Journal of Computer Applications
计算机应用,2017,37(3):876 - 882
ISSN 1001-9081
CODEN JYIIDU
2017-03-10
http://www.joca.cn

文章编号:1001-9081(2017)03-0876-07
DOI:10.11772/j.issn.1001-9081.2017.03.876

非规范化中文地址的行政区划提取算法

李晓林[1,2],黄　爽[1,2*],卢　涛[1,2],李　霖[3]
(1.武汉工程大学 计算机科学与工程学院,武汉 430205;　2.智能机器人湖北省重点实验室(武汉工程大学),武汉 430205;
3.武汉大学 资源与环境科学学院,武汉 430079)
(*通信作者电子邮箱 13986287758@163.com)

摘　要:由于互联网上中文地址的非规范化表达,导致互联网中的中文地址信息在地理位置服务中难以直接应用。针对此问题,提出一种非规范中文地址的行政区划提取算法。首先,对原始数据进行"路"特征词分组预处理;再利用行政区划字典和移动窗口最大匹配算法,从中文地址中提取所有可能的行政区划数据集;然后,利用中文地址行政区划元素之间具有层次关系的特点,建立行政区划条件集合运算规则,对获取的数据集进行集合运算;再利用行政区划匹配度建立一种行政区划集合解析规则,来计算行政区划可信度;最后,得到可信度最大信息量最完整的中文地址的行政区划。利用从互联网中提取的约25万条中文地址数据进行是否采用"路"特征词分组处理以及是否进行可信度计算处理,对算法的可用性进行了验证,并与目前的地址匹配技术进行对比,准确率达到93.51%。

关键词:集合运算;行政区划;中文地址;移动窗口;匹配度;解析规则
中图分类号:TP391.1　　**文献标志码**:A

Administrative division extracting algorithm for non-normalized Chinese addresses

LI Xiaolin[1,2], HUANG Shuang[1,2*], LU Tao[1,2], LI Lin[3]
(1. *School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan Hubei* 430205, *China*;
2. *Hubei Provincial Key Laboratory of Intelligent Robot (Wuhan Institute of Technology), Wuhan Hubei* 430205, *China*;
3. *School of Resource and Environmental Sciences, Wuhan University, Wuhan Hubei* 430079, *China*)

Abstract: Chinese addresses on the Internet are always non-normalized, which cannot be used directly in location-based services. To solve the problem, an algorithm to extract administrative divisions from non-normalized Chinese addresses was proposed. Firstly, preprocessing "road" feature word grouping for original data; using administrative division dictionary and moving window maximum matching algorithm, extract all possible administrative region data sets from Chinese address. Then, using the Chinese administrative divisions between the elements of the hierarchical relationship between the characteristics, the administrative set conditional set operation rule was established and the acquired data set was aggregated. using the administrative division of matching, a set of administrative division set rules were established to calculate the credibility of the administrative division. Finally, the credibility of the maximum amount of information the most complete Chinese address of the administrative divisions were obtained. By using the extracted from the Internet about 250 000 Chinese address data whether the use of "road" feature word packet processing and whether to carry on the credibility calculation process was verified for the availability of the algorithm, and with the current address matching technology for comparison, the accuracy rate of 93.51%.

Key words: set operation; administrative division; Chinese address; moving window; matching degree; analytical rule

0　引言

自然语言处理(Natural Language Processing, NLP)是人工智能领域的一个重要组成部分,它能实现人与计算机之间用自然语言进行有效通信的各种理论和方法[1]。随着互联网技术的飞速发展,网络上的信息量更是大得惊人,汉语是世界上使用人数最多的一门语言,那么中文信息处理自然也是NLP的重要分支,而中文分词又是中文信息处理的基础,是中文信息处理的第一步,只有做好中文分词,后面对信息处理的步骤才会精确,因此高效准确的中文分词意义重大。就自然语言处理这方面来说,西方由于其语言的天然便利性等因素,其处理发展得比较好,形成了不少成熟的技术[2],但是这些理论与方法常常都不能直接作用于汉语之上,原因是汉语自身的语言结构和西文差别较大,汉语不像其他外文如英语,没有天然的词汇分隔符,所以要对中文信息进行处理就必须首先完成中文分词[3]。

目前常见的中文分词算法主要分为三类[4]:基于词库的分词算法、基于统计的分词算法、基于理解的分词算法。

在基于互联网位置服务的领域中,中文分词也发挥了极大的作用。基于互联网位置服务是即时定位的位置服务,实

收稿日期:2016-08-26;**修回日期**:2016-10-18。　**基金项目**:测绘地理信息公益性行业科研专项(201412014);国家863计划项目(2013AA12A202);湖北省自然科学基金资助项目(2013CFA125);武汉工程大学第七届研究生创新基金资助项目(CX2015053)。
作者简介:李晓林(1962—),男,湖北孝感人,副教授,硕士,主要研究方向:数据挖掘、机器学习、人工智能;　黄爽(1992—),女,湖北武汉人,硕士研究生,主要研究方向:数据挖掘、机器学习、人工智能;　卢涛(1980—),男,湖北武汉人,副教授,博士,主要研究方向:图像/视觉处理、计算机视觉、人工智能;　李霖(1960—),男,湖北孝感人,教授,博士生导师,博士,主要研究方向:地理语义及本体、三维建模及可视化。

时为用户提供准确的地理位置信息,实现各种与位置相关的业务。在基于互联网位置服务中,地理位置可以有多种表达形式,中文文本表达是其中之一,用户可以通过中文地址信息获取他们所需的精确地址,更好地提高服务质量。随着地理信息系统在人们生活中的作用越来越重要,对于根据中文文本地址信息快速、准确查找其地理坐标的需求日益明显[5]。地址匹配技术能够在地理编码库中比对出相应的地理坐标,满足人们的需求。在地址匹配方面,采用分词的地址匹配技术,可以解决大多数非空间坐标地址的匹配问题。地址分词是地址匹配的基础,直接影响地址匹配的准确性,地址分词就是根据输入地址字符串、地址词典、地址模型,将地址切分转换为计算机能够理解的、结构化的词组。西文地址分解可以按照空格、标点等进行单词分割,中文地址分词需要借助地名语料库(地址词典)和中文分词算法进行中文地址分词。

