引文格式: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.(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

地址树模型的中文地址提取方法

亢孟军,杜清运,王明军
武汉大学资源与环境科学学院,湖北 武汉 430079

A New Method of Chinese Address Extraction Based on Address Tree Model

KANG Mengjun,DU Qingyun,WANG Mingjun
School of Resources and Environmental Science,Wuhan University, Wuhan 430079,China

Abstract: Address is a spatial location encoding method of individual geographical area. In China, address planning is relatively backward due to the rapid development of the city, resulting in the presence of large number of non-standard address. The space constrain relationship of standard address model is analyzed in this paper and a new method of standard address extraction based on the tree model is proposed, which regards topological relationship as consistent criteria of space constraints. With this method, standard address can be extracted and errors can be excluded from non-standard address. Results indicate that higher math rate can be obtained with this method.

Key words: standard address; geocoding; address tree model; Chinese address extraction; address math rate
Foundation support: The National Natural Science Foundation of China(No.41201403)

摘 要:地址是一种对个体地域空间位置信息的编码方法。在我国,由于城市快速发展,地址规划相对落后,非标准地址大量存在。本文在分析标准地址模型空间约束关系类型的基础上,提出了一种基于地址树模型的中文地址提取方法。该模型以拓扑关系作为空间约束关系是否一致的判断标准,可以从非标准地址中提取标准地址,并剔除非标准和错误地址元素。试验证明,该方法有较高的地址匹配率。

关键词:标准地址;地理编码;地址树;中文地址提取;地址匹配率
中图分类号:P208 **文献标识码**:A **文章编号**:1001-1595(2015)01-0099-09
基金项目:国家自然科学基金(41201403)

1 引 言

地址是一种采用自然语言组织描述个体地域空间位置的抽象的编码方法[1-2]。通过解析地址获取地理坐标是当前获取空间信息简单有效的手段[3]。这种方法称为地理编码(Geocoding),是指按照一定的规则赋予个体地域唯一、可识别的编码,建立个体地域与标准地址、空间坐标的映射关系,从而可将地址与空间坐标进行自动转换[3-5]。地理编码由 4 部分组成:输入数据、输出数据、处理算法和参考数据库[6]。地理编码服务由最初的专业应用已经完全融入普通公众的生活工作中。

地址匹配是地理编码的核心,主要有 3 种方法:①为待匹配地址分配一个地址单元(address parcel),例如网格单元;②基于点状地址模型(address point)的地址查询;③基于路网模型(street network),通过线性内插为门牌分配坐标[5,7]。欧美国家一般先进行城市总体规划,再建立详细的地址模型,地址数据规范,地理编码的难度和工作量小,地址匹配主要基于上述 3 种方法。而在我国,由于地址规划落后于城市建设,地址标准混乱以及中文地址表达的随意性,给地理编码工作带来了极大的难度和工作量,必须通过算法上的优化来解决[8-12]。

地址模型是地址匹配、地理编码的核心,决定地址编码的算法和地址匹配的质量。本文将从以下 4 部分进行论述:①讨论地址的概念,对常见的中文地址模型进行总结;②提出基于地址树模型的标准地址提取方法;③提出标准地址可靠性的评价方法;④通过地址匹配试验对本文提出的方

法进行验证。

2 地址的概念

地址是一种抽象的编码方法，通过自然语言组织描述个体地域的空间位置。地址是地址元素的集合，可表示为

$$A=\{x_i\in A\mid P(x_i,x_j)\neq\varnothing,x_i\neq x_j\} \quad (1)$$

式中，A 表示地址；x_i 表示地址元素；$P(x_i,x_j)$ 表示地址元素之间存在的空间约束关系，该约束不为空。

地址元素通常为地名，例如门牌号、街道名称、街道类型和邮政编码。狭义的地名指具有指位性和社会性的个体地域的指称[13]，广义地名指地理实体的指称。如图 1，地理实体有 3 个重要的性质：①地理实体“是什么”，涉及地理实体的语义、分类体系、空间关系等[14-15]；②地理实体“叫什么”，涉及地理实体的规范命名等[16-19]；③地理实体“在哪里”，涉及地理实体的空间位置描述和表达。地理坐标是空间位置的重要表达方式，但在导航应用中，空间坐标并不能提供给用户足够的空间指位功能，用户面临“最后 20 m”的难题，即根据导航数据只能找到目的地附近的位置，最后 20 m 的距离只能靠别的方式确定。地址是通过自然语言的编码方式表达地理实体的空间位置，它符合人的空间认知特点，便于进行位置的表达和交流。

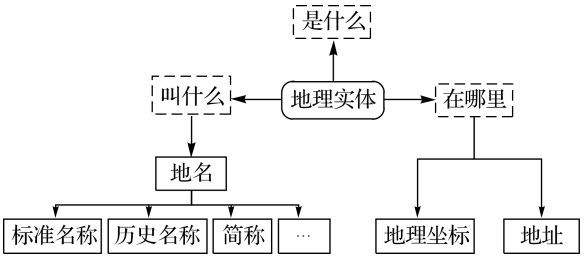


图 1 地理实体、地名、地址关系

Fig. 1 Relationship of geographical entity, place name and address

国内外城市规划部门都把地址作为城市规划的重要组成部分，产生了多种有特色的地址模型。如美国常用的地址模型包括以下几种地址元素[20]：门牌号；前缀方向、前缀类型、街道名称；街道类型，后缀方向；城市、州和邮政编码等信息，如图 2(a)所示。纽约皇后区规划东西走向的为路(avenue)，南北走向的为街(street)，同时记录道路交叉口信息，如图 2(c)。盐湖城的地址模型以“后期圣徒”教堂作为参考中心，为道路分配相对于教堂的距离和位置编码，如图 2(d)。伊利诺伊州地址模型把网格区域作为重要的地址元素，辅助确定门牌号的准确位置，如图 2(e)。德国一般地址模型同于美国，但地址元素的排列方式略有不同，如图 2(b)。

