化;文献<sup>[20]</sup>利用半监督机器学习方法,基于 HMM 训练模型实现地名地址标准化;文献<sup>[21]</sup>通过总结中文地址模型的内部规则与空间约束关系,提出基于可扩展地址树的标准地址提取方法。然而,由于汉语言文字固有的地址描述信息不带分隔符等特点,使得基于机器学习的方法也存在样本需求较大、训练周期较长、标准化准确率较低等弊端。

以上基于规则和基于机器学习的地址标准化方法,侧重从纯文本(地址文本)分析角度挖掘地址信息的组合规则,而对 POI 的位置属性却未充分加以利用。本文试图提出一种顾及空间位置关系的网络 POI 地址信息标准化处理方法,以可扩展中文地址树模型为指导,首先基于特征词对待处理 POI 的地址信息进行地址要素切分、识别并与地址树模型逐层匹配,其次将待处理 POI 的地理坐标与标准参考库进行位置关系计算并形成参考对象库,最后根据最佳匹配结果完成待处理 POI 地址信息的标准化处理。

## 1 中文地址模型

### 1.1 中文地址的层次模型

中文地址模型是一种基于层次关系的排列模型,可分为政区级地址要素、街区级地址要素、门牌级地址要素 3 个层级,其中:政区级要素可细分

为国家名、省名、市名、区\县名、乡镇名等;街区级要素一般表现为道路、街巷、住宅区等基础限定物;门牌级要素一般表现为楼牌号、单位名称、标志物等局部点位置描述。针对中文地址的结构特征,以及目前我国地址模型存在多套标准的现状,本文设计了一种包含行政区划、基础地址限定物、局部点位置描述的 3 层地址树模型,如图 1 所示。

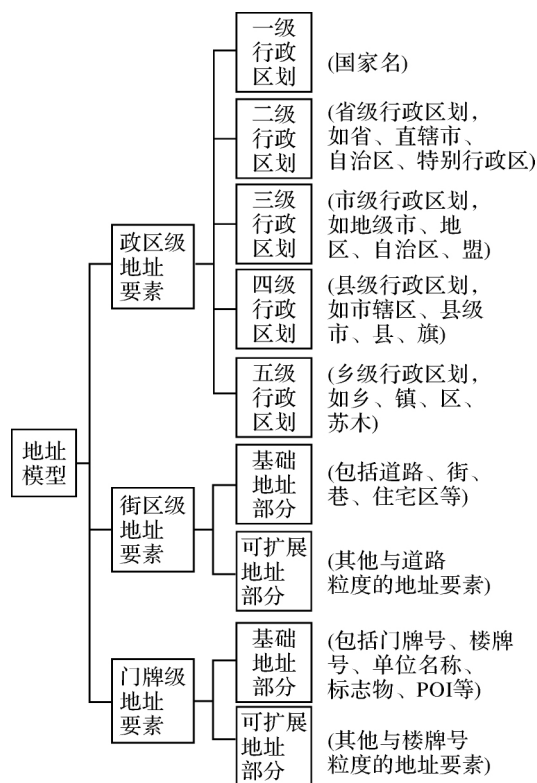


Fig.1 Composition of the address model

### 1.2 地址要素组合的限定关系

一个完整的中文地址由政区级、街区街、门牌级等 3 层要素构成,各层要素还可细分为不同的级别。对于某个具体的地址实例而言,上下级地址要素实例需要遵循一定的限定关系(通常为行政或管理意义上的隶属关系),如图 2 所示。这种要素实例的限定/映射关系普遍存在,是实现地址标准化尤其是缺失地址要素补全的重要依据。

## 2 POI 地址信息标准化处理

本文提出的 POI 地址信息标准化处理流程为:首先基于特征词典实现要素识别与切分,将输入的地址信息分割为多个地址要素;其次,通过匹配地址要素,构建各级要素的层次关系,形成地址树;再次,通过位置关系计算筛选出与待标准化