一个规范的中文地址应包含完整的行政区划,并按照行政区划(省/市/县/乡/村)、路街、牌号、建筑、户室的次序来表达[6],特征字明显,利用中文地址分词算法好切分,从而可以准确地与该地址的地理位置对应。然而,在互联网上,中文地址经常用非规范行政区划方式来描述,表述混乱与模糊,难以确定该地址所表达的地理位置,作为位置服务是无效的[7],因此,普通的中文地址分词算法无法很好地解决非规范的中文地址问题,需要在中文地址分词算法上研究一种优化的中文地址解析算法来解析非规范的中文地址。中文地址中与行政区划相关的不规范的表达方式有:省略行政区划特征词、省略部分行政区划、无行政区划、行政区划信息层次杂乱。此外,地址的非行政区划部分存在与行政区划同名的情况,主要表现在:路街的名称常用行政区划名称命名、建筑(或企业)名称中包含行政区划名称、地名与行政区划同名等。在互联网中纷杂的非规范信息中,辨别出相对于用户需要的信任度比较高的信息,在当今地理信息位置服务方面变得十分必要[8]。

因此,本文提出一种非规范中文地址的行政区划提取算法,对数据进行“路”特征词分组预处理,并根据中文地址具有层级关系的特点建立了条件集合运算规则,对通过移动窗口最大匹配算法中的提取的行政区划集合进行集合运算,并利用行政区划匹配度建立一种行政区划集合解析规则,计算行政区划可信度,从非规范的中文地址中提取出最完整准确的行政区划,可以有效地提高地址数据查找的速度和准确性,从而提高网络地图在线服务质量,为用户更好的定位。

1　相关工作

根据地址匹配算法的特征分类,迄今为止现有的中文地址匹配算法主要有三类[9]:

1)以地址要素层级模型为核心的地址匹配算法。此类算法以地址具有级别属性的特点来构建模型,这类算法的匹配率依赖于地址表述的规范性。文献[10]地址要素识别机制的地名地址分词算法,提出基于地址要素识别机制的地名地址分词算法,采用最大正向匹配算法,增加了基于地址要素的识别机制,提高了地址分词的准确度,但匹配速率却下降了。文献[11]基于分级地名库的层级结构,按照地址要素的等级进行迭代处理,匹配过程是逐级匹配。

2)以全文检索模型为核心的地址匹配算法。此类算法是将地址库作为文本库,将待匹配的地址作为检索条件,这类算法只考虑关键词匹配,匹配速率高,但是准确率不高。文献[12]建立了地址要素词库,利用正向最大匹配算法进行地址分词。文献[13]通过建立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分词匹配方法,但是对于大规模或大范围的地名地址数据,该算法不仅查找的速度慢,而且没有顾及地址的语义信息,导致查找的准确性较低,查找结果多样且往往不是用户所需要的结果。

3)以正则表达式匹配为核心的地址匹配算法。此类算法是以特征字为分界线使用正则表达式匹配的方法在地址库中进行查找,这类算法匹配速度慢,匹配率高,但准确率低。文献[14]通过系统分析地址要素的构词特征和句法模式,构建了各类地址要素的特征字库,提出了中文地址的数字表达方法,设计了基于规则的中文地址要素解析方法。但是部分解析规则存在冲突现象,导致部分信息无法正确解析,且对于不具备特征字的地址要素,只能解析出部分信息。文献[15]在中文地址编码研究中采用分段、组合、优先规则,对中文地址进行分段匹配,这些规则虽然减少了地址要素匹配次数,但是由于采用数据库查询的方式,算法总体匹配速率不高。

但这些算法大部分依赖于中文地址规范性、特征字以及地址词典,对规范的中文地址能够取得不错的成效,但对于非规范中文地址,成效不佳,因此为解决上述依赖规范中文地址、匹配速率、匹配准确度问题,同作者的文献[16],提出基于条件随机场的中文地址行政区划提取方法。该方法根据中文地址中行政区划的表达特点和特征,采用判别式概率模型,在观测序列已知的基础上对目标序列建模,通过构建语料训练集和建立相应的特征模板,得到行政区划的表达模型。对非规范中文地址的行政区划提取取得一定效果,但是此方法依赖于训练语料,需要进行人工标注,是有监督学习方法,因此在此研究基础上,本文提出一种非规范中文地址的行政区划提取算法,运用“路”特征词分组、行政区划集合运算以及可信度计算等方法,对原地址数据进行处理运算,避免人工预处理,提高了整个算法的运算速率,并且也提高了地址匹配的准确率。

2　非规范中文地址行政区划提取算法

互联网上的地址信息纷繁复杂,且由于人为书写原因以及各方面别的原因造成中文地址信息错误或者遗漏,所以本文先利用移动窗口匹配算法匹配行政区划得出所有可能行政区划结果集,再进行可信度计算得出可信度最大的中文地址,那么如何从匹配到的所有可能的行政区划集合中提取准确的中文地址信息是要解决的问题。一般是运用集合的交集运算对行政区划集合中的行政区划进行计算,来提取准确的行政区划结果。一般的交集运算是指按照行政区划中每一级行政区划所对应的行政区划元素是否相等,如果相等则取行政区划元素的值,如果不等,则该行政区划元素交的结果为空,但是在两个行政区划进行交集运算时,不能简单地按照各级元素是否相等来确定行政区划相交的结果,否则行政区划交运算的结果不是期望的结果。例如表 1(3 级行政区划,省、市、县)。