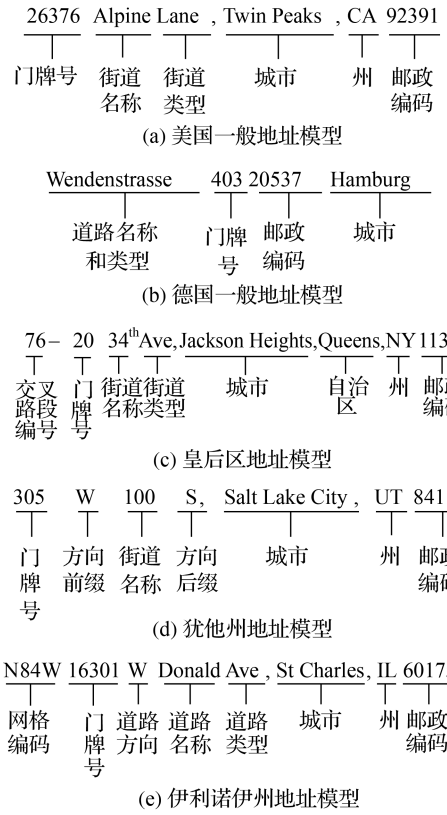


图 2 国外常用地址模型

Fig.2 Common foreign address models

以上地址模型的共同点：①重视邮政编码的指位功能，通过邮政编码即可定位到一定的空间区域；②门牌号作为最基本的地址元素，是地址最详细的位置指定元素，门牌号在西方文化中的重要性已经超越了规划的意义，例如“唐宁街 10 号”比“首相官邸”实体更有名气；③重视路网在宏观上的指位功能，描述更为详细的道路信息，例如通过道路后缀表示其走向、级别等；④地址元素的稳定性较高，所谓稳定性是指在一定时间段内的变化频率；⑤重视规划、超前规划、尊重规划，例如盐湖城的规划始于 19 世纪，皇后区的规划始于 20 世纪 20 年代，并且后期的地址编码都采用前期的规划原则。

在我国，地址模型尚未在城市规划中得到足够的重视，以门牌的管理为例，如图 3 所示，“武汉

大学信息学部”所在的地址为“武汉市洪山区珞喻路 129 号”，该地址模型为“市|区|道路|门牌号”，其中，“市|区”部分由民政部地名办公室管理，“道路”由武汉市规划部门管理，而“门牌号”由公安部门管理。目前全国到各省市，尚无一个统一的协调机构[21]，这种突出的矛盾已经无法满足城市规划和信息化发展的需求。

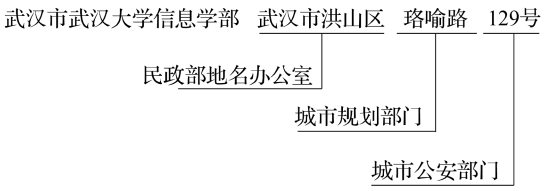


图 3　我国的地名管理体系示例

Fig.3　Sample of Chinese place names management system

中文地址模型研究已经得到越来越多学者关注[22-24]，在《深圳市地名总体规划》中，已经对深圳市的地址模型进行了全面的规范。表 1 通过分析深圳市部分地址，提取了几种常用的地址模型，可以得到以下 4 个结论：

表 1　深圳市地址模型示例

Tab.1　Address model samples of Shenzhen

地址	地址模型	备注
深圳市罗湖区泥岗东路 1118 号	行政区划＋道路＋门牌号	基本地址模型，行政区划不详细
深圳市罗湖区塘坑仔村 143 号	行政区划＋片区（自然村）＋门牌号	基本地址模型，行政区划不详细
深圳市罗湖区坭岗北村 22 栋	行政区划＋片区（自然村）＋楼栋号	楼栋号替代门牌，加大空间认知难度
深圳市罗湖区玉龙村 C 区 10 栋	行政区划＋片区（分区）＋楼栋号	对片区增加人为分区，加大空间认知难度
深圳市罗湖区笋西片区宝岗路嘉田大厦	行政区划＋片区＋道路＋公共设施名	分区、道路并存，加大空间认知难度
深圳市罗湖区红岗路 1005 号红岗西村小区	行政区＋道路＋门牌号＋公共设施	基本模型，增加地名描述，地址信息冗余

①行政区划在地址模型中作用重要，作为主要的空间区域约束元素；②由于缺乏唯一、标准的地址表述，在描述地址时，人们总是提供尽可能多地描述信息，从而导致地址描述的信息冗余，这种冗余亦可能导致地址歧义；③在有明确门牌号的情况下，人们愿意选择门牌描述地址，但是由于门牌规划、标示不充分，导致用户不得不选择公共设施、单位名等稳定性低的地名作为地址描述；④新旧城区无统一标准，部分城中村编码方案混乱，有采用门牌号的，也有采用楼栋号的。

