基于近似串匹配的地名数据库建设探析①

范立新, 黄龙军

(绍兴文理学院 计算机系, 绍兴 312000)

摘 要: 地址编码数据库在城市信息化建设中具有极其重要的地位. 本文从绍兴市地名实际出发, 就地址编码数据库的关键技术: 地址标准化和地址匹配及数据库设计等方面进行了探讨, 并重点介绍了多模式近似串匹配算法在地址匹配阶段的应用. 在地址标准化中, 讨论了规范化地址内容的表达形式, 把标准地址表达为行政区划名、主地址、子地址三部分构成, 建立层级地址数据模型和地址输入模型, 基于行政区划代码进行地址代码编制; 讨论了地址标准化的过程, 给出了标准化示例. 最后还给出了近似串匹配算法在地址匹配阶段应用的伪代码.

关键词: 近似串匹配; 地址编码; 地址标准化; 地址匹配; 地址内容

Construction of Geocoding Database Based on Approximate String Matching

FAN Li-Xin, HUANG Long-Jun

(Department of Computer Science, Shaoxing University, Shaoxing 312000, China)

Abstract: The geocoding database is important to informatization construction of city. This paper discusses the address standardization, address matching technology and design of address database based on the actual situation of address in Shaoxing. This paper mainly introduces the multiple approximate string matching algorithm in application of the address matching phase. About address standardization, this paper discusses the expression form of standardized address content, including administrative area, main address and subaddress, the hierarchical address data model and input model, the mode of geocoding based on administrative area code, the process and sample of address standardization. Finally, this paper shows the pseudo code of approximate string matching algorithm in the address matching phase.

Keywords: approximate string matching; geocoding; address standardization; address matching; address content

1 引 言

我国地址存在命名复杂、结构无序、街道门牌编号混乱[1]等情况; 个人在填写地址时, 随意性大, 书写形式多样. 地址信息书写形式的多样性与标准地址的唯一性匹配是迫切需要解决的问题. 建立综合的标准化地理编码数据库是解决问题的有效方法. 地理编码也可以被称为地址编码[2], 提供一种把描述性地址信息转换为地理坐标的方式. 地址编码过程[2]主要包括地址标准化和地址匹配. 地址标准化是实现地址编码的前提, 其中包括对地址数据现状的分析, 设计合理可行的地址数据结构, 逐步实现地址数据的标准化. 地址匹配的目标是为任何输入的地址数据返回最准确的匹配结果[2]. 本文就绍兴市地名数据库建设提出构思, 并就地址匹配阶段的匹配算法作了较深入的探讨.

2 地址标准化

江洲[3]等人认为地址数据内容标准和规范中应该对地址要素、地址表述形式(地址结构)等作详细的说明和描述.

规范化的地址数据内容描述应通过地址要素名称及其组合实现[4]. 地址要素一般是由一个或多个表达不同地理区域范围的字段所构成, 其中每一部分地址字段都可以看作是一个地址要素, 每个地址要素都由通名和专名两部分组成[4]. 例如, 地址数据“绍兴市环城西路 508 号”由 3 个地址要素构成, 分别是“绍兴市”、“环城西路”、“508 号”; 地址要素“绍兴市”中“绍兴”是

① 基金项目:浙江省教育厅科研项目(Y201327816)
收稿时间:2013-10-25;收到修改稿时间:2013-11-21

专名, “市”是通名. 这里涉及的专名和通名必须是不可再分的有意义的名称, 例如“绍兴”不能再分为“绍”和“兴”.

规范的地址表达形式[4]可以划分为城市地址和乡村地址两种表达形式, 城市地址表达形式中行政区划名一般描述到区/县级, 乡村地址表达形式中行政区划名一般描述到行政村级.

2.1 地址层级模型

采用地址要素表达地址信息的方法, 我们把标准地址分为三部分构成: “地址=行政区划名部分+主地址部分+子地址部分”. 结合绍兴市地址实际, 把地址分为城市地址和乡村地址, 地址表达形式分别为“城市地址=区/县级行政区划名部分+主地址部分+子地址部分”, “乡村地址=乡村级行政区划名部分+主地址部分+子地址部分”, 分别设计地址的层级模型如图 1 和图 2 所示.