POI 紧密相关的参考样本;最后利用最小粒度回溯法,基于参考 POI 实现地址信息中缺失要素自动填充与标准化。

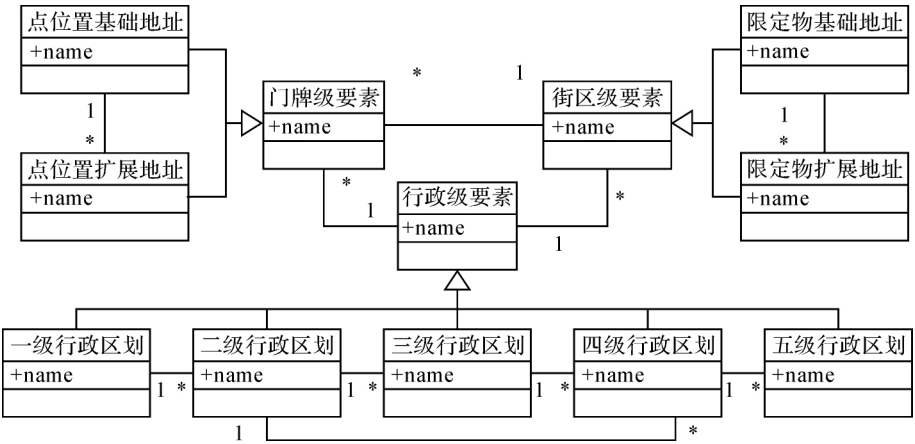


图 2 地址要素组合关系图

Fig.2 Relations of elements in the address tree

2.1 地址要素识别与切分

中文地址要素通常采用“专名+通名”的组合方式进行描述,如“北京市”、“海淀区”、“中关村创业大厦”。其中,通名是表征地址要素级别或类型的特征词,如“市”、“区”、“大厦”;一个地址要素中

除掉通名部分如“北京”、“海淀”、“中关村创业”即为专名,专名通常与通名相配合来完整表达一个地址要素。利用通名特征词可以很好地实现地址要素的切分和识别,本文使用的特征词库如表 1 所示。

表 1 地址要素类别与通名(特征词)列表

Tab.1 Type of address elements and feature words

地址要素大类	地址要素子类	通名(特征词)
政区级	一级行政区划	中国
	二级行政区划	省、市、县、区、直辖市、特别行政区……
	三级行政区划	市、地区、区、自治州、盟……
	四级行政区划	市辖区、区、市、县、自治县、特区、旗、自治旗……
	五级行政区划	乡、民族乡、苏木、镇、街道……
街区级	街道	路、街、道、巷、胡同、大街、大道、中路、条……
	住宅小区	庄、屯、里、弄、区、苑、园、院、坊、城、居……
门牌级	门牌号	号、楼、#、宿、公寓、堂、馆、斋……
	标志性建筑	广场、大厦、大楼、饭店、酒店、场、局、中心、公馆……

2.2 地址要素逐层匹配

在对地址信息进行要素切分后,需要根据地址树模型匹配处理,具体匹配方法是:读取一个待处理的地址信息后,首先按照 2.1 节所述的要素组成规则及特征词,将其切分为若干最小粒度的地址要素,然后顺次将各个地址要素与地址树模型的各个层次进行匹配。一旦某要素与地址树模型中的某一级别匹配成功,就将待处理的下一个地址要素与当前匹配级别的下级节点进行逐层比较直至成功匹配;若匹配失败,则将其作为成功匹配出的上级要素的下级节点。如此循环,直到所