地址模型是地址标准化的核心，也是实施地理编码的核心。地址模型的确立需要有完善的规划方案作为前提，同时要兼顾用户的空间认知习惯，以引导为主，逐步推进地址规范化的有效实施。而针对目前的非标准地址大量存在的现实，有效的地址提取算法是唯一解决办法。

3　地址树模型及标准地址提取

非标准地址要与空间坐标进行转换，要经过地址解析和标准化的过程，如图 4 所示。首先经过地址分词，形成可识别的地址元素集合，这里的词库是收录具有词汇意义的地名词典(Gazetteer)[25-26]。由于存在地名重名，因此需要消解地名歧义，构建符合空间约束关系的地名元素集合；经过定歧义消解，地址元素的空间语义较为明确，形成子地址集合；任一子地址根据其地址元素的类别，可明确该子地址的详细指位含义。此时，可以直接进行地址标准化或地址匹配操作，如图 4。

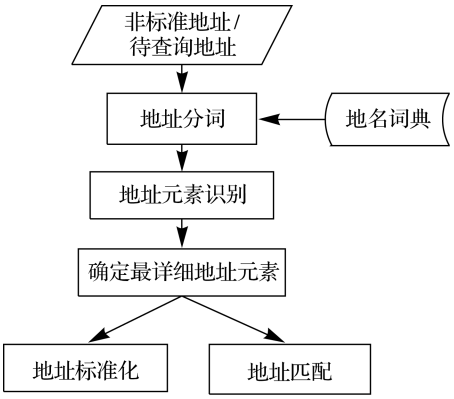


图 4　标准地址提取流程

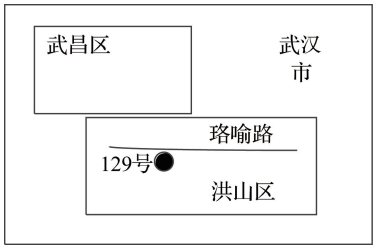
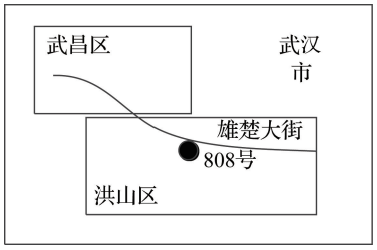
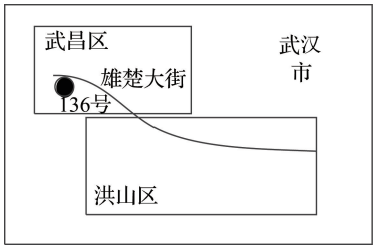
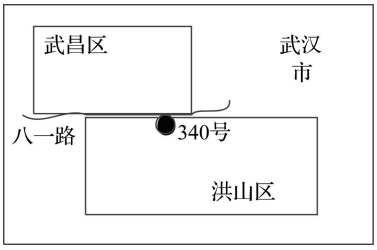
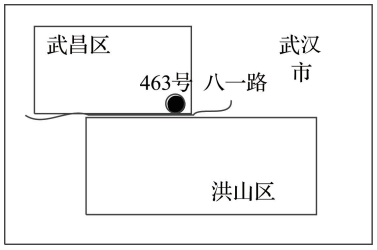
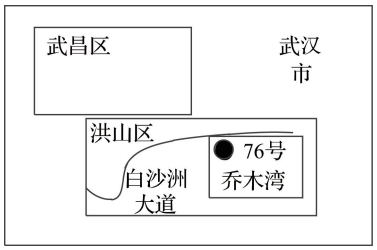
Fig.4　Process of standard address extraction

3.1　地址模型空间约束关系描述

组成标准地址的地址元素之间需要具有空间约束关系，如式(1)中的 $P(x_i,x_j)\neq\varnothing, x_i\neq x_j$ 可用拓扑关系表示这种约束，具体的拓扑关系类型要根据地址元素的几何类型确定，一般要保证地址元素间的包含或关联关系[27-29]。本文分别以“行政区划|道路|门牌号|公共设施”(street network model)和“行政区划|片区|门牌号|公共设施”(address parcel model)两种常用地址模型说明(表 2)。

表 2 地址元素空间约束关系的九交模型表达

Tab.2 The 9-intersection model expression on space constraint relationship of address elements

序号	关系示例	地址	九交模型	模型
1	武汉市洪山区 珞喻路 129 号		$\begin{matrix} & B^0 & \partial B & B^- \\ A^0 & -\varnothing & -\varnothing & -\varnothing \\ \partial A & \varnothing & \varnothing & -\varnothing \\ A^- & \varnothing & \varnothing & -\varnothing \end{matrix}$	路网模型
2	武汉市洪山区 雄楚大街 808 号		$\begin{matrix} & B^0 & \partial B & B^- \\ A^0 & -\varnothing & -\varnothing & -\varnothing \\ \partial A & -\varnothing & \varnothing & -\varnothing \\ A^- & -\varnothing & -\varnothing & -\varnothing \end{matrix}$	路网模型
3	武汉市洪山区 雄楚大街 136 号		$\begin{matrix} & B^0 & \partial B & B^- \\ A^0 & -\varnothing & -\varnothing & -\varnothing \\ \partial A & -\varnothing & \varnothing & -\varnothing \\ A^- & -\varnothing & -\varnothing & -\varnothing \end{matrix}$	路网模型
4	武汉市洪山区 八一路 340 号		$\begin{matrix} & B^0 & \partial B & B^- \\ A^0 & \varnothing & \varnothing & -\varnothing \\ \partial A & -\varnothing & -\varnothing & -\varnothing \\ A^- & \varnothing & \varnothing & -\varnothing \end{matrix}$ $\begin{matrix} & B^0 & \partial B & B^- \\ A^0 & \varnothing & \varnothing & -\varnothing \\ \partial A & -\varnothing & -\varnothing & -\varnothing \\ A^- & -\varnothing & -\varnothing & -\varnothing \end{matrix}$	路网模型
5	武汉市武昌区 八一路 463 号		$\begin{matrix} & B^0 & \partial B & B^- \\ A^0 & \varnothing & \varnothing & -\varnothing \\ \partial A & -\varnothing & -\varnothing & -\varnothing \\ A^- & \varnothing & \varnothing & -\varnothing \end{matrix}$ $\begin{matrix} & B^0 & \partial B & B^- \\ A^0 & \varnothing & \varnothing & -\varnothing \\ \partial A & -\varnothing & -\varnothing & -\varnothing \\ A^- & -\varnothing & -\varnothing & -\varnothing \end{matrix}$	路网模型
6	武汉市洪山区 乔木湾 76 号		$\begin{matrix} & B^0 & \partial B & B^- \\ A^0 & -\varnothing & -\varnothing & -\varnothing \\ \partial A & \varnothing & \varnothing & -\varnothing \\ A^- & \varnothing & \varnothing & -\varnothing \end{matrix}$	分区模型