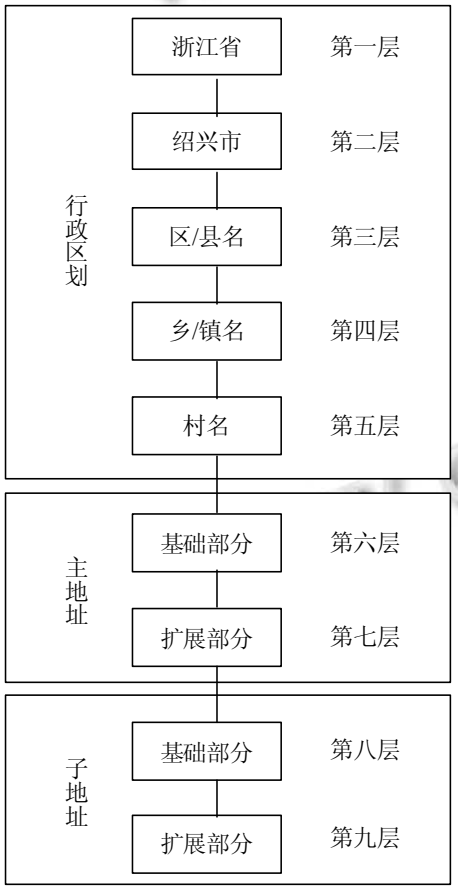

图 1 乡村地址层级结构

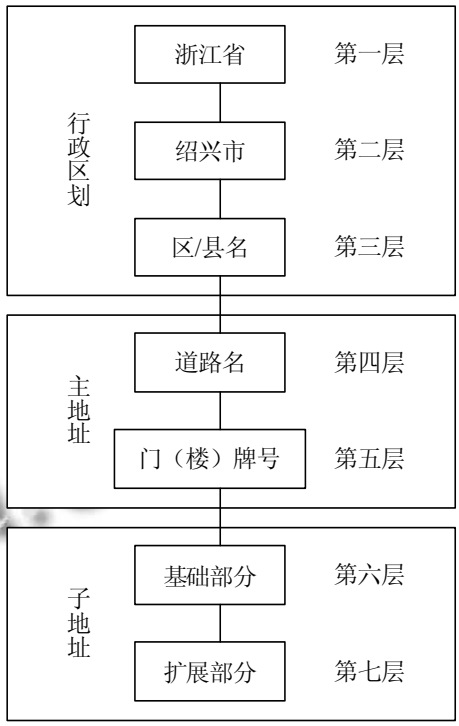

图 2 城市地址层级结构

在乡村地址结构中, 主地址的基础部分一般为住宅区名、建筑群名、自然村名等, 扩展部分一般为编号; 子地址的形式一般为“XXX 幢 XXX 室”. 主地址部分不能整体为空, 但主地址的基础部分及子地址部分可以根据情况确定是否为空, 例如, 地址“浙江省绍兴市绍兴县富盛镇倪家楼村 90 号”的主地址的基础部分和子地址部分为空.

对于乡镇级地址, 一般需要省略图 1 中第 5 层结构, 主地址的基础部分一般为标志性建筑名(如“镇政府”、“邮电局”等).

在城市地址结构中, 主地址部分的第四、五层是必须的, 子地址部分的表达形式一般为“XXX 幢 XXX 室”; 子地址部分可以根据情况确定是否为空, 如“浙江省绍兴市越城区环城西路 508 号”的子地址部分为空.

2.2 地址输入模型

根据地址层级结构, 我们可以建立地址输入模型. 吴海涛[5]等人提出“省(直辖市)→市(地级市)→区(县、县级市)→镇(乡、街道)→村(路)→门牌号”的地址输入模型, 这里, 我们采用 “省→市→区/县/县级市→道路名→门(楼)牌号→子地址”、“省→市→区/县/县级市→

镇/乡→村→主地址→子地址”这两种输入模型, 分别用于城市地址、乡村地址的输入. 这样, 地址标准化需要解决的主要问题就转换为去除冗余信息、填充缺省项、别名转换、修改书写错误等问题.