有地址要素都匹配成功或都已经加入到地址树中。

地址要素匹配主要有 3 种情况:完全匹配、粗粒度匹配、细粒度匹配。

完全匹配:当能够从地址树中完全匹配到从地址信息中切分出的地址要素时,该地址树无须进行扩展,具体情况见匹配路径(图 3(a)),这属于完全匹配情况。

粗粒度匹配:根据切分出的地址要素的上下层次关系,上层较粗粒度的地址要素匹配成功,但下层细粒度的地址要素无法匹配成功。此时,可

自动将细粒度地址要素添加到地址树中,匹配过程见图 3(b)、(c)、(d) 3 条路径,其中虚线为扩展。

细粒度匹配:在匹配过程中,地址树中间某层

的地址要素无法匹配成功,该情况下可将未匹配成功的地址要素,插入到地址树中,并建立地址树的父子语义关系,匹配过程见图 3(e) 的匹配路径。

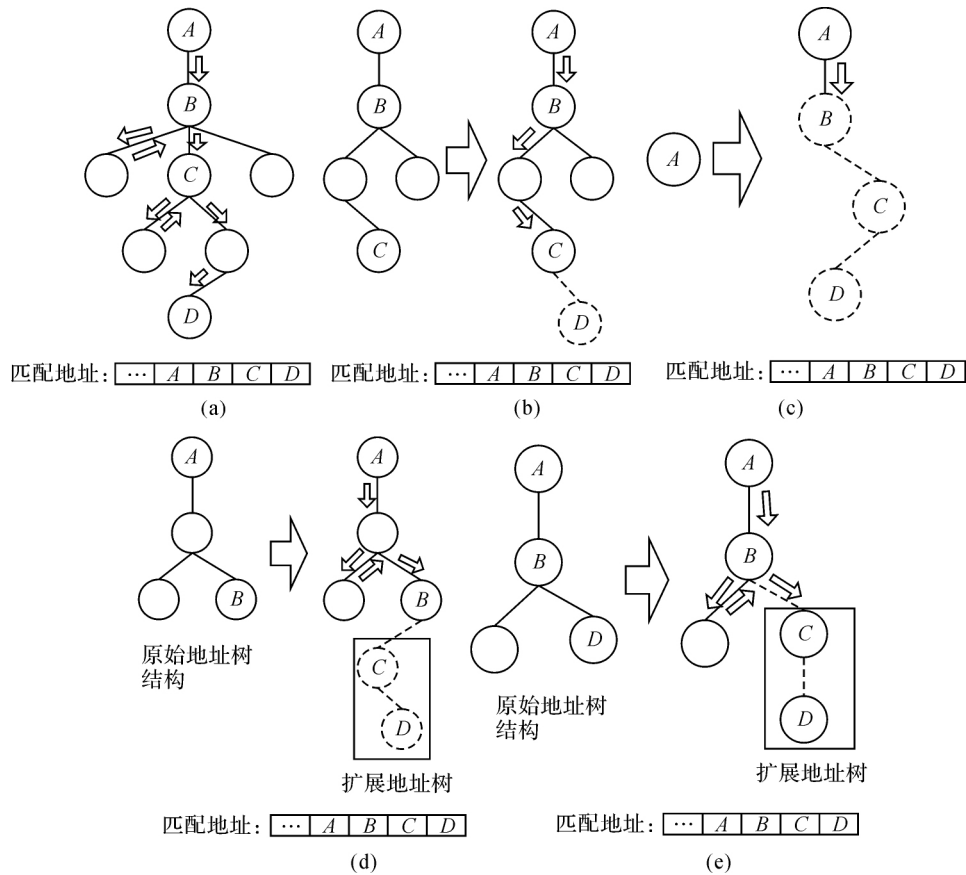


图 3 地址树中地址要素的匹配示意图

Fig.3 Matching of address elements in address tree

### 2.3 顾及位置关系的参考样本选取

POI 的地理位置与其地址描述具有强烈的关联关系,因此,待处理 POI 与标准化 POI 之间的位置关系对提升地址标准化效果具有重要参考价值。根据对地址标准化的影响程度,本文重点考

虑欧氏距离、从属同一区域、从属同一线状要素和从属同一点状要素等 4 类位置关系。假定  $P_1$  为待处理的 POI,  $P_2$  为地址信息已经标准化的 POI,  $P_i xq$ 、 $P_i y$  分别代表  $P_i$  点的地理坐标,则 4 种位置关系(图 4)的定义及计算方法如下:

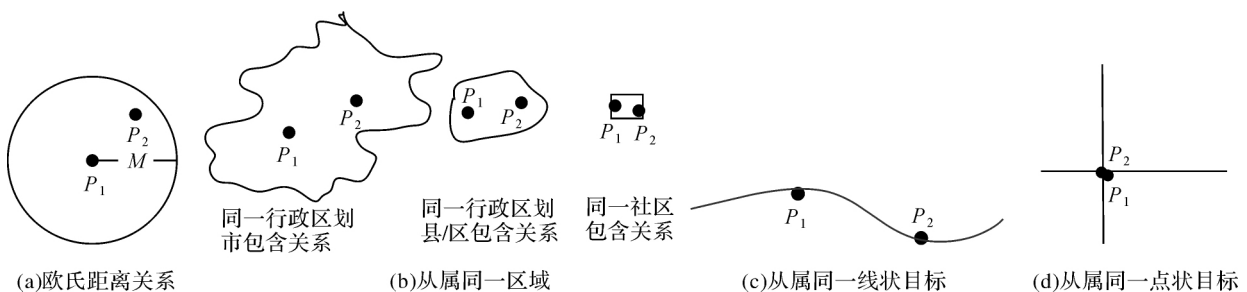


图 4 4 种位置关系示意图

Fig.4 Four types of positional relations

欧氏距离:以 POI 之间的直线距离来表示, 其计算公式为

$$\text{Dist}(P_1, P_2) = \sqrt{(P_1x - P_2x)^2 + (P_1y - P_2y)^2}$$
(1)