路网模型(street network model)是约束关系最复杂的一种模型，道路是地址信息的主要载体，行政区划与道路关系主要有 3 种：包含、关联和相交，表 2 中，例 1 是最常见的地址模型。中文地址的组织，往往从高级别行政区划开始，以空间上的包含关系来逐步限定地址表述目标，这种特点比较符合点状模型或者分区模型，但也被应用于路网模型。多数道路也适合这种“包含于”行政区划的特征，但是，道路经常作为行政区划的分界，或者出现跨越行政区划的现象，如表 2 中的示例 5，这时，地址元素的层次关系表达不代表其“包含”的空间关系，只代表其空间上的关联关系，这种组织方式可以明确路段信息，使地址指向更加明确。

门牌号与道路是拓扑关联关系，总体上沿道路按照线性特征分布。通过对部分城市门牌数据的分析发现，绝大多数门牌分布在道路 400 m 以内，部分区域由于路网稀疏、居民点密集、门牌呈聚集状分布。

分区模型(address parcel model)是以居住区为单位的面状区域地址元素，例如城中村、社区分区或工业区等，如表 2 中的示例 6。这类地址元素一般“包含于”行政区划，同时分区也包含一定的门牌号或楼栋号，这种空间约束不同于道路门牌的线状关系，一般呈面状聚集特征，因此这类匹配一般把分区的行政中心或几何中心作为结果返回。

表 2 的示例 4、5 中，“八一路”作为武昌区和洪山区的行政区划边界，地址描述根据门牌具体所属的行政区划组织，从而出现了同一道路门牌，行政区划的限定地址元素不同的现象。这种地址组织方式说明了拓扑包含这种空间关系在人们进行空间认知和表达中的重要性，同时，增强了地址的指向性。

3.2　地址模型的错误空间约束关系

标准的地址模型是指地址描述中包含完整的行政区划信息、详细地址元素，并且指向性明确。但在实际基础地理信息普查或地址应用中，非标准地址或错误地址大量出现，严重影响了地址匹配的精度。非标准地址或错误地址主要有以下 4 种情况：

(1) 行政区划地址元素不完善，但整条地址指向性明确。这类地址在实际应用中出现较多，属于非标准地址，需要标准化。

(2) 行政区划地址元素不完善，整条地址指向性不明确。例如肯德基或银行类公共设施，在一定行政区划内分布数量较多，需要补充附加描述信息。地址匹配时，可提供该类公共设施结果集或上一级地址元素作为查询结果。

(3) 地址元素空间约束级别倒置、混乱。这类地址由于书写的随意性，或对其空间位置的不确定性，将地址元素错误排列并增加其他相关位置描述信息。在地址匹配过程中，需要对地址元素识别并重建其空间约束关系，同时过滤关联关系弱的描述信息，是地址匹配需要重点解决的一类错误情况。

(4) 地址元素空间约束关系错误。这类错误较多出现在基础地理信息普查过程中，地址元素子集的空间指向性和整条地址是分离、不相关的。地址匹配时，需要识别地址的真实指向，并剔除错误地址元素，是地址匹配的难点。

实际应用的地址多是以上几种情况的混合，增加了地址匹配的难度和工作量。本文提出一种地址树模型，通过地址元素的识别、空间约束关系的重构、地址原始指向的识别、错误地址元素剔除以及地址冗余信息的过滤，提取标准地址，提高地址匹配的准确性。

3.3　地址树模型及提取算法

定义 1：地址是地址元素的集合，也是子地址集合。一个地址描述可能具有多个指向目标，可表示为

$$A=\{x_i\in A\mid P(x_i,y_i)\neq\varnothing,x_i\neq x_j\}=\{A_i\mid A_i\neq\varnothing,0\leqslant i\leqslant n\}\tag{2}$$

