一种基于规则的模糊中文地址分词匹配方法

程昌秀,于 滨
(中国科学院地理科学与资源研究所,北京 100101)

摘要:在研究分析地址模型的基础上,建立了存储标准地址数据集的标准地址库和自定义的地址匹配规则库,提出了一种基于规则的模糊中文地址编码方法。该方法在依据标准地址库分词的同时,也沿着自定义的地址匹配规则进行推理,从而缩小了下次分词所用到的目标数据集,提高了系统执行效率。另外,通过借助构建的规则树与歧义栈,提高了文中定义的两类模糊地址匹配的成功率。最后,基于该算法建立了一个地理编码原型系统,并利用经济普查项目中的相关数据对算法的可用性进行了验证。

关键词:地理编码;模糊地址;规则库;地址分词
中图分类号:TP391 **文献标识码**:A **文章编号**:1672-0504(2011)03-0026-04

0 引言

随着电子地图的应用与普及,许多行业都需要将大量自然语言描述的中文地址映射为地理坐标,并定位到电子地图上,从而使原有非空间数据获得空间坐标信息,实现各部门和各地理范围的数据整合与共享。地理(地址)编码就是一种把文本地址转换成地理坐标的技术[1]。

地理编码一般包括地址标准化、地址分词、数据库匹配、空间定位等步骤。其中,地址标准化是指通过更改格式和纠正拼写等方法对地址进行规范化处理;地址分词是指通过某种中文分词算法将地址分解成多个最小地址要素[2];数据库匹配是指在标准地址数据库中找到与该地址最接近的标准地址;空间定位则是根据这个标准地址的地理位置推理该地址的空间位置并定位。上述步骤是地理编码的核心内容,也是国内学者研究的重点。陈细谦等尝试在地址标准化阶段对数据进行数据清洗,通过总结地址错误模型和使用有穷自动机实现数据的规范化[3];孙亚夫等提出利用中文分词词典查找地址要素,利用最大正向匹配算法进行分词的同时查询判断地址要素的父地址[4];张铁燕等先使用最大逆向匹配算法进行分词,然后通过总结的几种地址模型进行地址匹配[5];张作华等尝试通过先分词再在数据库匹配时选择不同的地址公式以减少数据库的查询次数[6];Sengar 等针对地理编码中输入的不同数据格式进行了研究,并利用文本相似性与空间一致性实现了空间与非空间信息的相互转换[7];Goldberg 等初步探讨了地理编码过程中存在的误差、不确定性以及评测标准等常见问题[8];郭会提出了使用自动机对中文地址进行描述的方法,并研究提出了基于中文地址自动机的中文地址分词算法[9]等。但是,这些研究仍然存在着许多需要改进的地方和问题:

(1)地址标准化本身工作量非常繁重,并且很难穷举和定义出所有错误模型,故应设法提高对于非标准的模糊地址的匹配成功率,降低对地址标准化步骤的要求。其中,本文涉及的模糊地址主要包含两类:一类是指在数据库匹配时可能产生语义歧义的地址,本文将其定义为"第一类模糊地址";如"文慧园2号"中"文慧园"一词,匹配可能会出现"文慧园西路"、"文慧园东路"、"文慧园小区"等多种情况。另一类是指信息残缺地址,本文将其定义为"第二类模糊地址";如"清河中街69号"相对于标准地址"清河中街力度家园69号"缺失了"力度家园"这一地址要素。

(2)目前中文地址分词和数据库匹配主要基于字符串匹配的方法,但由于以往研究中地址分词与数据库匹配是两个不同的步骤,往往需要先对地址进行分词,再进行数据库匹配,这样往往导致数据查询和比较的次数过多,造成效率低下。因此,有必要探索一种高效的地址分词、匹配方法。