中文地址中的一个行政区划是一组有序的行政区划元素组成,行政区划元素是指中文地址中的词可以与行政区划字典成功匹配出一个或多个行政区划的词。行政区划包含有省、市、县、乡、村 5 级,可表示为:行政区划 = {省,市,县,乡,村},用 D 表示行政区划,$d_i(i = 1,2,3,4,5)$ 表示行政区划中

的每个元素,则行政区划 D 表示为:$D = \{d_1, d_2, d_3, d_4, d_5\}$。

表 1 行政区划交运算示例

Tab. 1 Administrative division intersection calculation example

示例	行政区划		结果 $D_1 \cap D_2$	
	D_1	D_2	交集	期望
11	江苏省,徐州市,鼓楼区	江苏省,南京市,鼓楼区	江苏省, ,鼓楼区	江苏省, ,鼓楼区
22	, ,鼓楼区	江苏省,南京市,鼓楼区	, ,鼓楼区	江苏省,南京市,鼓楼区

表1中 D_1 和 D_2 分别表示2个行政区划。期望是指根据给出2个行政区划的各个元素值推理出的一个合理行政区划。可以看出示例2中,当两个行政区划中其中一个行政区划缺失2级行政区时,交集运算得到的结果不是所期望的结果,期望的结果是 $D = D_1 \cap D_2 = \{$江苏省,南京市,鼓楼区$\}$。

根据以上的行政区划一般的交集算法无法得出期望的结果,因此需要有一个能够适应行政区划交集运算方法使计算结果达到期望的行政区划。为解决这个问题,本文在一般的集合运算的基础上提出一种条件集合运算。

2.1 行政区划集合运算

2.1.1 一般的集合运算

常见的行政区划区划集合运算是以下几种:

1)2个行政区划的交集运算。

若有2个行政区划 $D_1 = \{d_{11}, d_{12}, d_{13}, d_{14}, d_{15}\}$ 和 $D_2 = \{d_{21}, d_{22}, d_{23}, d_{24}, d_{25}\}$,则行政区划的交为各级行政区划元素的交,记为:$DI$,用式(1)表示。2个行政区划元素的交记为:$dI_i(i = 1,2,3,4,5)$。

$$DI(D_1, D_2) = D_1 \cap D_2 = \{d_{11}, d_{12}, d_{13}, d_{14}, d_{15}\} \cap \{d_{21}, d_{22}, d_{23}, d_{24}, d_{25}\} = \{dI_1, dI_2, dI_3, dI_4, dI_5\} \tag{1}$$

由于行政区划元素之间存在包含关系,即除了省级区划外,其他各级区划都属于1个或 n 个上级行政区划,所以行政区划交集运算时先计算省级行政区化元素的交,再计算非省级区划元素的交。

① 省级行政区划元素的交。

$$dI_1 = d_{11} \cap d_{21} = \begin{cases} d_{11}, & d_{11} \cap d_{21} \\ \varnothing, & d_{11} = \varnothing \wedge d_{21} = \varnothing \\ \rho, & d_{11} \neq d_{21} \wedge (d_{11} = \varnothing \vee d_{21} = \varnothing) \end{cases} \tag{2}$$

其中:$\wedge$ 为与运算符,$\vee$ 为或运算符。省级区划元素交的结果为 ρ 时,ρ 表示不确定,即2个行政区划中存在一个行政区划的省级区划元素为空 $\varnothing$。此时需要对省级区划元素为空的行政区划利用行政区划字典查询得到省级区划元素非空的行政区划。

假设,两个相交的行政区划 D_1 和 D_2 中,其中一个行政区划 $D_i(i = 1,2)$ 中的省级区划元素 $d_{i1}(i = 1,2)$ 为空,即 $d_{i1} = \varnothing, \exists d_{ik} \neq \varnothing(i = 1,2, k = 2,3,\cdots,5)$,选取一个区划元素 d_{ik},此行政区划元素 d_{ik} 是此行政区划元素中等级最小的一个,用式(3)表示:

$$d_{ik} = \arg\min_k \{d_{ik} \mid d_{ik} \neq \varnothing\};\ i = 1,2 \tag{3}$$

则用行政区划字典查询 d_{ik} 得到 m 个包含行政区划元素 d_{ik} 的行政区划的集合:

$$query(d_{ik}) = DS(d_{ik}) = \{D_{i1}, D_{i2}, \cdots, D_{im}\} = \{\{d_{i11}, d_{i12}, \cdots, d_{i1k}\}, \{d_{i21}, d_{i22}, \cdots, d_{i2k}\}, \cdots, \{d_{im1}, d_{im2}, \cdots, d_{imk}\}\};\ i = 1,2 \tag{4}$$

此时两个行政区划 D_1 和 D_2 省级区划元素交的计算应为省级行政区划元素为空的行政区划求得集合 $DS(d_{ik})$ 中每一个省级行政区划元素与另一个省级行政区划元素不为空行政区划的省级行政区划元素进行依次交运算,求并集:

$$d_{11} \cap d_{21} = \begin{cases} \{(d_{111} \cap d_{21}) \cup \cdots \cup (d_{1m1} \cap d_{21})\}, & d_{11} = \varnothing, \text{即 } i = 1 \\ \{(d_{11} \cap d_{221}) \cup \cdots \cup (d_{11} \cap d_{2m1})\}, & d_{21} = \varnothing, \text{即 } i = 2 \end{cases} \tag{5}$$

其中:$d_{11} = \varnothing$ 表示 D_1 的省级区划元素为空,$d_{21} = \varnothing$ 表示 D_2 的省级区划元素为空。