式中，A_i 表示地址 A 的一个指向目标。

定义 2：每个地址元素对应 n 个地址语义，地址语义指地址元素实际指向的个体地域目标，对应于实际的同名不同址问题。可表示为

$$x_i=\{s_i\in S\mid s_i\neq\varnothing\quad 0\leqslant i\leqslant n\}\tag{3}$$

式中，x_i 表示任意地址元素；S 表示地址元素 x_i 的语义集合；s_i 表示 x_i 的任意地址语义。

定义 3：语义级别指按照地址元素类型的分级信息，行政区划级别高，详细地址元素级别低，语义 s_i 的语义级别表示为 AddrLevel(s_i)。

如图 5 所示，地址提取的过程是在地址元素的语义集合中，寻找一条符合空间约束关系的连通路径，每条子地址可看作地址描述的一个子树，这种特点适合用树模型进行地址解析。

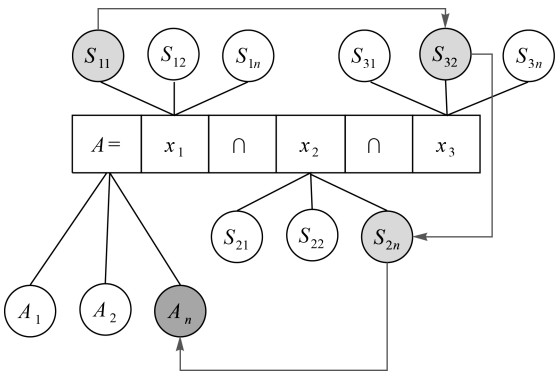


图 5 地址、地址元素、地址语义关系

Fig.5 Relationship among address, address elements and address semantic

基于地址树模型的标准地址提取算法如图 6 所示，具体步骤描述如下：

(1) 假设地址字符串已经分词、识别，生成地址元素集合 X 以及地址元素语义集合 S。

(2) 建立根节点 root，提取地址元素 x_1，遍历 x_1 的语义集 S_1，构建地址语义节点，并依次连结到根节点。

(3) 提取后续地址元素 x_i，遍历其语义节点 S_i。对于节点 S_{i1}，依次与当前地址树的叶子节点 l_i 比较。首先比较其语义级别，若 S_{i1} 语义级别低于 l_i，则比较 S_{i1} 与 l_i 的空间约束关系一致性，若空间约束关系一致，则 S_{i1} 连结到当前叶子节点 l_i。

若不一致，则沿该子树上溯，直到找到该子树的结点 l_i'，满足 l_i' 语义大于 S_{i1}。此时比较两节点的空间约束一致性，若不一致，则比较 S_{i1} 与地址树的下一叶子节点，重复步骤(3)；若一致，比较 S_{i1} 与 l_i' 后一结点的空间约束关系，若一致，则把 S_{i1} 插入该子树当前位置，若不一致，则比较 S_{i1} 与地址树的下一叶子节点，重复步骤(3)。

若 S_{i1} 上溯到根节点，仍未连结，则把该节点连结到地址树的最右边，作为一条新的子树。

(4) 对于同一地址元素，若 $AddrLevel(s_i)\neq AddrLevel(s_j)(i\neq j)$，并且 s_j 已经成为地址树的叶子节点，则跳过该叶子节点。

空间约束关系一致是指地址元素拓扑关系符合表 2 的约束规则。具体实施可采用两种方法：①实时计算地址元素拓扑关系，该方法运算量大、响应时间长，但反映拓扑关系准确；②对地址元素预处理，通过一定地理编码方案，记录地址元素的拓扑关系。一般的编码方案只记录拓扑“包含”，对于路网并不适用，可扩展编码方法，对路网的“关联”关系进行适当记录。该方法运算量小、响应时间短，但地址元素变化后，需要更新编码以维护其空间关系。对于实际应用而言，主要地址元素，如行政区划等，其稳定性较高，采用第 2 种方法更方便。

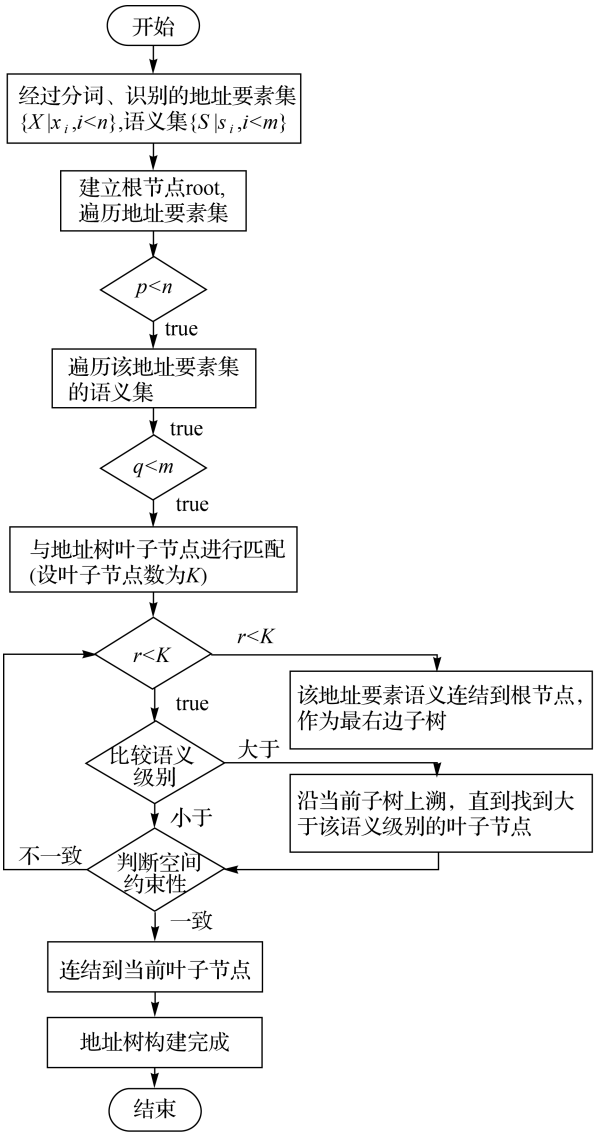


图 6 基于地址树模型的标准地址提取算法流程

Fig.6 Process of standard address extraction algorithm based on address tree model