2.3 地址标准化过程

在地址标准化过程中, 需要去除非标准地址的冗余信息, 如“绍兴市越城区府山街道常禧路 25 号”中“府山街道”为冗余信息, “绍兴市越城区环城西路 508 号绍兴文理学院”中“绍兴文理学院”为冗余信息; 需要补充缺省项, 如“绍兴市环城西路 508 号”中, 需补充“越城区”信息为“绍兴市越城区环城西路 508 号”; 需要把地标名、别名等转换为“道路名+道路号”的形式, 如“绍兴市越城区绍兴文理学院”转换为“绍兴市越城区环城西路 508 号”.

地址标准化的过程为“初始地址→地址拆分→地址规范化→生成关键字地址字段”, 例如, “绍兴环城西路 508 号”的规范化过程如图 3 所示.

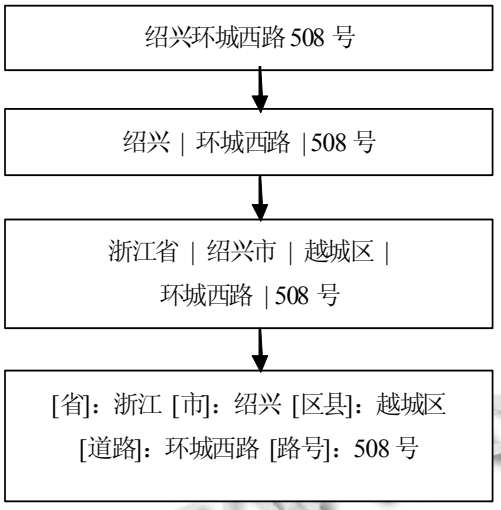

图 3　地址规范化过程示意图

2.4 地址代码编制

基于上述地址表达方式, 在地址代码的编制时, 对于定位到户的地址服务要求, 可以基于国家统计局统计用的行政区划及城乡分类代码进行非定长编码. 例如, 地址“浙江省绍兴市绍兴县富盛镇倪家楼村 90 号”的代码为“33062111220200090”, “绍兴市越城区环城西路 508 号”的代码为“33060203000508”, 具体含义如表 1、表 2 所示.

表 1　乡村地址代码示例

3306	21	112	202	00090
绍兴市区划码	绍兴县区划码	富盛镇区划码	倪家楼村区划码	门牌顺序编号

表 2　城市地址代码示例

3306	02	030	00508
绍兴市区划码	越城区划码	道路顺序号	门(楼)牌号

在区划代码中, 用 4 位代表市级代码, 用 2 位代表县/区级代码, 用 3 位代表街道/乡/镇级代码, 用 3 位代表社区/居委会/村级代码; 道路采用 3 位代码按顺序编号, 门(楼)牌号采用 5 位代码也按顺序编号, 不足部分左补 0.

3　数据库设计

地名数据库采用 Oracle 11g 数据库管理系统实现, 主要包括标准地址表和地址分词表(含行政区划、主地址、子地址等部分). 其中行政区划分为市、区县、乡镇、村等, 表之间的关系通过外键字段 FatherId 关联. 例如“绍兴市”的 ID 为 1, 则“越城区”、“绍兴县”等区县级的 FatherId 为 1, 表示“越城区”、“绍兴县”等属于“绍兴市”. 而层次字段 LevelCode 的取值范围可以在 1~9 之间, 表示该要素的层次. 地址分词表、标准地址表主体结构如表 3 和表 4 所示.

表 3　地址分词表结构

属性	类型	描述
ID	number	序号
LevelCode	number	层次
Code	varchar2	行政区划码编码
Name	varchar2	名称
FatherId	number	父级ID, 外键

表 4 标准地址表结构

属性	类型	描述
ID	number	序号
Code	varchar2	地址代码
StandardName	varchar2	标准地址
ZipCode	varchar2	邮编
CoordinateX	number	X坐标
CoordinateY	number	Y坐标
TipName	varchar2	别名
Supplement	varchar2	辅助信息

4 地址匹配

地址匹配[6]是指根据用户输入的包含地址信息的文字描述, 按照一定的地址匹配策略, 与地理编码库中的地址信息进行比对, 从而获得对应的空间地理坐标, 并定位到电子地图的相应空间位置的过程, 即地址匹配过程为“初始地址→地址拆分、标准化→匹配地址编码数据库→赋予地址地理坐标/地址编码”.