② 非省级区划元素的交。

$$dI_i = d_{1i} \cap d_{2i} = \begin{cases} d_{1i}, & d_{1i} = d_{2i} \\ d_{1i}, & d_{1i} \neq d_{2i} \wedge d_{2i} = \varnothing \wedge \exists dI_j \neq \varnothing (j < i) \\ d_{2i}, & d_{1i} \neq d_{2i} \wedge d_{1i} = \varnothing \wedge \exists dI_j \neq \varnothing (j < i) \\ \varnothing, & d_{1i} \neq d_{2i} \wedge d_{1i} = \varnothing \wedge d_{2i} = \varnothing \end{cases} \tag{6}$$

当区划元素相等时,则交的结果为区划元素;

当区划元素不相等,且区划元素都不为空,则结果为空;

当区划元素不相等,且区划元素有一个为空时,如果存在非空的交父元素($\exists dI_j \neq \varnothing$),结果为非空区划元素值。

2)1个行政区划集合的交集运算。

一个行政区划集合 $DS = (D_1, D_2, \cdots, D_m)$,并且 $D_1, D_2, \cdots, D_m$ 的省级区划元素都不为空,则行政区划集合 DS 的交集为 $D_1, D_2, \cdots, D_m$ 相交,记为 $DI(D_1, D_2, \cdots, D_m)$,用式(7)表示:

$$DI(D_1, D_2, \cdots, D_m) = \cap DS = \cap (D_1, D_2, \cdots, D_m) = D_1 \cap D_2 \cap \cdots \cap D_m \tag{7}$$

其中:$\cap DS$ 表示集合 DS 里面的元素相交。

3)多个行政区划集合的交集运算。

多个行政区划集合的交为多个行政区划集合分别两两相交结果的交,记为 DSI,用式(8)表示:

$$DSI = (DS_1, DS_2, \cdots, DS_n) = \{\{DS_1 \cap DS_2\}, \{DS_1 \cap DS_3\}, \cdots, \{DS_1 \cap DS_n\}, \{DS_2 \cap DS_3\}, \cdots, \{DS_2 \cap DS_n\}, \cdots, \{DS_{n-1} \cap DS_n\}\} \tag{8}$$

2.1.2 条件集合运算

由于中文地址的混乱和无序性,会有多个集合运算中行政区划得出的结果集没有任何关联的可能,导致集合运算的结果为空集。如果式(8)中多个行政区划集合的交集运算结果为空,即 $DSI(DS_1, DS_2, \cdots, DS_n) = \varnothing$,则会造成地址的行政区划信息的丢失。为了避免行政区划信息的丢失,本文提出一种条件集合运算。

当 $DSI(DS_1, DS_2, \cdots, DS_n) = \varnothing$ 时,将行政区划的交运算变成并运算,即 $DSI(DS_1, DS_2, \cdots, DS_n) \rightarrow DSU(DS_1, DS_2, \cdots, DS_n)$,用式(9)表示:

$$DSI(DS_1, DS_2, \cdots, DS_n) \rightarrow \cup DSU(DS_1, DS_2, \cdots, DS_n) =$$

$$DS_1 \cup DS_2 \cup,\cdots, \cup DS_n = \cup \begin{pmatrix} D_{11},D_{12},\cdots,D_{1p} \\ D_{21},D_{22},\cdots,D_{2q} \\ \vdots \\ D_{n1},D_{n2},\cdots,D_{nm} \end{pmatrix} = \{D_{11},D_{12},\cdots,D_{1p}\} \cup \{D_{21},D_{22},\cdots,D_{2q}\} \cup,\cdots, \cup \{D_{n1},D_{n2},\cdots,D_{nm}\} \quad (9)$$

2.2　行政区划可信度

当集合运算的结果依然是一个集合时,为提取出这个集合中最正确最完整并与原中文地址最为匹配的行政区划,本文提出行政区划可信度计算。行政区划可信度是根据移动窗口算法中完全匹配与部分匹配规则与行政区划的层次关系建立一个规则,计算集合中每个行政区划的可信度,选取可信度最大的行政区划作为最终提取结果。

完全匹配就是将中文地址中的行政区划字符串与得出的行政区划集合中的行政区划进行匹配,每个字符都全部匹配。部分匹配是指中文地址中的行政区字符串与得出的行政区划集合中的行政区划进行匹配,只能匹配出除去“省”“市”“区”“县”“乡”“村”特征词外的部分。

完全匹配度用 a 表示,部分匹配度用 p 表示,完全匹配的概率大于部分匹配的概率,且均为正数,即 $0 < p < a$。由于中文地址的行政区划之间具有层级关系,区划范围是逐级递减的,而由于部分匹配是根据关键字进行匹配,容易将以行政区划命名的路街匹配成行政区划,形成干扰结果,因而考虑到完全匹配以及部分匹配因素以及行政区划的层级关系,用几何级数 x 表示,则省、市、县、乡、村分别表示为 x、x^2、x^3、x^4、x^5,可信值用 R 表示,可信度用 Re 表示,集合中的任意一个行政区划的可信度为此行政区划的可信值与集合中所有行政区划可信值的和的比,则行政区划可信值与可信度计算如式(10):

$$\begin{cases} R = k_1x + k_2x^2 + k_3x^3 + k_4x^4 + k_5x^5 \\ Re = R_i / \sum_{i=1} R_i \end{cases} \quad (10)$$

其中:$k_n = \begin{cases} a, & \text{完全匹配} \\ p, & \text{部分匹配;} n = 1,2,3,4,5,\text{此处给出一个集} \\ 0, & \text{不匹配} \end{cases}$

合中行政区划可信度比较规则:

1) 行政区划中全部是完全匹配的行政区划的可信度最大;全部是部分匹配的行政区划可信度最小;

2) 两个行政区划中级数大的是完全匹配的行政区划可信度大;