1 基于规则的模糊中文地址分词匹配算法

为解决上述3类问题,本文提出一种基于规则的

收稿日期:2010-12-22; **修订日期**:2011-03-03
基金项目:国家863项目"经济普查与基本单位统计遥感应用系统"(2006AA120106)、"地理空间数据库管理系统总体设计"(2007AA120401)
作者简介:程昌秀(1973-),女,博士,主要研究方向为空间数据库、GIS应用。E-mail:chengcx@lreis.ac.cn

模糊中文地址分词匹配算法。该算法的核心是在调用最大正向匹配算法进行地址分词的同时,在标准地址库中进行地址匹配。通过借助每次分词时对标准地址库的搜索,并实时参照地址匹配规则树,达到不断缩小目标数据集的目的,最终当满足规则库中某一规则时,终止算法,返回目标数据集,完成地址匹配。

1.1　中文地址与标准地址库

参考文献[2,10]对地址的理解,本文将中文地址分为两部分:1)行政区划:根据国家行政区划代码定义,行政区划可分为省、市、县、街道、居委会 5 个级别,其编码方式与码段意义如图 1 所示。图 1 中前六位中每两位代表一级,后六位中每三位代表一级,如果已知的行政区划部分描述不够完整,导致转化为代码后不足 12 位,则后面位数用 0 补齐。例如,北京市海淀区四季青镇对应的行政区划代码为 110108102000。2)街道地址:是由一系列的地址要素词组组合而成。根据类型不同,可细分为:道路、门牌号、住宅小区、楼牌号、突出建筑物(POI)5 种类型。另外,地址要素又由专名与通名共同构成[6],如地址要素"北海花园小区"中,"北海花园"为专名,"小区"为通名。针对不同类型地址要素的特点与规律,可分别归纳总结出几种常见通名(表 1)。

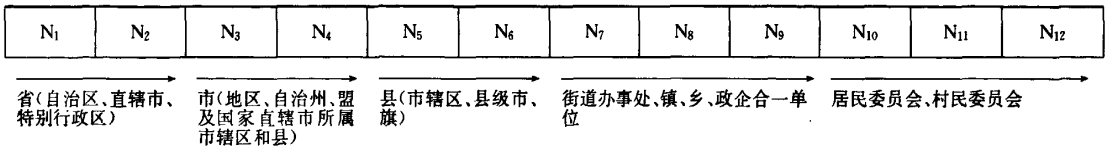

图 1　国家行政区划编码与码段含义
Fig. 1　Means of administrative divisions code

表 1　地址要素通名
Table 1　General names of address element

编号	类型	通名
1	道路	路/大道/道/大街/街/巷/胡同/条/里
2	门牌号	号
3	住宅小区	里/区/园/村/坊/庄/居/寓/苑/墅/小区/弄/公寓
4	楼牌号	号/号楼/楼/宿舍/斋/馆/堂
5	POI	大厦/广场/饭店/中心/大楼/场/广场/馆/酒店/宾馆/市场/花园/招待所

根据对地址结构的研究,本文设计了一个标准地址库结构,如表 2 所示。其中,第 0 项(ADCode)为地址的行政区划代码;第 1～5 项对应街道地址的通名,由于实际街道地址可能仅由上述几项构成(如道路+门牌号、小区+楼牌号、POI 等),故其可以部分字段为空;Geometry 字段为地址的空间坐标信息,该字段用于地址匹配完成后的空间定位。

表 2　标准地址库结构
Table 2　Field names of standard address database

编号	列名称	说明	类型	可否为空
0	ADCode	行政区划代码	文本	否
1	RoadName	道路名	文本	是
2	HouseID	门牌号名	文本	是
3	RCName	住宅小区名	文本	是
4	BuildingID	楼牌号	文本	是
5	POI	兴趣点	文本	是
6	Geometry	空间坐标	地理几何数据类型	否

1.2　基于规则的模糊中文地址分词匹配算法

1.2.1　匹配规则的定义　考虑到详细街道地址常用部分地址通名就能表达,为了提高效率,在地址匹配之前需要定义地址的匹配规则。表 3 给出了本实验中北京市地址匹配规则的定义。以表 3 中的第 1 条规则为例,若输入地址的道路名、住宅小区名、楼牌号与标准库中的相应字段相符则判定为匹配成功。若输入地址满足表 3 中规则一到规则六中的任意一条规则都视为匹配成功。

