基于规则的中文地址匹配系统

庄海东，张鸿恩
（厦门精图信息技术股份有限公司 福建 厦门 361008）

【摘 要】:地址是日常生活中最常用的一种地理信息描述。但目前地址信息的描述未形成统一标准,很多地址信息无法共用。地址匹配技术为地址信息的通用提供了途径。本文从地址模型建立到地址拆分再到地址匹配,提供一套较为完整的地址匹配的系统建设方案。实践证明,该建设方案准确率较高,在数字城市建设过程中的政府业务系统间的数据互通起到一定的作用。

【关键词】:地址模型 地址拆分 地址匹配 地理编码 标准地址

1 引言

城市信息日常管理中，存在大量的地址信息,但这些信息并未形成统一的标准,各个单位对于地址有各自的描述方式,同时未形成地址信息的空间位置表示。

地址匹配技术通过将各单位地址与标准地址的匹配能够实现数据的标准化和空间定位。

本文采用基于规则的中文要素解析方法进行地址解析,引入基于权重的地址匹配算法,并针对实际应用过程中碰到的地址匹配问题引入地址评价规则库,提高匹配的精确度。

2 关键技术分析

本文通过对地址模型的定义,识别出地址中的特征类型,根据该特征类型整理出地址特征字,通过地址特征字的解析实现对中文地址的拆分。

地址拆分成不同特征部分词组后,通过计算相似度以及基于权重和规则库评价机制对不同的地址进行匹配。

系统需解决地址模型、地址解析和地址匹配三项关键技术。

2.1 地址模型及其从属关系

地址模型建立参照 GB/T 23705-2009 文件 [1],主要构成部分如图 1 所示:

图 1 地址模型

构建地址模型如下：

1、行政区域地名+街巷名+门楼址

2、行政区域地名+街巷名+标志物名

3、行政区域地名+小区名+门楼址

4、行政区域地名+小区名+标识物名

5、行政区域地名+街巷名+门楼址+标志物名

6、行政区域名+街巷名+小区名

7、行政区域地名+小区名

8、 行政区域地名+街巷名+小区名+门楼址+标志物名

根据地址不同构成部分的地理空间特性将地址构成分为从属关系和跨从属关系两类[2]。行政区域、小区名、门楼址、标志物名存在从属关系,街巷名由于其地理位置跨度较大,为跨从属关系。

根据地址的从属关系建立标准地址数据索引,方便对标准地址进行查找[3]。

地址间的从属关系定义如表 1 所示:

表 1 地址构成部分从属

地址构成	级别	从属关系
行政区域	10	-1
街巷名	20	10
小区名	30	10,20
门楼址	40	20,30
标志物名	50	20,30,40

例如：

厦门市思明区吕岭路 1819 号精图数码大厦

厦门市思明区吕岭路三安电子

厦门市思明区岭兜小区 141 号根据地址拆分规则,形成地址索引数据如图 2 所示:

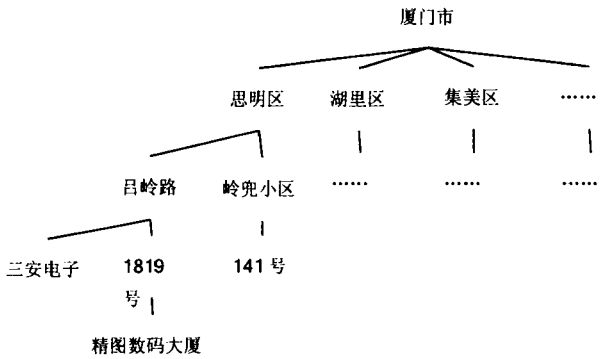

图 2 地址数据索引

2.2 地址解析

地址信息的解析重点在于地址拆分。地址由固定部分构成，系统基于地址模型解析出特征字进行地址拆分[4]。

根据地址构成，同时对大量样例数据进行分析，增加“方位”构成部分，提取出不同构成部分对应的特殊字（参见表 2）。地址解析过程中通过对特征字的解析，达到对地址的不同构成部分的提取，实现地址拆分。