欧氏距离一般只用于 POI 点比较稀少且路网、居民地较为稀少的农村或边远地区,主要作为一种弱空间相关的参考 POI 样本选取据。在地址标准化参考样本选取时,可以设定一个距离阈值  $N$ ,当标准化 POI 与待处理 POI 的距离大于阈值  $N$  时,将不作为地址标准化处理的样本。对我国县级行政区的面积进行统计发现最小面积为  $56 \text{ km}^2$ ,本文以面积相当的圆反算对应半径,因此将距离阈值设置为  $N=4.2 \text{ km}$ 。

从属同一区域:表示两个 POI 点处于同一个面状地理对象范围内,即被同一个面状地理对象包含,如同一行政区划市包含关系、同一行政区划区包含关系、同一社区包含关系等。

$$\text{Area}(A_i, P_m, P_n) = \text{PtInArea}(P_m, A_i) \& \text{PtInArea}(P_n, A_i)$$
(2)

式中,  $\text{Area}(A_i, P_m, P_n)$  表示点  $P_m, P_n$  同时被面对象  $A_i$  包含范围;  $\text{PtInArea}$  用于判断某点  $P$  是否被面对象  $A$  包含,计算公式如下

$$\text{PtInArea}(P, \text{Area}) = \{ \forall \text{Area}[i], \text{Area}[j] | (P_x - \text{Area}[i]x) * (\text{Area}[j]y - \text{Area}[i]y) - (\text{Area}[j]x - \text{Area}[i]x) * (P_y - \text{Area}[i]y) < 0 \}$$
(3)

式(3)通过计算  $P$  与  $\text{Area}$  中任意两点

$\text{Area}[i], \text{Area}[j]$  的向量叉积是否小于 0,判断点  $P$  是否被面对象  $\text{Area}$  包含。

从属同一线状要素:表示两个 POI 点同处于某一个线性地理对象上,如相同道路附属关系、相同街道附属关系等

$$\text{Line}(L_i, P_m, P_n) = \text{PtOnLine}(P_m, L_i) \& \text{PtOnLine}(P_n, L_i)$$
(4)

式中,  $\text{Line}(L_i, P_m, P_n)$  表示点  $P, P_2$  同属于线对象  $L_i$ 。  $\text{PtOneLine}$  用于判断是否位于某个线对象上,计算公式如下

$$\text{PtOnLine}(P, \text{Line}) = \{ \exists \text{Line}[i] | \text{Dist}(P, \text{Line}[i]) = 0 \}$$
(5)



式中,  $\text{Line}[i]$  为构成  $\text{Line}$  的任一线段;  $\text{Dist}(P, \text{Line}[i])$  表示点与线段的欧氏距离。

从属同一点状要素:表示两个 POI 点处于同一点状对象或同一地理位置,如属于同一座大厦、位于同一个单元号、或位于同一个地理坐标

$$\text{Dist}(P_1, P_2) = 0 \parallel \text{Dist}(P_1, P_2) < M$$
(6)

式中,  $\text{Dist}(P_1, P_2)$  表示点  $P_1, P_2$  的欧氏距离;  $M$  为实际计算中判断为共点关系的阈值。在地址标准化处理中,可作为参考 POI 的一般为相对固定的点状交通管线要素(如公交站、电线杆)和地标物(如大厦、广场等)。因此本文在重点参考城市道路、建筑设计等相关规范(详见表 2)的基础上,设定阈值  $M=3.5 \text{ m}$ 。

表 2 共点距离阈值  $M$  设置的主要参考依据  
Tab.2 References for assignment of threshold  $M$

点状要素类型	常见形态	参考依据	共点距离阈值 $M$
公交站、门牌号、电线杆……		该类点状要素,距离最近的为往返公交站点。根据《城市道路交通规划设计规范》(GB50220—95)规定,单条机动车道的宽度为 3.5~5.0 m。	3.5 m
大厦、广场、发射塔……		根据《建筑设计防火规范》(GB50016—2014)规定,建筑物之间需要设置消防通道,且其宽度应该不小于 4.0 m。	