地址匹配的核心要素可分为:

1)分词.

对用户输入的初始地址进行初始化, 也就是将地址字符串切分成一组记录级别的地址要素或标志物通名[7]. 不少论文中都提到了分词的算法, 如吴海涛[5]等人在其论文中就有对地址分词算法的详细描述.

2)匹配.

将切分后的词段或地址要素和地址分词表进行比对, 找出符合条件的最佳结果, 并最终形成完整的地址代码.

在实际生活中, 由于下述原因, 均会导致经过分词阶段后产生的地址要素可能产生误差:

1)用户在手工书写地址时由于客观或主观的原因会出现地址不规范或笔误的情况;

2)将已经书写在纸质材料上的地址利用 OCR 技术进行转换时也可能产生一定的误差;

3)操作人员将地址信息输入到电脑的过程中, 也会因为各种原因而导致输入错误.

4)分词算法本身不够完美而产生误差.

由于上述原因的存在, 所以经过分词阶段后所产生的地址要素仍有可能发生无法在标准的地名数据库中进行精确匹配的可能.

5 利用多模式近似串匹配算法MBPM-BM解决匹配问题

理想状态下, 分词阶段后所产生的地址要素可以进行精确地唯一匹配, 但实际上会因前述各种原因使得无法进行精确匹配.

5.1 总体策略

为此, 本文采取层次控制结合多模式字符串近似匹配(模糊匹配)的策略予以解决. 为便于说明, 以输入城市地址信息“浙江省绍兴市越程区府山街道常禧路 25 号”作为示例, 经过标准化处理去除冗余信息, 再经过分词后得到的地址要素包括“浙江省”、“绍兴市”、“越程区”、“常禧路”、“25 号”. 具体描述如下:

步骤 0: 将分词阶段完成后得到的分词结果放入地址要素数组 S[ ](若为城市地址, 该数组的长度为 7; 若为乡村地址, 该数组的长度为 9), 并设置初始的匹配层次 level=1, 设置当前的候选集 CS(即待匹配集合)为地址分词表中所有层次为 1 的分词信息, 也就是全部的省、自治区、直辖市. 其中候选集的查询语句伪代码为:

```
Select Id,Name Into CS From 地址分词表
Where LevelCode =1
```

本步骤完成后的说明: 候选集 CS 的内容可能包括: {浙江省, 江苏省, 上海市, 北京市, 等等}及相应的 Id, 而分词的结果则得到如下数组 S 的内容:

含义	层次	值
省	1	浙江省
市	2	绍兴市
区/县	3	越程区
道路	4	常禧路
门牌	5	25号
基础子地址	6	(空)
扩展子地址	7	(空)

步骤 1: 以 S[level]的内容为条件(关键字), 在候选集 CS 中对该关键字进行匹配. 计算结果集 IDS, 其计算思想可用 SQL 语句伪代码表示为:

```
Select Id Into IDS From CS Where Name=S[level]
```

1)若能精确匹配, 则取得的唯一 Id 作为结果集 IDS 的内容.

2)若不能精确匹配, 则利用近似匹配算法, 取得一个或多个 Id 作为结果集 IDS 的内容(近似匹配算法的说明详见后述).

3)若 S[level]为空, 则将候选集中的全部 Id 作为结果集 IDS 的内容.

步骤 2: 根据已经得到的结果集 IDS, 在地址分词表中更新候选集 CS, 更新候选集 CS 的 SQL 语句伪代码如下:

```
Select Id,Name Into 新的 CS From 地址分词表
Where LevelCode =level+1 And FatherId In IDS
```

步骤 3: 若数组 S 的元素均已处理完毕, 则匹配结束, 输出结果(候选集); 否则, 当前的匹配层次 level 加 1.