表 2 地址构成特征字

地址构成	特征字
行政区域	省、市、区、街道、乡、镇……
街巷名	路、街、大道、巷、横路、横街、纵路、纵街、弄、线……
小区名	村、里、坊、横、队、社……
门楼址	栋、座、型、阁、号、#、楼、层、室、房……
标志物名	大厦、商场、商城、城、公司、宾馆、别墅、商店、所……
方位	东、西、南、北、部、侧、外、端、段……

利用正则表达式实现关键字词正向最大值匹配，形成地址信息的拆分。

在进行地址拆分时由于部分地址构成特征字存在重复，需要对这些数据进行冲突检测处理，并根据检测后的结果进行地址构成的正确划分。

地址拆分、冲突检测后仍存在部分无法正确归类的地址数据，该部分数据与现有的 POI 数据进行对比分析，从而实现更精确的拆分。

2.3 地址匹配

对现存的标准地址数据进行地址解析，根据地址模型中的地址从属关系构建地址索引数据库，为后续的地址匹配服务提供基础数据。

为验证两个地址的匹配程度，引入地址不同构成部分的权重设置（参见表 3）。

定义匹配程度计算方法如下：

$$M=\sum_{i=0}^{n}(W_i\times M_i)/\sum_{i=0}^{n}W_i \qquad (1)$$

其中地址构成权重值，根据地址构成从权重表中获取，取值范围为 $0\leqslant W_i\leqslant 1$，为两个字符串的匹配程度值，根据编辑距离计算得出。

当两个地址完全匹配时其匹配程度为 1，完全不能匹配时其匹配程度为 0。

表 3 地址构成部分权重

地址构成	权重值(%)
行政区域	10
街巷名	15
小区名	15
门楼址	30
标志物名	30

通过对地址不同构成部分赋予的不同权重值实现对地址匹配程度的控制。通过实际生产经验来看，该权重值设置具备较理想的匹配效果。

实际匹配过程中，考虑城市建设过程中可能进行的行政区划调整、道路命更名以及建筑物的命更名等过程，建立同名词库来规避因为命更名所产生的同一地点不同命名的问题，提高地址匹配的包容度。

通过权重及同名词库对地址进行匹配后发现存在因为地址相差较大，但却完全匹配的情况。原因在于不同地址描述信息由不同的部分构成，导致匹配度计算上有所偏差。如：“厦门市思明区”[地址 1]跟“厦门市思明区吕岭路 1819 号 A 座”[地址 2]同“厦门市思明区吕岭路 1819 号 B 座”[标准地址]进行对比，根据权重配置原则，则“厦门市思明区”的匹配程度要高于“厦门市思明区吕岭路 1819 号 A 座”。具体匹配值如表 4 所示：

表 4 权重匹配结果

	地址 1		地址 2		标准地址	
权重值	拆分结果	匹配结果	拆分结果	匹配结果	拆分结果	匹配结果
1	厦门市	1	厦门市	1	厦门市	1
1	思明区	1	思明区	1	思明区	1
1	——	——	吕岭路	1	吕岭路	1
1	——	——	1819 号	1	1819 号	1
0.3	——	——	A 座	0.5	B 座	1
匹配值	1		0.97 式(1)		——	

由表可见，尽管“厦门市思明区”并不是一个有实际意义的地址，但是其根据权重值得出的匹配值要比地址二的匹配程度要高。为解决该问题，引入“评价规则库”。

评价规则库实现对由不同地址构成的地址信息

的匹配程度的定义。

一个地址由哪几个部分构成，比如“厦门市思明区”由行政区域这个部分构成，“厦门市思明区吕岭路1819号A座”由行政区域、街巷名、门楼址等三个部分构成，“厦门市思明区吕岭路1819号B座”也同样由行政区域、街巷名、门楼址三个部分构成。构成部分值越高的其匹配率也相应的越高。根据该思路，定义评价规则库如表5所示：