2.4 地址要素填充与标准化处理

根据位置关系对地址标准化的影响程度,给出如下强弱关系排序为:共点关系>共线关系>从属同一区域关系>欧氏距离关系。在给定一个具有标准化地址信息的 POI 数据集后,可以为某个待处理的 POI 计算出对应于 4 种位置关系的参考 POI 集合,分别为  $M_{pt}$  (满足共点关系的集合)、 $M_{ln}$  (满足共线关系的集合)、 $M_{ar}$  (从属同一区域关系的集合)、 $M_d$  (符合欧氏距离阈值条件的集合)。依次从 4 个数据集中选取标准地址作为参考,对待处理 POI 地址信息中的“缺位”的地址

要素进行自动填充,并使地址信息标准化尽可能达到地址要素的最小粒度。具体分为 3 种情况:

(1) 基于共点匹配的地址标准化:当  $N_{pt} > 0$  时,可以根据参考 POI 的地址信息实现门牌级地址要素标准化。首先通过文本相似度计算,筛选出与待处理地址具有最大文本相似度的 POI 作为候选;若候选 POI 个数大于 1,则取距离最近的作为标准化依据。后续的地址标准化处理流程为:以待处理 POI 地址的最小粒度要素为起点,逐层追溯参考 POI 的上级地址要素直至最顶层,然后将各级要素的名称顺序串联起来。

(2) 基于共线/共面关系的地址标准化: 当  $N_{pt}=0 \& (N_{ln}>0 \parallel N_{ar}>0)$  时, 门牌级地址要素匹配失败, 但可以根据共线或共面位置关系匹配到关联 POI。在这两种情况下, 可以回溯到门牌级地址要素的上一级, 再根据上一级地址要素与地址树的匹配情况进行处理: ① 如果该要素的上一级地址要素匹配成功, 则找出所有以该上一级地址要素为父节点的地址要素, 并依次与当前地址要素进行相似度计算, 选取相似度最高的地址要素作为地址标准化的参考节点, 然后再从该参考地址要素为起点, 逐层追溯其所有的上级地址要素, 直至地址树的最顶层, 从而实现中文地址的标准化处理; ② 如果该要素的上一级地址要素仍然匹配失败, 则依次循环, 继续回溯到更上一级的地址要素进行匹配, 直到匹配成功, 最终完成地址标准化处理。

(3) 基于欧氏距离的地址标准化: 当  $N_{pt}=0 \& N_{ln}=0 \& N_{ar}=0$  即不存在与该 POI 共点/共线/共面的参考 POI 资源时, 可以通过欧氏距离计算来选择参考 POI。根据  $M_d$  中 POI 对象的地址信息, 利用文本相似度进行匹配。如果匹配成功, 则以该参考地址要素为起点进行地址标准化处理; 如果失败, 则不以该 POI 地址作为标准化参考。

### 3 试验与分析

#### 3.1 算法试验

本文以北京市为例, 选取 4 家互联网地图商的 POI 数据进行试验, 以其中 2 家互联网地图商的地址数据作为基础匹配资源库, 另外 2 家地址数据作为待处理的测试数据。测试中, 基础匹配库分别设置了 3 万和 6 万两个级别的数据量, 待处理测试数据的数量分别为 5000、8000、10 000、15 000、20 000、25 000, 测试结果如图 5 所示。其中,  $a_1$ 、 $b_1$  表示基础 POI 资源库有 3 万条地址数据时的匹配率曲线,  $a_2$ 、 $b_2$  则表示基础 POI 资源库数量增加至 6 万条时的匹配率曲线。

从图 5 可看出: ① 不同来源的地址数据标准化的正确率不完全相同, 其原因是由于不同来源的网络 POI 地址表达方式不尽相同, 地址表达相对规范或与某一地址模型更为接近的数据源, 其地址标准化正确率也相对高些; ② 随着基础 POI 资源库数量的增大, 尤其是能基本覆盖整个试验地区后, 地址标准化将获得更高的正确率, 可达 90% 左右。