5.2 有关说明

结合上例, 关于步骤 1、步骤 2 的说明如下:

1)当 level 等于 1 时, 由于能够在候选集 CS 中精确匹配“浙江省”, 因此, 结果集 IDS 的内容就是“浙江省”所对应的 Id. 而更新后的候选集 CS 就是浙江省内的城市及其 Id, 其内容可能包括: {绍兴市, 杭州市, 宁波市, 温州市, 等等}及相应的 Id.

2)当 Level 等于 2 时, 由于能够在候选集 CS 中精确匹配“绍兴市”, 因此, 结果集 IDS 的内容就是“绍兴市”所对应的 Id. 而更新后的候选集 CS 就是绍兴市下面的区、县、县级市, 其内容可能包括: {越城区, 绍兴县, 新昌县, 上虞市, 嵊州市, 诸暨市, 等等}及相应的 Id.

3)当 Level 等于 3 时, 由于无法精确匹配“越程区”, 则利用近似匹配算法进行匹配(算法说明见下一节). 其核心思想就是将“越程区”和当前的候选集 CS={越城区, 绍兴县, 新昌县, 上虞市, 嵊州市, 诸暨市, 等等}, 进行近似匹配, 计算编辑距离. 为提高匹配效率, 考虑采用 MBPM-BM 算法[8]进行多模式近似匹配.

MBPM-BM 算法[8]的主要思想是: 首先, 在过滤阶段利用多模式跳跃引理实现快速略过不可能匹配的区域. 其次, 在可能符合匹配条件的区域, 采用位并行算法, 进行多关键并行匹配.

对于本例, 当候选集元素为{越城区, 绍兴县, 新昌县, 上虞市, 嵊州市, 诸暨市}时, 可将这 6 个元素打包成一个计算机的物理字长. 目前主流计算机操作系统的字长是 32 位或 64 位, 而这 6 个关键字只需占用 18 位二进制位, 因此, 只需进行一次匹配就可以寻找出全部可能的匹配项. 若候选集 CS 元素打包后超出了计算机的字长, 则需对元素进行分组后再进行匹配.

近似匹配的误差 k 从 1 开始尝试, 若近似匹配失败, 则调整 k 的大小, 一般采用每轮加一的方式逐步放宽误差值, 并重复这个过程. 当超过最大能够容忍的 k 值, 则匹配失败. 实际上由于 S[level]的长度一般均不大, 因此, 本匹配过程的效率还是非常高的. 若在某一轮近似匹配中成功匹配, 则将匹配成功的一个或多个地址要素作为条件来计算新的候选集 CS.

本例中, 当 k=1 时, MBPM-BM 算法能够定位出的地址要素只有一项“越城区”, 因此, 结果集 IDS 的内容就是“越城区”所对应的 Id. 更新后的候选集 CS 就是越城区下面的道路信息, 其内容可能包括: {常禧路, 环城西路, 解放路, 中兴路, 等等}及相应的 Id.

顺便指出, 若设置误差 k 为 0, MBPM-BM 算法也可实现对多个关键字进行并行精确匹配的功能. 假设可以将候选集 CS 中的多个元素打包到一个物理字长(本例中绍兴市下面的区县共有六个, 每个长度均为 3, 总长度为 18, 小于计算机物理字长), 然后进行一次匹配即可.

4)当 Level 等于 4 时, 由于能够在候选集 CS 中精确匹配“常禧路”, 因此, 结果集 IDS 的内容就是“常禧路”所对应的 Id. 而更新后的候选集 CS 就是常禧路下面的门牌号, 其内容可能包括: {1 号, 2 号, …, 25 号, 等等}及相应的 Id.

5)当 Level 等于 5 时, 由于能够在候选集 CS 中精确匹配“25 号”, 因此, 结果集 IDS 的内容就是“25 号”所对应的 Id. 而更新后的候选集 CS 为空.