非标准地址标准化后由两个大部分构成，其组合原则如下：

〈标准地址〉::=〈行政辖区名称〉〈局部地址描述〉…〈局部地址描述〉

其中〈行政辖区名称〉为政区类地名，局部地址描述可以是多级的，如：深圳市福田区福田街道

口岸社区福田南路 34 号皇岗海关生活区新 2 栋，为 2 级局部地址型地址。

3.4　匹配地址集的筛选和评价方法

非标准地址经过地址模型提取后，得到子地址集合，需要对子地址集合按照空间指向相关系进行评价、排序，得到最符合原始地址指向目标的标准地址。评价方法主要从以下 3 个指标进行评价：有意义比例、完整度和基于地址元素权重的评分。

有意义比例，是指可识别的地址元素占所有地址元素的比例，反映当前子地址的可靠性，若有意义比例过低，则该子地址的指向目标是不可靠的；完整度，是子地址树的深度和地址元素集合数目的比。完整度为 1 的子地址与原始地址的指向性完全一致；基于地址元素权重的评分方法，假设一个地址严格按照标准地址模型组织，每个地址元素对应一个严格的位置，若子地址中某地址元素偏离其标准位置越远，则得分越低，反之，则得分越高，再结合该地址元素的权重，可计算子地址打分。

在实际应用中，可先设置有意义比例的阈值，高于阈值的子地址集合比较其完整度，完整度小于 1 的子地址集合，再计算其评分，按该流程对子地址排序，获取最符合原始地址的指向目标的子地址。

地址匹配度是指原始地址的匹配地址与其目标地址的契合程度，经过上述处理，可以得到地址库中最契合原始地址描述的候选地址。地址匹配度可表达为

$$D=\frac{M_S}{O_S}=\frac{s(t_1+t_2+t_3+\cdots+t_n)}{t_1+x_1+t_2+t_3+x_2+\cdots+t_n+x_i}\approx\frac{s_1+s_2+\cdots+s_n}{t_1+x_1+t_2+t_3+x_2+\cdots+t_n+x_i}=\frac{s_n}{t_1+x_1+t_2+t_3+x_2+\cdots+t_n+x_i}$$

式中，D 为地址匹配度；M_S 表示原始地址的匹配地址；O_S 表示原始地址的目标地址，即地址本身指向的空间位置；M 是原始地址分词结果集的空间语义集，表示为 $s(t_1+t_2+t_3+\cdots+t_n)$，也可表示为 $s_1+s_2+\cdots+s_n$，该空间语义集按照拓扑关系约束等价于 s_n，即由最详细、地址语义级别最低的要素表示该地址的指向。原始地址经过分词可拆分一组可识别和不可识别的地址要素集，其中 t 表示可识别的要素，x 表示不可识别的要素，$t_1+x_1+t_2+t_3+x_2+\cdots+t_n+x_i$ 为该拆分结果的表示，未知元素 x 可分布在地址描述的任意位置。为简化地址匹配度计算，假设地址描述不存在乱序现象，即地址描述按照行政区划级别从高到底排列。

量化地址匹配度，一般采用向量空间模型（vector space model，VSM），传统 VSM 的分词项被假设为彼此相互独立，权重由词频决定，这种方法未考虑地址要素的空间约束关系。本文采用改进的 VSM，分词项权重设置为 $w=\mathrm{e}^i$，其中 e 为自然对数的底数，i 为分词项在原始地址分词集合中的顺序数，对各分词项权重进行归一化处理：$w_i=\frac{\mathrm{e}^i(1-\mathrm{e})}{\mathrm{e}(1-\mathrm{e}^n)}$，$n$ 为分词集合的数量，此时 w_i 的值域为 $(0,1]$。此权重的设置可以保证两点：①原始地址描述中顺序靠后的地址元素有更高的权重；② $\sum_{i=1}^{n-1}\mathrm{e}^i<\mathrm{e}^n$，即前 $n-1$ 项的权重和小于第 n 项的权重，确保顺序靠后的地址要素具有足够的权重。

则地址匹配度表达为：$D=\cos\theta=(M\cdot O)/(\|M\|\|O\|)$，其中，$M$ 为原始地址的匹配地址要素的权重矢量；O 为原始地址描述的分词权重矢量。

4　试验结果及分析

本文以深圳市 2012 年地址编码库为参考，选取深圳市建筑物普查 377 条数据作为试验数据（多数为标准地址），原始数据包含建筑物的面状空间信息和地址描述。其中，主要存在两种非标准地址：①一个建筑物对应多个门牌，例如分布在道路交叉口、呈 L 形的建筑物；②多个建筑物属于同一个门牌，例如面街的大型建筑物。这两种情况会影响到地址匹配精度。

提取建筑物的地址，利用自主开发的 TeleG-Coder 软件进行地址匹配，生成建筑物所属地址的点状要素数据和地址匹配统计信息，匹配情况如图 7 所示。

表 3 为建筑物地址匹配率统计结果，试验结果显示地址匹配度 100％的条目占到总条目的 94.96％，这类地址的特征为：①描述相对规范，无别字错字；②行政区划完善，详细地址部分描述准确；③符合基本的地址模型规则，对地址匹配的干扰较小。