表 3　匹配规则
Table 3　Match rules list

规则序号	规则一	规则二	规则三	规则四	规则五	规则六
匹配条件	1,3,4	1,2	1,4	1,5	3,4	5

在基于规则的模糊中文地址分词匹配算法中,规则库与算法本身脱离,极大地提高了规则库的灵活性,方便用户对规则进行修改与顺序调整,从而可以根据地区和数据源的差异提高算法的运行效率。

1.2.2　匹配算法　对输入的中文地址,匹配算法是依据标准地址库进行分词处理,同时根据自定义的规则进行匹配推理,不断缩小目标数据集的范围,最终将匹配结果返回。若 R 为当前目标数据集、S 为输入的字符串、S1 为子串、S2 为当前的剩余字符串(S－S1)、Stack 为存储语义歧义的栈、Struct(i)用于存储歧义子串的多个变量、MatchField 为各个字段是否匹配的标记数组。匹配算法的具体流程如下:

第 1 步:读入地址字符串 S 和标准地址库,并将全部标准地址库记录作为目标数据集 R。

第 2 步:判断地址中是否存在行政区划部分。如果不存在,则转步骤 3;如果存在,则判断并分割出该子串 S1,然后在相应级别的《行政区划代码表》中查找 S1。根据查询结果,将所能查找到的最低级别行政区划的对应 12 位代码返回。在目标数据集 R 中,利用返回的 12 位代码对 ADCode 字段进行过滤,横向缩小目标数据集 R 的查询范围。

第 3 步:将字符串 S 赋值给子串 S1。

第 4 步：查询地址匹配规则树（图 2），根据规则树，限定下一步搜索时的备选字段范围，缩小匹配查询的纵向查找范围。

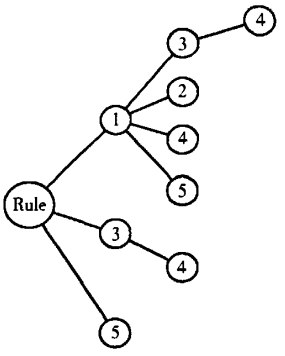

图 2　地址匹配规则树
Fig. 2　Rule-tree of address matching

第 5 步：判断 S1 是否为空，如果为空则转步骤 10；如果不为空，则调用最大正向匹配算法进行分词，在目标数据集 R 的指定字段中查找子串 S1。如果查询成功，则转步骤 6；查询失败，则继续调用最大正向匹配算法进行分词，直到 S1 为空，则转步骤 10。

第 6 步：判断查找到与子串匹配的字段个数：如果匹配字段数为 1，则在 MatchField 数组中标记该字段为匹配，并转步骤 7；如果匹配字段数大于 1，则说明产生语义歧义，转步骤 8。

第 7 步：查询规则库，比对标记数组 MatchField 与每条规则，查看是否有满足条件的规则存在，如果存在则将目标数据集 R 中的剩余记录返回，算法终止；如果不存在，则将查询到的子串 S1 在字符串中去掉，重新赋值为 S1＝S－S1，返回步骤 4。

第 8 步：子串产生歧义设置游标，指定与子串匹配的字段，并在 MatchField 数组中标记该字段。将此时的目标数据集 R、剩余子串 S－S1、标记数组 MatchField 3 个变量写入结构体 Struct(i)中，并把该结构体变量放入歧义栈 Stack 中。移动游标，重复该步骤，直至将所有歧义情况都作为结构体变量依次放入栈中。