通过可信度比较规则,可得出 a,p,x 的关系为 $x > a/p$,由于完全匹配是指字符串全部匹配,不妨设完全匹配概率为 1,部分匹配概率为 0.6,即 $a = 1$,$p = 0.6$,则 $x > 5/3$,取 $x = 2$。

2.3　“路”特征词分组

由于中文地址中路街名称大量使用行政区划的名称来命名,比如“洪山园路,“洪山”是“洪山区”的简称,在对行政区划进行移动窗口匹配时容易把街道匹配成行政区划,从而对下一步可信度计算造成干扰。为了提高行政区划的准确率,本文对地址中的路街名称过滤。路街名称的一般命名规则是“名称 + 路街特征词”。常用的特征词有“路”“街”“大街”“道”“大道”等。地址中的行政区划一般位于路街名称的前面,将中文地址以路街特征为参照分组,取第一个分组。然后截取第一个分组前半部分作为计算地址行政区划的地址字符串,匹配行政区划元素词。

2.4　非规范中文地址行政区划提取算法

对于输入的中文地址,本算法先对原数据进行“路”特征词分组预处理,再根据基于移动窗口算法的地址匹配对地址进行匹配,返回中文地址中所有可能的行政区划结果集,然后进行集合运算,最后对集合运算出的结果进行可信度计算,解析出可信度最大的中文地址的行政区划。非规范中文地址的行政区划提取算法的流程如图 1。

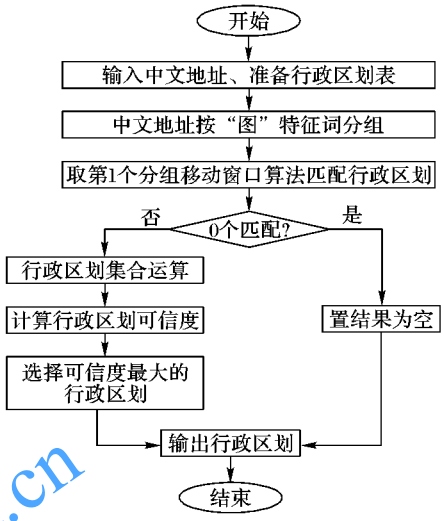

图 1　非规范中文地址的行政区划提取算法流程

Fig. 1　Process for extracting administrative division of non-normalized Chinese address

基于移动窗口匹配算法的地址匹配方法,首先建立用于行政区划匹配字典,然后根据地址数据表达的语义特点,建立行政区划的匹配规则,将字符串中的字符比作一个可滑动的滑动窗口对行政区划表进行匹配查询,返回对应的行政区划结果集,包含与该行政区划匹配父行政区划,直到省级,从而得到所有可能的行政区划。

非规范中文地址的行政区划提取算法步骤如下:

输入:原始中文地址;

输出:完整的行政区划地址。

步骤 1　读入行政区划表。

步骤 2　对原始数据进行“路”词分组预处理,取第一个分组,若 0 个匹配,则置结果为空,直接输出。

步骤 3　利用移动窗口算法对行政区划表进行匹配查询,根据分组后地址中文地址中包含的行政区划元素词匹配出这个地址字符串所包含的可能行政区划结果集 DS。

步骤 4　判断行政区划集合个数,分为以下三种情况:

若 DS 仅仅是一个行政区划,则直接输出。

若 DS 是一个集合,则转到步骤 5。

若 DS 是多个行政区划集合,则转到步骤 6。

步骤 5　利用式(1) ~ (7) 进行 1 个行政区划集合的交集运算得到 DI,转到步骤 7。

步骤 6　利用式(8) 进行多个行政区划集合的交集运算得到 DSI,当 $DSI = \varnothing$ 时,利用式(9) 进行多个行政区划结合的条件交集运算得到新的 DSI。

步骤 7　利用式(10) 对集合运算的结果进行可信度计算,选择可信度大的行政区划。

步骤 8　输出行政区划结果。

3 实验设计与分析

3.1 实验设计

为了验证本算法的有效性,本文做了以下准备工作:

准备一个行政区划字典,该字典是规范的表达。

给定一个中文地址,该地址没有其他参考信息,如邮编、电话区号等。

以地址“福州鼓楼洪山园路”为例,该地址存在以下几方面问题:1)该地址的行政区划部分不完整且没有规律;2)该地址中的地址要素残缺,无法推出完整地址;3)该地址不是按照省、市、县的规则形成的,无法使用一般的中文地址匹配方法进行匹配;4)该地址的路名中包含行政区划名称。由此可见,该地址存在要素残缺和语义模糊等问题,具有代表性。

首先对该地址进行“路”特征词分组得到“福州鼓楼”,在通过移动窗口匹配查询得到所有可能的行政区划集合,然后进行集合运算,再进行利用可信度公式进行可信度计算,提取出可信度最大的行政区划,集合运算结果及可信度如表 2。

表 2 行政区划可信度

Tab. 2 Credibility of administrative division

行政区划	可信度	行政区划	可信度
江苏省,南京市,鼓楼区	0.22	福建省,福州市,鼓楼区	0.33
江苏省,徐州市,鼓楼区	0.22	河南省,开封市,鼓楼区	0.22

可以看出,“福建省,福州市,鼓楼区”的可信度最大,因此选取此地址作为行政区划结果。

3.2 实验分析

本文利用网络爬虫从互联网上提取约 25 万条地址数据进行中文地址行政区划匹配实验。从三个方面验证实验:1)通过对本文算法的数据的预处理过程,比较不同处理方法对实验结果的影响,从而选取最佳方案。2)通过加入可信度计算,比较加入可信度对实验结果的影响。3)通过对比分析不同的算法来验证本文算法的有效性。