表5 评价规则表

标准地址 / 输入地址	Q-J-M	Q-J-B	Q-X-M	Q-X-B	Q-J-M-B	Q-J-X	Q-X	Q-J-X-M-B
Q-J-M	1	0.25	0.4	0.1	0.55	0.25	0.1	1
Q-J-B	0.25	1	0.1	0.4	0.55	0.25	0.1	1
Q-X-M	0.4	0.1	1	0.25	0.4	0.25	0.25	1
Q-X-B	0.1	0.4	0.25	1	0.4	0.25	0.25	1
Q-J-M-B	0.55	0.55	0.4	0.4	1	0.25	0.1	1
Q-J-X	0.25	0.25	0.25	0.25	0.25	1	0.6	1
Q-X	0.1	0.1	0.6	0.6	0.1	0.6	1	1
Q-J-X-M-B	1	1	1	1	1	1	1	1
……	……	……	……	……	……	……	……	……
Q	0	0	0	0	0	0	0	0

其中Q代表行政区域、J代表街巷名、X代表小区名、M代表门楼址、B代表标志物，可以通过调整该表格的具体数值及其扩展表格输入地址的种类来匹配不同输入地址类型及其权重值。

上面的问题通过评价规则库，进行如下处理：

“厦门市思明区”与“厦门市思明区吕岭路1819号B座”之间的权重匹配值为1，通过评价规则库表找到其对应规则为输入地址Q，标准地址为Q-J-M，对应的评价值为0，则这两条数据的最终值为0，而“厦门市思明区吕岭路1819号A座”与“厦门市思明区吕岭路1819号B座”之间的权重匹配值为0.96，通过评价规则库找到其评价值为1，则最终值为0.96。“厦门市思明区吕岭路1819号A座”相对“厦门市思明区”的最终匹配值要高。

3 系统实现

3.1 系统架构

系统以地址模型为基准，通过对标准地址进行拆分建立地址索引库数据，同时对输入的地址进行拆分，形成输入地址的拆分数据，而后将输入地址的拆分数据与建立起来的中间索引数据进行匹配，得出输入地址的匹配结果数据。

3.2 处理过程

地址匹配系统以.Net Framework 4.0为平台，应用Visual C#开发。系统通过选择标准地址与待匹配地址，系统实现对于待匹配地址与标准地址的匹配工作。

在进行人工匹配的时候，可以查看该地址的拆分结果（见图3），并可进行调整，同时支持对该拆分结果进行同义词词库保存。

图3 地址拆分结果

3.3 结果分析

系统已经在实际生产中投入使用，通过对海口市23460地址数据的进行匹配，测试结果的分析如表6所示：

表6 匹配结果分析

匹配度	匹配个数	占比例
匹配度100	14820	0.6317135549872123
匹配度90-99	3880	0.16538789428815
匹配度80-89	1580	0.0673486786018755
匹配度70-79	800	0.0341005967604433
匹配度70以下	2190	0.0933503836317136
匹配度0	190	0.0080988917306053

从表中数据可知：有效匹配率（匹配度80以上）达到了86.45%。该结果为程序运行后的匹配结果统计。在生产过程中，因实际地址信息的复杂性，在匹配度小于100%的情况下，根据情况有可能需要进行一定的人工干预。

4 结束语

本文构建了完整的地址匹配处理过程。根据地址模型提炼出地址中的特征字，通过特征字进行地址拆分操作，较好的解决了根据词库进行分词操作时需要全面、准确词库的问题。利用地址模型构建地址不同构成部分间的从属关系，方便地址匹配查询。而后为地址特征字解析出来的不同地址部分设置权重值来处理匹配度的计算问题，以及根据地址模型生成的匹配模型来解决匹配算法中因为匹配 （下转第146页）