第 9 步：取出位于栈顶的结构体 Struct(i)，并将结构体中各变量的值赋给相应变量，然后转步骤 7。

第 10 步：判断歧义栈 Stack 是否为空，如果为空则算法终止，查询失败；如果不为空，则转步骤 9。

以“北京市海淀区安宁庄 22 号楼”为例，首先，将该地址字符串 S 与整个标准地址库即目标数据集 R 作为参数传入。判断 S 中存在行政区划部分“北京市海淀区”，将其查询转换为相应的 12 位代码“110108000000”，并将目标数据集 R 中 ADCode 字段与该代码前六位“110108”不符的记录过滤掉。把剩余字符串“安宁庄 22 号”赋给 S1，通过查询地址匹配规则树确认备选匹配字段为 1、3、5。调用最大正向匹配算法，查询到“安宁庄”一词分别与 1(“安宁庄东路”)和 3(“安宁庄小区”)两个字段模糊匹配，因此产生语义歧义。将 1 字段与 3 字段先后入栈，然后取栈顶元素，先将“安宁庄”匹配到 3 字段“安宁庄小区”，当前数据集 R 缩小为 8 条记录，此时由于没有满足条件的规则，故对“22 号楼”继续进行分词匹配。第二次查询规则树，确定备选字段为 4，在 R 的 4 字段(楼牌号)中查询“22 号楼”，无匹配结果。故重新选取栈顶元素，将“安宁庄”匹配到 1 字段(“安宁庄东路”)，过滤后当前数据集 R 缩小为 11 条，查询规则库无满足的终止条件，故继续分词匹配。重新查询规则库并确定备选字段为 2、3、4。在 R 的备选字段中搜寻字符串“22 号楼”，发现 4 字段(楼牌号)有一条匹配记录，过滤后数据集 R 缩小为 1 条记录。此时继续查询规则库，发现有满足条件的规则，即规则三(1 和 4 字段匹配)。返回目标数据集 R 中的剩余记录，分词匹配算法成功终止。运算结果：对于模糊地址“北京市海淀区安宁庄 22 号楼”，在目标数据集 R 中找到了一条模糊匹配记录“安宁庄东路 15 号 22 号楼”。

1.3　讨论

与以往方法相比，该方法的改进如下：1)将地址分词、地址匹配两个环节整合，在依据标准地址库分词的同时，沿着自定义的地址匹配规则进行推理，从而缩小了下次分词所用到的目标数据集，提高了系统执行的效率。例如，根据表 3 定义的匹配规则，可以生成图 2 所示的地址匹配规则树，根据此规则树，在第一次搜索时，仅查出标准地址库中的 1、3、5 项即可；若第一次搜索在第 3 项字段中找到匹配的值，根据规则树，第二次搜索时，仅查出第 3、4 项字段，且其第 3 项字段等于第一次分词结果的数据集，从而可以从横、纵方向缩小目标数据集，加快分词与匹配速度。2)对于分词匹配时产生的语意歧义，算法中通过建立一棵歧义树并利用栈进行临时存储，然后依据深度优先原则，对歧义树进行遍历访问，直至满足匹配规则终止算法，否则继续完成整棵歧义树的遍历。从而解决了第一类模糊地址的分词匹配问题。3)利用树中的匹配规则，对于残缺地址要素的第二类模糊地址数据，也可以提高其匹配成功率。

2　应用实验分析

基于上述算法，本文利用 Visual Basic 6.0、

MySQL 以及 SuperMap Object 组件,开发了一个地理编码原型系统。为了验证算法的可用性,本文在经济普查项目中,选取了海淀区的 1 827 条未经标准化的企业地址数据,在项目中建立的北京市标准地址库中进行地址匹配,统计结果如表 4 所示。

表 4　实验结果统计
Table 4　Statistic of experiment results

地址类型	匹配结果		单项占比率(%)	总占比率(%)
成功匹配地址	完全匹配	1 292 条	70.7	83.6
(共 1 527 条)	模糊匹配	235 条	12.9	
无法匹配地址	过于模糊	291 条	15.9	16.4
(共 300 条)	错误地址	9 条	0.5	

根据实验结果,83.6%的匹配成功率说明该算法基本达到了经济普查项目中对于地址匹配的要求。图 3 是地理编码原型系统中对于地址"北京市海淀区安宁庄 22 号楼"的匹配与定位结果。