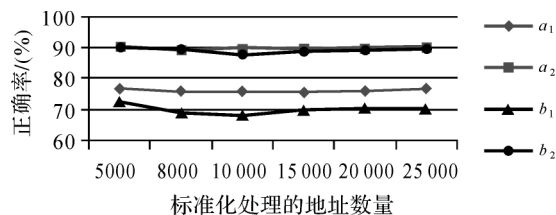


图 5 中文地址标准化试验结果对比

Fig.5 Comparison of experimental result for address standardization

#### 3.2 算法讨论

本文提出的地址标准化方法, 其处理效果与参考 POI 数据的丰富程度密切相关, 因为参考数据越多, 为待标准化地址的各级要素匹配到对应参考对象的几率就越大, 从而使缺失的地址要素得以补全、较粗粒度的地址信息也得以提升到更细粒度。在网络数据环境中, 由于地图服务提供的 POI 数量庞大且地址信息较为规范, 使得本文提出的基于位置关系 POI 地址标准化方法具有相当的可行性。

以北京市为例, 当基础参考信息为行政区划数据即北京市所辖各区时, 待标准化处理的 POI 数据如表 3 所示, 其标准化匹配遍历及结果如图 10。主要存在两种情况:

(1) 成功标准化:  $POI_4$ – $POI_8$  的地址标准化处理可以以  $POI_1$ – $POI_3$  的地址要素为参考样本, 同时也能自动修正与填充  $POI_1$ – $POI_3$  中地址缺失的地址要素。

(2) 标准化失败: 对于  $POI_{10}$  地址为“北京西绒线胡同 33 号”, 由于地址要素“西城区”与“西绒线胡同”在其他 POI 中从未出现, 导致该条 POI 地址标准化失败。

#### 3.3 与传统方法比较

基于规则匹配和纯文本机器学习等传统中文地址信息标准化处理方法<sup>[11,20]</sup>, 都聚焦在对“地址文本”进行分析处理, 而对因地理坐标派生的“位置关系”及其参考资源考虑较少。此外, 基于规则匹配的方法由于需要人工构建地址信息的规则库和专家库, 较为耗时耗力, 通用性较差, 地址标准化效果受规则库质量的影响较大; 纯文本机器学习方法多采用半监督学习方法, 具有较高的通用性, 可以获得较高的准确率。

表 3 POI 地址标准化匹配示例数据

Tab.3 Sample data for poi address standardization

名称	原始地址文本	切分的地址要素	地址要素匹配路径	参考 POI	标准化地址
POI <sub>1</sub>	海淀区北蜂窝路 6 号	海淀区   北蜂窝路   6 号	中国→北京市→海淀区→北蜂窝路→6 号	—	中国北京市海淀区北蜂窝路 6 号
POI <sub>2</sub>	海淀区莲花池西路 1 号	海淀区   莲花池西路   1 号	中国→北京市→海淀区→莲花池西路→1 号	—	中国北京市海淀区莲花池西路 1 号
POI <sub>3</sub>	海淀翠微路 19 号	海淀   翠微路   19 号	中国→北京市→海淀区→翠微路→9 号	—	中国北京市海淀区翠微路 19 号
POI <sub>4</sub>	北蜂窝路 62 号	海淀区   北蜂窝路   62 号	中国→北京市→海淀区→北蜂窝路→62 号	POI <sub>1</sub>	中国北京市海淀区北蜂窝路 62 号
POI <sub>5</sub>	莲花池西路 15 号	莲花池西路   15 号	中国→北京市→海淀区→莲花池西路→15 号	POI <sub>2</sub>	中国北京市海淀区莲花池西路 15 号
POI <sub>6</sub>	北京市北蜂窝中路 6 号	北蜂窝中路   6 号	中国→北京市→北蜂窝中路→6 号	—	中国北京市海淀区北蜂窝中路 6 号
POI <sub>7</sub>	翠微路凯德 Mall 大厦	翠微路   凯德 Mall 大厦	中国→北京市→海淀区→翠微路→凯德 Mall 大厦	POI <sub>3</sub>	中国北京市翠微路凯德 Mall 大厦
POI <sub>8</sub>	海淀区北蜂窝中路 15 号院	海淀区   北蜂窝中路   15 号院	中国→北京市→海淀区→北蜂窝中路→6 号	POI <sub>6</sub>	中国北京市海淀区北蜂窝中路 15 号院
POI <sub>9</sub>	海淀区羊坊店路 6 号	海淀区   羊坊店   6 号	中国→北京市→海淀区→羊坊店→62 号	—	中国北京市羊坊店 6 号
POI <sub>10</sub>	北京西绒线胡同 33 号	北京   西绒线胡同   33 号	中国→北京市→西绒线胡同→33 号	—	中国北京市西绒线胡同 33 号