3.2.1 “路”特征词分组处理

本文实验对于中文地址的预处理计算分为直接地址处理和“路”特征词分组地址处理 2 种。直接地址是将原始地址作为计算的字符串直接用于匹配计算。分组地址是依据中文地址中行政区划表示的特点选取路街前面的地址部分进行“路”特征分组处理后作为行政区划匹配计算的字符串。本文将直接地址、分组地址与完全匹配查询(F)、完全匹配查询+部分匹配查询(P)进行组合,进行实验,实验结果如表 3。

根据上述实验数据,从两个方面进行分析,首先从选择完全区划匹配以及选择完全+部分行政区划匹配方面分析。在正确率方面,由表 3 可以看出,对原始数据,选择完全区划匹配的正确率高于选择完全+部分区划匹配,是因为完全+部分查询匹配是可以对关键字进行匹配,例如“南京鼓楼区上海路”,直接进行完全+部分查询匹配,会将“上海”匹配成“上海市”,将以行政区划命名的道路匹配成行政区划,导致结果错误,从而降低正确率。在时间消耗方面,选择完全区划匹配查询的时效要远高于选择完全+部分区划匹配,是由于完全+部分匹配查询对关键字匹配,查询次数比选择完全匹配查询次数多,导致消耗的时间多。

从选择“路”特征词分组处理以及不选择“路”特征词分组处理方面分析,由表 3 可以看出,采用完全行政区划匹配方法时正确率和时效不受“路”特征词分组处理的影响基本维持不变,因为行政区划完全匹配方法已经把所有用行政区划命名的路街全都过滤掉了,但对于一些中文地址中行政区划区划省略了特征词“省市县”的也过滤了。所以,无论是否对原始地址进行“路”特征词分组,对计算结果没有太大的影响。而选择完全匹配查询+部分匹配查询测试的地址字符串进行“路”特征词分组处理后正确率有明显的提升,大约提升了 20%,达到 93.51%,是因为省略了地址中的道路、街道等,匹配时只需要匹配行政区划,避免了将道路、街道名匹配成行政区划,所以时效和正确率都有明显的提升,但是有些以“道”“街”等特征词命名的行政区划经过“路”特征词分组处理后也被过滤掉了,导致解析正确率无法到达 100%。

表 3 “路”特征词分组对比

Tab. 3 “Road” feature word grouping comparison

地址	总数	完全区划匹配			完全+部分区划匹配		
		正确数	正确率/%	时耗/ms	正确数	正确率/%	时耗/ms
原始	254459	210977	82.91	12579	187423	73.66	57347
“路”特征词分组	254459	213604	83.94	12001	237221	93.51	10872

3.2.2 可信度计算

由于行政区划集合运算得到的结果有可能是集合形式,无法得到确切的行政区划,本节实验从两个方面进行:一方面选择最后求得的集合中最大非空行政区划元素作为计算结果,一方面对集合运算的结果作可信度计算,选择可信度最大的行政区划,实验结果如表 4。

表 4 可信度对比

Tab. 4 Credibility contrast

地址	总数	完全区划匹配				完全+部分区划匹配			
		选择最大非空元素		选择可信度最大元素		选择最大非空元素		选择可信度最大元素	
		时耗/ms	正确率/%	时耗/ms	正确率/%	时耗/ms	正确率/%	时耗/ms	正确率/%
原始	254459	6453	82.91	6329	83.07	32048	73.66	32682	78.69
“路”特征词分组	254459	5469	82.91	5219	83.07	8766	93.23	8594	93.51

由表 4 可以看出,在正确率方面,选择完全区划匹配查询时,对数据进行“路”特征词分组处理与不进行处理后,选择最大非空区划元素或者选择可信度最大的行政区划元素作为结果的正确率并没有发生变化,因为完全区划匹配就是对行政区中的最大元素进行匹配,所以选择可信度最大的或者选择最大非空行政区划元素得到的结果相同,因此对结果无影响;同时可以看出,选择完全+部分区划匹配查询,对结果进行可信度处理的正确率是要高于选择最大非空行政区划元素处理的正确率,对原始数据而言,由于完全+部分匹配查询是对关键字进行匹配查询,会匹配出干扰行政区划,比如“上海路”会匹配成“上海市”,对结果的选择造成影响,而选择可信度最大的行政区划作为结果是对完全匹配以及部分匹配以及

行政区划层次结构等因素进行考虑后得出的结果,所以得出的结果正确率高于单纯选择最大行政区划的结果,并且选择“路”特征词分组处理 + 完全 + 部分匹配查询 + 可信度处理,能够使正确率提高到93.51%,是由于“路”特征词分组处理省略了地址中的道路、街道等,匹配时只需要匹配行政区划,避免了将道路、街道名匹配成行政区划。时间消耗是受完全匹配查询或者完全 + 部分匹配查询以及是否进行“路”特征词分组的影响,上一节已进行分析。

3.2.3　算法对比

通过分析中文地址解析在各种算法中的应用,将采用基于分级地名库的中文地理编码[11]、基于分词的地址匹配技术[12]、基于规则的中文地址要素解析方法[14]与本文算法进行对比。基于分级地名库的中文地理编码是通过 TRIE 树词典对地址要素字段创建索引,地址匹配的过程就是在每个级别的 TRIE 索引树中查询最大地址要素的过程。基于分词的地址匹配技术是建立地址要素词库,采用基于“正向最大匹配分词”的地址分词算法对中文地址进行切分。基于规则的中文地址要素解析方法通过系统分析地址要素的构词特征和句法模式,构建了各类地址要素的特征字库,提出中文地址数字表达方法,设计了基于规则的中文地址要素解析方法。本文从解析的正确率与效率对四种算法进行了比较,算法对比表如表5。

表 5　不同算法的处理效率与正确率比较

Tab. 5　Processing efficiency and correctness comparison of different algorithms

算法	效率/(ms·条$^{-1}$)	正确率/%
基于分级地名库的地址匹配算法	0.039	89.21
基于分词的地址匹配技术	0.038	85.63
基于规则的中文地址要素解析方法	0.220	83.23
本文算法	0.043	93.51