6)当地址要素数组 S 的内容已被处理完毕, 则匹配结束. 需要说明的是, 此时只是得到了最终匹配的某个 ID 及相应的编码(Code), 完整的地址代码信息则可以通过简单的回溯, 不断查找当前记录的父级 ID(FatherId)并将对应的编码进行串接即可. 有了完整的地址代码, 若还需要获得地址的邮编、坐标等信息, 则可以非常方便地在标准地址表中找到相关信息.

5.3 匹配阶段的流程图

图 4 是匹配阶段的总体流程图.

5.4 匹配算法伪代码

```
完成地址要素数组 S 的准备工作
Level←1
计算 CS        // 计算初始候选集, 详见步骤 0
```

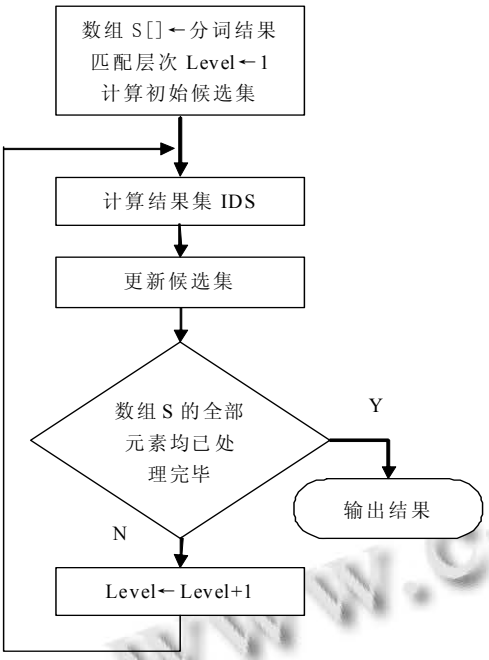

图 4 匹配阶段的总体流程图

While true Do
在 CS 中精确匹配 S[level]
If 精确匹配失败 Then
设置误差初值 k←1
设置误差上限 uk
Repeat
MBPM-BM(CS,S[Level],k)
If 近似匹配成功 Then Break
k←k+1
Until k> uk
End If
利用上面的精确或近似匹配结果获得 IDS
利用 IDS 更新候选集 CS
Level←Level+1
If 数组 S 的内容处理完毕 Then Break
End While
输出最新的 CS(回溯串接后得到完整地址代码)

6 结 语

地址编码数据库在城市信息化建设、数字化城市中具有极其重要的地位. 由于我国地名情况复杂, 不易建立统一的地址数据模型和标准. 国内的发展趋势是进行地址数据内容规范研究[4], 建立全国地址信息标准化的标准, 实现全国地址信息标准化, 建立全国地址编码数据库. 本文从绍兴市地名实际出发, 把地址分为城市地址和乡村地址两种形式进行表达, 建立层级地址模型, 基于行政区划码进行地名代码编制; 在地址编码数据库的地址标准化、地址匹配、数据库设计等方面进行了探讨, 并提出了应用多模式近似串匹配算法在地址匹配阶段的应用方案.

参考文献

1 李军,李琦,毛东军.北京市地理编码数据库的研究. 计算机工程与应用,2004,41(2):1–4.
2 江洲,李琦.地理编码(Geoceding)的应用研究.地理与地理信息科学,2003,19(3):24–25.
3 江洲,李小林,刘碧松.地理信息系统地址编码技术标准化研究. 世界标准化与质量管理,2007,(5):22–25.
4 佟文会,江洲,李小林.地址编码关键技术—地址数据内容规范研究.标准科学,2009,(11):39–42.
5 吴海涛,俞立,张贵军.基于模糊匹配策略的城市中文地址编码系统.计算机工程,2011,37(2):194–196,199.
6 钱敏,顾国强,鲁明.用于地址(地理位置) 匹配的关键路径法.计算机应用与软件,2012,29(1):211–214,219.
7 马照亭,李志刚,张伟,印洁.一种基于地址分词的自动地理编码算法.测绘通报,2011,(2):59–62.
8 范立新,谢晓能,吴飞.基于过滤的中文多模式近似字符串匹配算法.计算机工程,2006,32(20):48–50,58.