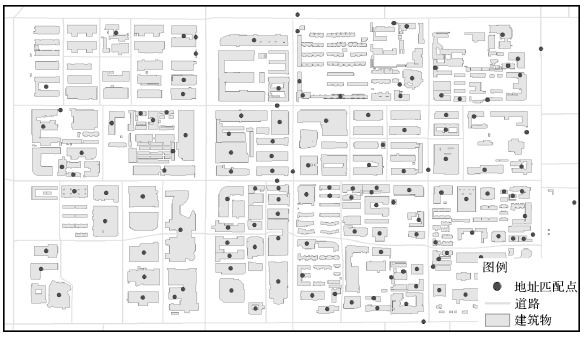


图 7 建筑物所在地址匹配结果

Fig.7 The match result of building addresses

表 3 建筑物地址匹配率统计结果

Tab.3 Statistics of building address match rate

匹配度区间	匹配条目	匹配率/(%)
0～50%	3	0.8
50%～60%	0	0
60%～70%	4	1.1
70%～80%	0	0
80%～90%	12	3.18
90%～100%	0	0
100%	358	94.96
总条目	377	

表 4 列举了几种低匹配度地址情况，主要有以下原因：①地址参考库的数据不完善；②地址含有非标准的公共设施名称；③采用相对位置关系、方位关系描述地址；④采用路口描述地址。因此，在地址参考库不变的情况下，规范地址描述，可以极大地提高地址匹配率。

表 4 低匹配度地址匹配结果

Tab.4 Samples of low match degree address

建筑物地址	匹配地址	匹配度/(%)
福田区上步中路北侧	深圳市福田区上步中路	85
福田区深南中路华联大厦旁边	广东省深圳市深南中路	85
福田区振兴路体委宿舍 2 栋	深圳市福田区振兴路	85
福田区燕南路外国语学校 1 号平房	深圳市福田区燕南路	80
福田区深南中路华联商业街地铁进出口	广东省深圳市深南中路	80
福田区深南中路与上步路交汇处	福田区深南中路	60
福田区税务局宿舍 2 栋旁	深圳市福田区	60

5 结 论

地址匹配是建立专题数据与空间信息关联的有效手段，由于缺乏完善的城市地址规划，人们对规范地址的认知度不高，地址表达不规范，导致地址匹配率不高。本文在论述地址概念的基础上，提出了几种标准地址模型的组织方式，并用九交模型对其拓扑关系进行归类。鉴于非标准地址大量出现在基础地理信息普查数据中，本文提出了一种基于地址树模型的标准地址提取方法，以地址元素的空间约束关系为条件，提取标准子地址集合并剔除非标准子地址或错误地址元素。试验结果表明，该方法可以获得较高的地址匹配率。由于非标准地址类型较多，要进一步提高地址匹配率需要在地址规范普及和算法两个方面进行更多的研究。

参考文献：

[1] ZHANG Xueying, ZHU Shaonan, ZHANG Chunju. Annotation of Geographical Named Entities in Chinese Text[J]. Acta Geodaetica et Cartographic Sinica, 2012,41(1):115-120.(张雪英，朱少楠，张春菊. 中文文本的地理命名实体标注 [J]. 测绘学报，2012，41(1)：115-120.)

[2] PALKOWSKY B, METACARTA I. A New Application Information Discovery:Geography Really Does Matter [C]// Proceedings of the SPE Annuual Technical Conference and Exhibition. Dallas:[s.n.],2005.

[3] ROONGPIBOONSOPIT D, KARIMI H A. Comparative Evaluation and Analysis of Online Geocoding Services [J]. International Geographical Information Science, 2010, 24(7): 1081-1100.

[4] ZHANG Xueying, LÜ Guonian, LI Boqiu, et al. Rule-based Approach to Semantic Resolution of Chinese Addresses [J]. Geoinfomation Science, 2010, 12(1): 9-17.(张雪英，闾国年，李伯秋，等. 基于规则的中文地址要素解析方法 [J]. 地球信息科学学报，2010，12(1)：9-17.)

[5] ZANDBERGEN P A. A Comparison of Address Point, Parcel and Street Geocoding Techniques [J]. Computers, Environment and Urban Systems, 2008, 32(3): 214-232.

[6] 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.

[7] RUSHTON G, ARMSTRONG M P, GITTLER J, et al. Geocoding in Cancer Research: A Review [J]. American Journal of Preventive Medicine, 2006, 30(2): S16-S24.

[8] ZHU Jianwei, WANG Zemin. The Principle of Geocodifying and Its Solution on Localization[J]. Beijing Surveying and Mapping, 2004(2):24-27.(朱建伟，王泽民. 地理编码原理及其本地化解决方案 [J]. 北京测绘，2004(2):24-27.)

[9] WANG Xiuming. Address Automatic Matching of Geographic Information System[J]. Journal of Minxi Vocational and Technical College, 2007, 9(2): 75-77.(王秀明. 地理信息系统地址自动匹配 [J]. 闽西职业技术学院学报，2007，9(2)：75-77.)

[10] HU Qing, XU Jianhua, WANG Zhihai. Study on the Method

of Address Automatically Matching in GIS Database[J]. Geomatics and Spatial Information Technology, 2008, 31(6): 50-52.(胡青, 徐建华, 王志海. GIS 数据库中地址自动匹配方法研究 [J]. 测绘与空间地理信息, 2008, 31(6): 50-52.)