现用户在使用交易查询功能时能够回溯到历史交易数据，并且以一定的方式，诸如报表等加以展示。

(2) 交易统计

根据对交易对象，地域，交易产品种类，数量，交易金额的计算和统计，分析出交易趋势，并以直观的方式展示给用户。用户只需要根据需要设定参数，就能得到需要了解的采购或销售方面的烟草市场趋势。

(3) 在线支付

在线支付是一种通过第三方提供的与银行之间的支付接口进行支付的方式。由于烟草专卖品属于特殊商品，需要在线支付的安全等级相对较高，因此本平台门户网站选取信誉较好，规模较大的第三方支付平台作为合作方，包括支付宝、财付通、快钱。通过调用第三方平台公开的支付接口来实现支付功能。

三、关于烟草交易平台门户网站的权限控制

1. 根据用户的级别呈现不同的内容

在登陆界面点击确定登陆时进行页面重定向，根据角色的不同，确定跳转页面的地址以实现这一功能。角色所对应的权限(可以跳转至的页面范围)可以通过管理员配置。

2. 防止非系统用户绕过登录页面

解决思路是利用 session 存取当前登录的用户 ID，每当进入需要权限验证的页面时，访问该 session 变量以判断是否已登录。该 session 变量第一次赋值是在用户合法登录后。

3. 防止低权限用户越权访问

每当进入需要权限验证的页面时，首先判断该角色是否具有查看该页面的权限。通过遍历角色权限表，若包含则予以通过，若不包含则重定向至登陆页。同样的，该表示可以通过管理员进行配置的。

四、结束语

平台经济，作为近期互联网领域、电子商务领域、金融领域炙手可热的概念已经悄无声息的进入日常生活中，依托大数据技术的成熟运用将烟草的线下销售平台化渐成“平台经济”趋势发展的方向。本文正是利用平台经济概念、大数据技术提出了烟草流通平台的架构设想与实现方法，未来市场前景十分看好。

参考文献：

[1]陈威如[J]. 平台战略:正在席卷全球的商业模式革命[M]. 中信出版社.?2013-01-01

[2]王[illegible][J]. 平台战争[M]. 中国纺织出版社.?2013-01-01

[3]徐晋[J]. 平台经济学:平台竞争的理论与实践[M]. 上海交通大学出版社.?2013-01-01

[4]维克托·迈尔－舍恩伯格[J]. 大数据时代:生活、工作与思维的大变革[M]. 浙江人民出版社.?2013-01-01

[5]涂子沛[J]. 大数据:正在到来的数据革命[M]. 广西师范大学出版社.?2012-07-01

[6] 成由根 [J]. 掘金大数据 [M]. 北京时代华文书局.?2013-09-01

(上接第 132 页)

模型不同而导致的匹配度计算错误的问题。

系统主要基于标准地址库来进行地址数据的匹配，用来匹配的标准地址库要求完整、可靠，匹配结果的准确率及其可靠性才能更高。同时由于匹配过程经历了多个复杂的对比和计算的，匹配效率并不是很高，这些需要再优化。

参考文献：

[1] GB/T 23705-2009 .数字城市地理信息公共平台地名/地址编码规则[S]

[2] 孙亚夫 陈文斌. 基于分词的地址匹配技术[EB/OL]. http://wenku.baidu.com/view/8ee9deef5ef7ba0d4a733b7f.html, 2011-05-26

[3] 吴海涛 俞力 张贵军. 基于模糊匹配策略的城市中文地址编码系统[J]. 计算机工程,2011,37(2):194-196

[4] 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):9-16

[5] 马照亭 李志刚 孙伟等. 一种基于地址分词的自动地理编码算法[J]. 测绘通报,2011,(2):59-62

[6] 刘哲. ETL 过程中的数据清洗技术研究与应用 [D]. 沈阳:沈阳航空工业学院,2007

[7] 杨青 陈薇 闻彬. 面向语义信息查询的模糊本地模型[J]. 计算机工程,2010,36(8):188-190