由于用来实验的中文地址数据来源于互联网,大部分是特征字模糊、顺序混乱的非规范地址。根据实验结果可以看出基于规则的中文地址要素解析方法与基于分词的地址匹配技术的正确率相差不大,原因是基于分词的地址匹配技术是基于地址要素的词典进行切分,而基于规则的地址要素解析方法是基于特征字库设计的规则与算法,它们全部要求地址是完全匹配才能匹配准确,只对规范的特征字明显的中文地址有作用,对非规范的中文地址解析的正确率不高。基于分级地名库的地址匹配算法虽然可以同时进行模糊匹配和完全匹配,但是最终的匹配结果可能有多个,无法对最后得出的集合进行计算得出确定的地址,需要人工选择准确的地址。本文算法不仅可以能够对非规范的地址匹配查询出完整的行政区划集合,且当返回结果有多个时,可以利用集合运算计算出最准确的地址,增加了正确率。在效率方面,由于基于分级地名库的地址匹配算法、基于分词的地址匹配技术和本文算法都是利用中文地址具有层次结构特征构建的层级式词典进行查询匹配,所以效率高。通过实验对比,可以看出,本文算法在正确率上具有极大优势,且具有高效率,证明了本算法的有效性。

3.2.4　数据分析

根据上述“路”特征词分组处理实验,以及算法对比分析实验,可以看出,在中文地址行政区划解析方面,影响中文地址解析效率的方面有以下三点:“路”特征词、尾特征词、词典结构以及可信度计算。

在时效上,影响速率的第一个因素是“路”特征词。根据本文在对原始地址进行“路”特征词处理的实验中,由表 3 可以看出将地址进行了“路”特征词处理之后,解析速度明显提高,因为对地址进行“路”特征词处理后,过滤掉了地址中的道路、街道等信息,只对行政区划进行解析,大幅度提高了中文地址行政区划的解析速率。而在本文方法与基于分级地名库的中文地理编码、基于分词的地址匹配技术以及基于规则的中文地址要素解析方法的对比实验中可以看出,影响速率的第二个因素是词典结构,在本文方法与其他三个算法的对比实验中,利用中文地址的层次结构特征,对有从属关系的地址建立逐级的父子关系,从而建立起层级式词典,减少了查询次数,加快了地址的查询匹配速率。

在正确率上,第一个影响因素是“路”特征词。在表 3 中,虽然进行“路”特征词分组处理后,由于省略了道路、街道等信息,去掉了干扰项,正确率有一定提升,但是由于一些行政区划是以“路”特征词命名,导致有用信息被省略,从而导致正确率无法达到 100%。第二个影响因素是尾特征词。同样在表 3 中可以看出,进行完全匹配查询 + 部分匹配查询的正确率高于只采用完全匹配查询,由于实验数据来源于互联网,大多数是非规范的中文地址,很多地址缺乏关键字,而完全匹配查询是依赖于中文地址的规范性,依赖于关键字全部匹配,而部分匹配查询,可以有效避免缺乏尾特征词的非规范地址匹配不上,因此增大了正确率。而在表 5 中,同样可以看出依赖于尾特征词的基于规则的中文地址要素解析方法正确率较低,对非规范中文地址的解析准确率不高。第三个因素是可信度计算,从表 4 可以看出,在行政区划集合运算得出的结果是一个集合时,选择可信度大的计算结果的正确率比选择最大非空行政区划元素的结果正确率高。

4　结语

在目前无法用一般分词匹配算法匹配出正确的行政区划的情况下,本文提出了一种非规范中文地址的行政区划提取算法,本算法利用基于移动窗口算法的地址匹配算法,并顾及中文地址的语义,根据中文地址的表达特点,建立行政区划集合运算规则和可信度计算规则,提高了对中文地址行政区划解析的正确率和时效。本算法提出了一种对中文地址数据进行预处理的方法——“路”特征词分组处理,能够过滤掉干扰中文地址行政区划解析的路街信息,使中文地址行政区划解析的效率得到很大的提高。本算法还提出的行政区划条件集合运算和可信度计算,能够便捷地处理多个行政区划集合并解析出最完整、最准确的行政区划信息,不造成地址信息丢失,本算法不依赖于地址来源,对非规范的中文地址也能进行行政区划信息的提取,在性能上具有明显的优越性,因此,本算法在地理位置服务中具有实用性。

但是该算法还有一定的缺陷,在进行“路”特征词分组处理时,由于中文地址中有一部分行政区划以“道”“路”等特征词命名,如“哈尔滨道里区”,按照“路”特征词分组后会将“道里区”过滤掉,导致解析结果错误。还有中文地址中一些乡镇的名称与行政区划名称相同也会产生错误的结果,因此在未来的工作中,将处理更加复杂的、辨别度不高的非规范中文

地址,改进算法,从而设计出适应各种不同类型地址的算法。

参考文献(References)

[1] 李生. 自然语言处理的研究与发展[J]. 燕山大学学报, 2013, 37(5): 377 - 384. (LI S. Research and development of natural language processing [J]. Journal of Yanshan University, 2013, 37(5): 377 - 384.)

[2] 吕雅娟, 赵铁军, 杨沐昀, 等. 基于分解与动态规划策略的汉语未登录词识别[J]. 中文信息学报, 2001, 15(1): 28 - 33. (LYU Y J, ZHAO T J, YANG M J, et al. Leveled unknown Chinese words resolution by dynamic programing [J]. Journal of Chinese Information Processing, 2001, 15(1): 28 - 33.)