[11] SUN Yafu, CHEN Wenbin. Address Matching Technology Based on Word Segmentation[C]// Proceedings of China Association of Geographic Information Systems Fourth Congress of the 11th Annual Meeting. Beijing: [s. n.], 2007: 114-125.(孙亚夫, 陈文斌. 基于分词的地址匹配技术 [C]//中国地理信息系统协会第四次会员代表大会暨第十一届年会论文集.北京:[s.n.], 2007: 114-125.)

[12] HUANG Song. Research on Chinese Address Coding Technology[D].Beijing:Beijing University,2005.(黄颂. 中文地址编码技术的研究 [D].北京:北京大学,2005.)

[13] CHU Yaping, YIN Junke, SUN Donghu. The Toponymy Basis Tutorial[M]. Beijing:Sinomap Press,1994.(褚亚平, 尹钧科, 孙冬虎. 地名学基础教程 [M]. 北京:中国地图出版社, 1994.)

[14] ZHANG Xueying, ZHANG Chuju, LÜ Guonian. Design and Analysis of a Classification Scheme of Geographical Named Entities[J]. Geoinformation Science, 2010, 12(2): 220-227.(张雪英, 张春菊, 闾国年. 地理命名实体分类体系的设计与应用分析 [J]. 地球信息科学, 2010, 12(2): 220-227.)

[15] CHEN Jianjun, ZHOU Chenhu, WANG Jinggui. Advances in the Study of the Geo-ontology [J]. Earth Science Frontiers, 2006, 13(3): 81-90.(陈建军, 周成虎, 王敬贵. 地理本体的研究进展与分析 [J]. 地学前缘, 2006, 13(3): 81-90.

[16] CHU Yaping. The City Names Commercialization of Geographic Names Legalization[J]. Chinese Toponym, 1996(1): 4-6.(褚亚平. 城市地名商品化与地名管理法制化 [J]. 中国地名, 1996(1): 4-6.)

[17] CHU Yaping. Urban Planning and Development Can not Ignore the Toponym Planning [J]. Beijing Planning Review, 2004(6): 112-113.(褚亚平. 城市规划发展不能忽略地名规划 [J]. 北京规划建设, 2004(6): 112-113.)

[18] QIN Xuexiu. Three Forms of Placename Data and Their Demand[J]. Bulletin of Surveying and Mapping, 2011(10): 68-69.(秦学秀. 地名数据的 3 种形式及其质量要求 [J]. 测绘通报, 2011(10): 68-69.)

[19] ZHANG Li. Analysis of Chinese Signposts Language Usage [J]. Lanzhou Academic Journal,2007(3): 206-208.(张黎. 我国地名标志语言文字使用状况分析 [J]. 兰州学刊, 2007(3): 206-208.)

[20] ESRI. ArcGIS Resource [EB/OL].[2013-07-12].http://help.arcgis.com/zh-cn/arcgisdesk-top.

[21] ZHAO Guozhou. To Talk about the Doorplate Reform[J]. Research and Exploration, 1998, 2(1):34-36.(赵国洲. 谈谈门牌改革 [J]. 决策探索, 1998, 2(1):34-36.)

[22] GUO Xiaolin. Discussion on the Management of Doorplate in City[J]. Shandong Economic Strategy Research, 2008(3): 61-62.(郭晓琳. 略论城市建设中的楼门牌设置与管理 [J]. 山东经济战略研究, 2008(3): 61-62.)

[23] LI Qimin. The Social Function of the City Doorplate[J]. Construction Science and Technology, 2002(2): 46-47.(李启明."城市门牌"的社会功能 [J]. 建设科技, 2002(2): 46-47.)

[24] LI Yongheng. The Integration Process of Macau Geography Information Data: Taking Street Door Number Data as an Example[J]. Geomatics World, 2013, (1):87-91.(李永恒. 澳门地理数据的整合进程——以街道门牌数据为例 [J]. 地理信息世界, 2013, (1):87-91.)

[25] HILL L, FREW J, ZHENG Q. Geographic Names: The Implementation of a Gazetteer in a Georeferenced Digital Library [J/OL].[2013-07-13].http://dblp.uni-trier.de/db/journals/dlib.

[26] HILL L L. Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints [M]. Berlin: Springer, 2000: 280-290.

[27] EGENHOFER M J, HERRING J. A Mathematical Framework for the Definition of Topological Relationships [C]// Proceedings of the Fourth International Symposium on Spatial Data Handling, Zurich:[s.n.], 1990: 803-813.

[28] EGENHOFER M J, HERRING J. Categorizing Binary Topological Relations between Regions, Lines, and Points in Geographic Databases [R]. Orono: University of Marine,1999.

[29] DENG Min, LIU Wenbao, FENG Xuezhi. A Generic Model Describing Topological Relations among Area Objects in GIS[J]. Acta Geodaetica et Cartographica Sinica, 2005, 34(1): 85-90.(邓敏, 刘文宝, 冯学智. GIS 面目标间拓扑关系的形式化模型 [J]. 测绘学报, 2005, 34(1): 85-90.)

(责任编辑:陈品馨)

收稿日期: 2014-01-08

修回日期: 2014-10-05

第一作者简介: 亢孟军(1983—),男,讲师,主要研究方向为电子地图、地理编码。

First author: KANG Mengjun(1983—), male, lecturer, majors in digital maps and geocoding.

E-mail: mengjunk@gmail.com