与之相比,本文提出的顾及位置关系的地址信息标准化处理方法考虑了待处理 POI 与参考资源库的位置关系,充分利用网络 POI 数量庞大、样本丰富的优势,在有效克服地址要素缺失、标准化粒度较粗等问题的同时达到较高准确率(表 4);无监督学习方式也使得该方法具有较高通用性,可以很好地解决我国大部分城市的 POI 地址标准化问题。但在偏远地区,由于受参考 POI 样本数量限制,标准化效果与纯文本机器学习方法相当。

表 4 本文方法与传统地址标准化方法的比较

Tab.4 Comparison between methods of this paper and others

地址标准化方法	考虑位置因素	学习方式	通用性	匹配准确率
本文方法	是	无监督	高	高(>90%)
纯文本机器学习方法	否	半监督	较高	较高(<85%)
基于规则匹配的方法	否	不支持	差	受规则数量和质量影响大

4 结 论

本文提出一种顾及空间位置关系的网络 POI 地址信息标准化处理方法,该方法基于可扩展中文地址树模型,首先在对 POI 地址信息进行要素

切分和匹配,其次通过基于地理坐标衍生的 4 种位置关系从标准参考库中抽取参考对象库,最后根据共点、共线、共面等不同情况完成待处理 POI 地址信息的细粒度要素匹配和缺失要素填充。与传统地址标准化方法相比,该方法充分利用了 POI 的坐标信息及其衍生位置关系,能够明显改善机器学习、规则匹配方法等传统方法训练和归纳成本较大、耗时耗力等问题,尤其在具有大量参考 POI 样本资源的互联网数据环境中具有更好的适用性和更高的准确率。目前本方法使用的位置关系较为简单,相关阈值设定也主要为经验取值,在后续工作中将考虑增加更多的位置关系(如通达性),并就相关阈值设置进行更多的讨论,以使筛选出的候选目标对 POI 地址标准化具有更好的参考价值。

参考文献:

[1] GOLDBERG D W, WILSON J P, KNOBLOCK C A. From Text to Geographic Coordinates: The Current State of Geocoding[J].URISA Journal, 2007,19(1): 33-46.

[2] 黄颂. 中文地址编码技术的研究[D]. 北京: 北京大学, 2005.

HUANG Song. Research on Chinese Address Coding Technology[D]. Beijing: Beijing University, 2005.

[3] 陈细谦, 迟忠先, 金妮. 城市地理编码系统应用与研究[J]. 计算机工程, 2004, 30(23): 50-52.