[3] 李庆虎, 陈玉健, 孙家广. 一种中文分词词典新机制——双字哈希机制[J]. 中文信息学报, 2003, 17(4): 13 - 18. (LI Q H, CHEN Y J, SUN J G. A new dictionary mechanism for Chinese word segmentation [J]. Journal of Chinese Information Processing, 2003, 17(4): 13 - 18.)

[4] 于光. 中文分词系统的设计与实现[D]. 成都: 电子科技大学, 2012: 73. (YU G. Design and implementation of Chinese word segmentation system [D]. Chengdu: University of Electronic Science and Technology of China, 2012: 73.)

[5] 郭会, 宋关福, 马柳青, 等. 地理编码系统设计与实现[J]. 计算机工程, 2009, 35(1): 250 - 252. (GUO H, SONG G F, MA L Q, et al. Design and implementation of address geocoding system [J]. Computer Engineering, 2009, 35(1): 250 - 252.)

[6] 郭文龙. 基于 SNM 算法的大数据量中文地址清洗方法[J]. 计算机工程与应用, 2014, 50(5): 108 - 111. (GUO W L. Cleaning approach to large amounts of Chinese address based on SNM algorithm [J]. Computer Engineering and Applications, 2014, 50(5): 108 - 111.)

[7] 徐娟, 曹晔, 张奇. 面向自由文本的中文地址规范化[J]. 计算机应用与软件, 2015, 32(8): 22 - 24. (XU J, CAO Y, ZHANG Q. Chinese address standardisation for plain text [J]. Computer Applications and Software, 2015, 32(8): 22 - 24.)

[8] 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. (CHEN X Q, CHI Z X, JIN N. Application and study of city geocoding system [J]. Computer Engineering, 2004, 30(23): 50 - 52.)

[9] 宋子辉. 自然语言理解的中文地址匹配算法[J]. 遥感学报, 2013, 17(4): 788 - 801. (SONG Z H. Address matching algorithm based on Chinese natural language understanding [J]. Journal of Remote Sensing, 2013, 17(4): 788 - 801.)

[10] 赵阳阳, 王亮, 仇阿根. 地址要素识别机制的地名地址分词算法[J]. 测绘科学, 2013, 38(5): 74 - 76. (ZHAO Y Y, WANG L, QIU A G. An improved algorithm for address segmentation [J]. Science of Surveying and Mapping, 2013, 38(5): 74 - 76.)

[11] 孙存群, 周顺平, 杨林. 基于分级地名库的中文地理编码[J]. 计算机应用, 2010, 30(7): 1953 - 1955. (SUN C Q, ZHOU S P, YANG L. Chinese geo-coding based on classification database of geographical names [J]. Journal of Computer Applications, 2010, 30(7): 1953 - 1955.)

[12] 孙亚夫, 陈文斌. 基于分词的地址匹配技术[EB/OL]. [2016-01-05]. http://xueshu.baidu.com/s?wd=paperuri%3A%284105a7e9cf9ea8588730d99199975503%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fcpfd.cnki.com.cn%2FArticle%2FCPFDTOTAL-DLXX200711001019.htm&ie=utf-8&sc_us=16495669320387933132. (SUN Y F, CHEN W B. Address matching technology based on segmentation [EB/OL]. [2016-01-05]. http://xueshu.baidu.com/s?wd=paperuri%3A%284105a7e9cf9ea8588730d99199975503%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fcpfd.cnki.com.cn%2FArticle%2FCPFDTOTAL-DLXX200711001019.htm&ie=utf-8&sc_us=16495669320387933132.)

[13] 程昌秀, 于滨. 一种基于规则的模糊中文地址分词匹配方法[J]. 地理与地理信息科学, 2011, 27(3): 26 - 29. (CHENG C X, YU B. A rule-based segmenting and matching method for fuzzy Chinese addresses [J]. Geography and Geo-Information Science, 2011, 27(3): 26 - 29.)

[14] 张雪英, 闫国年, 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. (ZHANG X Y, LYU G N, LI B Q, et al. Rule-based approach to semantic resolution of Chinese addresses [J]. Journal of Geo-Information Science, 2010, 12(1): 9 - 16.)

[15] 唐静. 城市地名地址的编码匹配研究[D]. 昆明: 昆明理工大学, 2011: 76. (TANG J. Study on city names address matches the encoding [D]. Kunming: Kunming University of Science and Technology, 2011: 76.)

[16] 段艳会, 李晓林, 黄爽. 基于条件随机场的中文地址行政区划提取方法[J]. 武汉工程大学学报, 2015, 37(11): 47 - 51. (DUAN Y H, LI X L, HUANG S. Extraction of administrative division of Chinese address based on conditional random fields [J]. Journal of Wuhan Institute of Technology, 2015, 37(11): 47 - 51.)

[17] 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. (MA Z T, LI Z G, SUN W, et al. An automatic geocoding algorithm based on address segmentation [J]. Bulletin of Surveying and Mapping, 2011(2): 59 - 62.)

[18] GUO H, ZHU H, GUO Z, et al. Address standardization with latent semantic association [C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009: 1155 - 1164.

[19] COLDBERG 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.

This work is partially supported by Special Plan of Surveying and Mapping Geographic Information Public Welfare Scientific Research Special Industry (201412014), the National High Technology Research and Development Program (863 Program) (2013AA12A202), the Natural Science Foundation of Hubei Province (2013CFA125), the 7th Graduate Student Innovation Fund Projects of Wuhan Institute of Technology (CX2015053).

LI Xiaolin, born in 1962, M. S., associate professor. His research interests include data mining, machine learning, artificial intelligence.

HUANG Shuang, born in 1992, M. S. candidate. Her research interests include data mining, machine learning, artificial intelligence.

LU Tao, born in 1980, Ph. D., associate professor. His research interests include image/visual processing, computer vision, artificial intelligence.

LI Lin, born in 1960, Ph. D., professor. His research interests include geo-semantics and ontology, three-dimensional modeling and visualization.