一种基于综合分词和模糊匹配的地名地址匹配方法研究

周 浩，葛江涛
（天津市测绘院，天津 300381）

摘要：通过对国内地址匹配现状的全面研究，并结合对原始地址数据的分析，总结出了我国常用地址表述的习惯和规律，抽象出了地址匹配层级模型和地址匹配类型，用于地址匹配的各个过程。在此基础上，提出了基于先验知识的中文分词算法和基于规则的模糊匹配方法，并利用工商注册数据库为原始地址数据文件，建立了工商职能部门企业地址数据与空间数据的地址匹配映射。

关键词：地址匹配；层级模型；模糊匹配；先验规则
中图分类号：P208 **文献标识码**：A **文章编号**：1673-1131(2015)05-0010-01

随着我国城市建设脚步的加快，原有的地名地址已经无法与当前城市实际情况相适应，因此，对地名地址进行更新已经成为了城市发展中的一项重要工作。但就目前相关部门所采取的地名地址更新方法来看，并不十分完善，无法满足城市建设发展的根本需求。为此，对地名地址更新方法进行完善已经成为了城建部门的一项重要工作。

随着建设数字城市的兴起和发展，特别是随着GPS、北斗导航定位系统的推广应用，地址数据已经成为人们表达他们感兴趣点位的重要方式。如何将没有坐标信息的属性数据与包含坐标信息的空间数据对应起来，建立他们之间的关联关系，是地名地址匹配需要处理和解决的问题。

1 地名地址匹配原理分析

实现地名地址匹配，需要满足以下约束条件，首先是空间参考坐标系统，用于确定匹配的数据赋予何种空间坐标；其次是地址匹配层级模型，它是地名表述的抽象数据模型，是进行地址拆分和地址标准化的基础；第三是地址匹配规则，即标准化的地址数据与地理编码数据库进行匹配时的验证规则。以上约束条件是地址匹配的必备条件，是保证地址匹配准确率和精度的重要手段。

1.1 地址匹配层级模型

根据对地名地址信息的分析，总结抽象出了地址匹配层级模型。包括三个层级，依次是面状行政区级别、面状或线状地址级别、点状子地址级别。根据层级模型，标准地址的全写形式应该是：省（地级市）+市/县+道路/小区+门/楼牌/标志物/兴趣点。

（1）面状行政区级别：包括五层：国家层、省级层、地级市层、区县层、镇/乡层。

（2）面状或线状地址级别：由道路或小区组成，不允许为空。

（3）点状子地址级别：由门牌、楼牌、标志物和兴趣点组成，不允许为空。

1.2 地址匹配类型分析

由于我国地址表述的特殊性，在地址匹配中采用三种基本的地址匹配类型：街道定位，区域定位，街道+区域定位。衍生出五种基本地址结构：街道+门牌、小区+楼牌、街道+楼牌、街道+兴趣点、街道+小区+楼牌。地理编码是地名地址数据匹配及更新系统的一项重要内容，在进行编码之前，做好必要的准备工作是不容忽视的。目前，地理编码的准备条件主要包括两个步骤：一个是地址标准化，另一个是编码标准化。

2 关键技术研究

在地址匹配时，首先通过XML可扩展标记语言存储分词标志词库。通过字符串分词正向最大匹配算法根据 XML 行政区划词库进行原始地址拆分，对拆分结果划入地址匹配层级模型，并与行政区划地址库进行模糊验证，将结果写入模糊地址数组 A。剩余原始地址，按 XML 道路和小区词库进行拆分，分别与道路和小区地址库进行模糊验证，写入模糊地址数组 A，完成地址匹配层级模型中的第一级和第二级构建。对经过行政区划和道路小区拆分后的字符串。根据门牌、楼牌、兴趣点XML词库进行拆分，拆分时选用词库顺序根据地址匹配规则进行排序。对拆分的字符与门牌、楼牌、兴趣点地址库进行模糊验证。完成地址匹配层级模型第三级构建。在拆分门牌、楼牌、兴趣时，遍历模糊数组 A，并根据模糊匹配规则进行地址字段缺失和地址字段歧义的处理，直到字符串拆分完毕。完全匹配规则库的记录，查询对应属性地址库的空间坐标并返回给用户。模糊匹配规则库的记录，返回可能的地址全称供用户选择。无法匹配的记录，直接返回。如“奉化市广平路89号”，首先拆分出奉化市行政区一级，然后依次匹配辖区和道路，辖区库无“广平”字段记录，转而匹配道路，在数据库中找到“广平”一词，根据街道+门牌类型在找到匹配记录，反查广平路的上级编码为锦屏街道。重构后的标准地址就是“奉化市锦屏街道广平路 89 号”。又如“奉化市锦屏街道广平 2 号楼”，首先通过基于先验知识的中文分词算法切分出奉化市和锦屏街道两词，然后使用正向最大匹配算法对广平一词在道路和小区表中进行匹配，分别匹配到广平路和广平小区两词，剩余字段为 2 号楼。根据地址匹配类型中的街道+门牌和小区+楼牌两种类型，首先按街道+门牌对广平路 2 号楼进行假设性匹配，匹配失败，转由小区+楼牌的方式对广平小区 2 号楼进行假设性匹配，匹配成功。重构后的标准化地址是“奉化市锦屏街道广平小区 2 号楼”，赋予地址数据空间坐标，完成地址匹配。

3 地名地址匹配应用举例

本文选取工商注册数据库 2500 条实例数据作为实验对象，以地址普查成果为属性地址库进行地址匹配实验。完全匹配的记录 1934 个，模糊匹配的记录 256 个，无法匹配的记录 310 个，匹配成功率 87.6%。

参考文献：

[1] 林澍哲.简化分词的地址匹配技术[J].信息与电脑.2012(1):110-112
[2] 张林曼，吴升.地理编码系统中地址匹配引擎的设计与实现[J].测绘信息与工程,2008,33(6):13-16