CHEN Xiqian, CHI Zhongxian, JIN Ni. Application and

- Study of City Geocoding System[J]. Computer Engineering, 2004, 30(23): 50-52.
- [4] 江洲, 李琦, 王凌云. 空间信息融合与地理编码数据库的开发[J]. 计算机工程, 2004, 30(5): 1-2, 153.  
JIANG Zhou, LI Qi, WANG Lingyun. Geospatial Information Fusion and Implementation of Geocoding Database [J]. Computer Engineering, 2004, 30(5): 1-2, 153.
- [5] 李琦, 罗志清, 郝力, 等. 基于不规则网格的城市管理网格体系与地理编码[J]. 武汉大学学报(信息科学版), 2005, 30(5): 408-411.  
LI Qi, LUO Zhiqing, HAO Li, et al. Research on Urban Grid System and Geocodes[J]. Geomatics and Information Science of Wuhan University, 2005, 30(5): 408-411.
- [6] 程承旗, 关丽. 基于地图分幅拓展的全球剖分模型及其地址编码研究[J]. 测绘学报, 2010, 39(3): 295-302.  
CHENG Chengqi, GUAN Li. The Global Subdivision Grid Based on Extended Mapping Division and Its Address Coding [J]. Acta Geodaetica et Cartographica Sinica, 2010, 39(3): 295-302.
- [7] ZANDBERGEN P A. A Comparison of Address Point, Parcel and Street Geocoding Techniques[J]. Computers, Environment and Urban Systems, 2008, 32(3): 214-232.
- [8] 薛明, 肖学年. 关于地理编码几个问题的思考[J]. 北京测绘, 2007(2): 54-56.  
XUE Ming, XIAO Xuenian. Considering on Some Questions of Geocoding[J]. Beijing Surveying and Mapping, 2007 (2): 54-56.
- [9] 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.  
ZHANG Yifeng, WU Jianping, CHENG Yi, et al. The Improvement of Geocoding in ArcGIS[J]. Geomatics & Spatial Information Technology, 2007, 30(3): 116-119.
- [10] 朱前飞. MapInfo 中的地理编码及应用[J]. 四川测绘, 2001, 24(3): 117-119.  
ZHU Qianfei. Geocode and Its Application in MapInfo[J]. Surveying and Mapping of Sichuan, 2001, 24(3): 117-119.
- [11] GU Bin, JIN Yanfeng, ZHANG Chang. Study on the Standardized Method of Chinese Addresses Based on Expert System[C]//Proceedings of the IEEE 2nd International Conference on Cloud Computing and Intelligent Systems (CCIS). Hangzhou: IEEE, 2012: 1254-1258.
- [12] KOTHARI G, FARUQUIE T A, SUBRAMANIAM L V, et al. Transfer of Supervision for Improved Address Standardization[C]//Proceedings of the 20th International Conference on Pattern Recognition (ICPR). Istanbul: IEEE, 2010: 2178-2181.
- [13] CHEN Liyan, FANG Yuan. The Design and Research of Standard Address Database System Based on WebGIS in Panyu, Guangzhou[C]//Proceedings of 2008 International Seminar on Business and Information Management. Wuhan: IEEE, 2008: 233-235.
- [14] AUTHORITY T V. Address Data Content Standard Public Review Draft[S]. [S.l.]: Subcommittee on Cultural and Demographic Data, Federal Geographic Data Committee, 2003.
- [15] 高红, 黄德根, 杨元生. 汉语自动分词中中文地名识别[J]. 大连理工大学学报, 2006, 46(4): 576-581.  
GAO Hong, HUANG Degen, YANG Yuansheng. Chinese Place Names Recognition for Chinese Automatic Segmentation[J]. Journal of Dalian University of Technology, 2006, 46(4): 576-581.
- [16] 张春菊, 张雪英, 吉蕾静, 等. 地名通名与地理要素类型的关系映射[J]. 武汉大学学报(信息科学版), 2011, 36(7): 857-861.  
ZHANG Chunju, ZHANG Xueying, JI Leijing, et al. Relation Mapping between Generic Terms of Place Names and Geographical Feature Types[J]. Geomatics and Information Science of Wuhan University, 2011, 36 (7): 857-861.
- [17] 唐旭日, 陈小荷, 张雪英. 中文文本的地名解析方法研究[J]. 武汉大学学报(信息科学版), 2010, 35(8): 930-935, 982.  
TANG Xuri, CHEN Xiaohe, ZHANG Xueying. Research on Toponym Resolution in Chinese Text[J]. Geomatics and Information Science of Wuhan University, 2010, 35 (8): 930-935, 982.
- [18] BOURLAND F J, WALDEN S C, BAKER C A. Rich Browser-based Interface for Address Standardization and Geocoding: US, 20080065605[P]. 2008-03-13.
- [19] MASREK M N, RAZAK Z A. Malaysian Address Semantic: The Process of Standardization[C]//Proceedings of the 2nd International Conference on Computer Research and Development. Kuala Lumpur: IEEE, 2010: 77-80.
- [20] KALEEM A, GHORI K M, KHANZADA Z, et al. Address Standardization Using Supervised Machine Learning[C]//Proceedings of 2011 International Conference on Computer Communication and Management. Singapore: IACSIT Press, 2011, 5: 441-445.
- [21] 亢孟军, 杜清运, 王明军. 地址树模型的中文地址提取方法[J]. 测绘学报, 2015, 44(1): 99-107. DOI: 10.11947/j. AGCS.2015.20130205.  
KANG Mengjun, DU Qingyun, WANG Mingjun. A New Method of Chinese Address Extraction Based on Address Tree Model[J]. Acta Geodaetica et Cartographica Sinica, 2015, 44 (1): 99-107. DOI: 10. 11947/j. AGCS. 2015.20130205.
- [22] GUO Honglei, ZHU Huijia, GUO Zhili, et al. Address Standardization with Latent Semantic Association[C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009: 1155-1164.

(责任编辑: 宋启凡)

收稿日期: 2015-12-08

修回日期: 2016-03-22

第一作者简介: 王勇(1976—), 男, 副研究员, 研究方向为网络地理信息获取与挖掘。

First author: WANG Yong(1976—), male, associate professor, majors in retrieving and mining of Web geospatial information.

E-mail: wangyong@casm.ac.cn