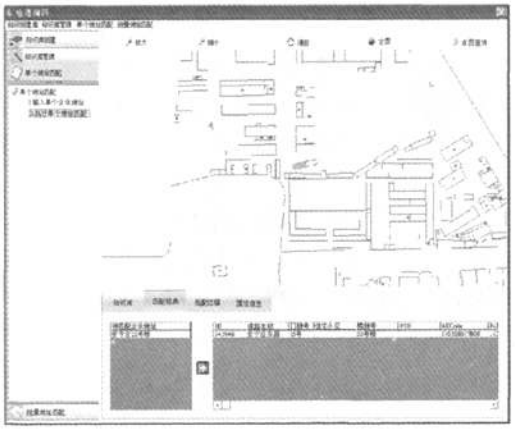

图 3　原型系统中地理编码结果演示
Fig. 3　A geocoding results in the prototype system

3　结语

本文提出的基于规则的模糊中文地址分词匹配算法,为地理编码项目中经常遇到两类模糊地址的匹配问题提供了新的解决思路,并在算法中尝试加入规则以及其他多种方法,提高了地址匹配的成功率与算法效率。在项目应用中,对于首次未能匹配的企业地址,可以尝试通过人机交互与知识学习不断扩充和完善标准地址库,进而达到不断提高匹配成功率的目的。另外,由于算法中对于分词匹配时每一步的目标数据集都进行了单独保存,如果利用每一步匹配的临时结果集并配合规则库中规则,可以进一步研究实现对备选地址记录设置匹配度或者进行评分,从而当匹配失败时,可以提供匹配度或评分较高的地址记录供用户选择。

参考文献:

[1] 王凌云,李琦,江洲. 国内地理编码数据库系统开发与研究[J]. 计算机工程与应用,2004(21):167－169.

[2] 李军,李琦,毛东军,等. 北京市地理编码数据库的研究[J]. 计算机工程与应用,2004(2):1－6.

[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.

[4] 孙亚夫,陈文斌. 基于分词的地址匹配技术[A]. 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[C]. 2007. 114－125.

[5] 张铁燕,翁敬农,黄坚. 城市地理编码方法的探索与实践[A]. 中国地理信息系统协会第九届年会论文集[C]. 2005.

[6] 张作华,孙凌宇. 基于城市地址编码技术的探讨[J]. 井冈山师范学院学报(自然科学),2005,26(3):42－25.

[7] SENGAR V, JOSHI T, JOY J, et al. Robust Location Search from Text Queries. http://research. microsoft. com/en-us/people/josephj/acm_gis_2007_robust_location_search. pdf, 2007－12－30.

[8] GOLDBERG D W, WILSON J P, KNOBLOCK C A. From text to geographic coordinates: The current state of geocoding[J]. Urban and Regional Information Systems Association, 2007, 19(1):33－46.

[9] 郭会. 基于自动机分词的中文地址地理编码技术研究与实现[D]. 中国科学院地理科学与资源研究所,2008.

[10] 高巍. 在大城市实现有线电视用户地址标准化的设想[J]. 广播与电视技术,2007(10):99－102.

A Rule-Based Segmenting and Matching Method for Fuzzy Chinese Addresses

CHENG Chang－xiu, YU Bin

(*Institute of Geographic Science and Natural Resources Research, CAS, Beijing 100101, China*)

Abstract: After analyzing Chinese address model, this paper built a standard address database and an address matching rules database, and then presented a rule-based Geocoding method for fuzzy Chinese addresses. This method used the standard address database to segment the input fuzzy Chinese address. At the same time, the method used the rules database to reduce and find a standard address that matched with that fuzzy address. The method used the customized rules to reduce candidate addresses so that it can participate in match reduction and save the matching executive time. In addition, the introduction of rule tree and semantic stacks also promote the matching of fuzzy address. Finally, a Geocoding prototype system was built, and then its availability was verified utilizing the data of natural economic census project.

Key words: Geocoding; fuzzy address; rule database; address segmentation