Normalizing Chinese Address for Internet Applications

Xiaolin Li[1, 2, a], Shuang Huang[1, 2, b], Tao Lu[1, 2, c] and Deng Chen[1, 2, d]

[1]Wuhan Institute of Technology, Wuhan and 430205, China

[2] Hubei Province Key Laboratory of Intelligent Robot, Wuhan and 430205, China

[a]lxl989898@163.com, [b]13986287758@163.com, [c]lut@wit.edu.cn, [b]dChen@wit.edu.cn

Keywords: Set operation; Administrative division; Chinese address; Moving window; Matching degree; Analytical rules

Abstract. Many Internet applications take addresses as input. However, addresses on the Internet are always non-normalized, which cannot be used directly. In this paper, we propose an Administrative Divisions Extracting Algorithm to normalize Chinese addresses on the Internet. Our approach proceeds as follows: 1) It began with the “Road” feature words processing and extracted all possible administrative divisions data set from Chinese addresses by using administrative divisions dictionary and Moving Window Algorithm. 2) According to the Chinese administrative divisions has the characteristics of hierarchical relationships between elements, the algorithm established the conditions set operations rules of administrative divisions, it carried on the set operations to administrative divisions data set. 3) The algorithm obtained Chinese address administrative divisions of the most integrity. In order to investigate the feasibility and effectiveness of our approach, we performed experiments that the paper verified the availability of whether to adopt the “road” feature words processing for about 250 thousands Chinese address data extracted from the internet. At the same time, the algorithm compared with the current address matching technology. Experimental results show that the accuracy reached 93.51%.

Introduction

In the Internet location-based services, location can have a variety of forms of expression, Chinese text expression is one of them, and users can through the Chinese address information to obtain the exact address they want, which is better for improving the quality of service. With geographic information systems playing more and more important role in people's lives, the demand of quickly and accurately find the geographical coordinates according to the Chinese text of address information has become increasingly apparent [1].

A specification of Chinese address should contains full of administrative divisions, and in accordance with administrative divisions (province/city/county/town/village), road street, number, building, and the order of the households room to express[2], it has obvious features of the word. However, on the Internet, Chinese address frequently used address non-normative way to describe the Administrative Division, expressed confusion and ambiguity, it is difficult to determine the location of the address expression, as the location is not a valid [3]. Therefore, ordinary Chinese address segmentation algorithm could not be a good solution to handle issues of non-normalized Chinese address, there need research on an optimized algorithm of Chinese address analysis algorithm to resolve the non- normalized address. Mixed in the non-normative information on the Internet, identifying the high credibility of the information users need becomes very necessary in today's geographic location information services [4].

Therefore the paper provides an administrative divisions extracting algorithm for non-normalized Chinese addresses. It dealt address data with "Road" feature word grouping processing. According to the Chinese administrative divisions has the characteristics of hierarchical relationships between elements, the algorithm established the conditions set operations rules of administrative divisions, it carried on the set operations to administrative divisions data set. The algorithm obtained Chinese address administrative divisions of the most integrity. It can effectively improve the speed and accuracy of address data lookup, thus improving the quality of online services network map, better positioning for the user.

Our technique

General approach. General intersection operation means each level of administrative division of elements compared for equality, and if elements are equal then the values of the elements of administrative division is taken, else the administrative elements of intersection operation results is empty. But at the intersection operation of the two administrative divisions, we cannot simply follow the equality of elements at all levels to determine the outcome of administrative divisions intersection, if we do it, the intersection results of administrative division is not the desired result. Such as table 1 (3-level administrative divisions, provinces, cities, counties).

Administrative divisions includes 5 levels of province, city, county, town and village, which can be expressed as $D.d_i, i=1,2,3,4,5$, it is on behalf of each element of administrative division, and $D=\{d_1,d_2,d_3,d_4,d_5\}$.

Table 1. The example of intersection operation of administrative division

The example	administrative division		The result $D_1 \cap D_2$	
1	D_1	Jiangsu Province, Xuzhou, Gulou	Intersection	Jiangsu Province, , Gulou
	D_2	Jiangsu Province, Nanjing, Gulou	Exception	Jiangsu Province, , Gulou
2	D_1	, , Gulou	Intersection	, , Gulou
	D_2	Jiangsu Province, Nanjing, Gulou	Exception	Jiangsu Province, Nanjing, Gulou

As can be seen in the example 2, when one of two administrative divisions lack two levels of administrative divisions, the intersection operation result is not the desired result, and the desired result is $D = D_1 \cap D_2 = \{\texttt{Jiangsu Province,Nanjing,Gulou}\}$.

To resolve that the general intersection algorithm is unable to conclude the expected result, the paper proposes a conditional set operation based on the general set operations.

General set operations. Common administrative divisions set operations are as follows:

（1） The intersection operation of two administrative divisions

If there are two administrative divisions $D_1=\{d_{11},d_{12},d_{13},d_{14},d_{15}\}$ and $D_2=\{d_{21},d_{22},d_{23},d_{24},d_{25}\}$, the intersection of administrative divisions is the intersection of elements of each level, it denoted as DI and represented by Eq. (1). The intersection of elements of two administrative divisions denoted as $dI_i, i=1,2,3,4,5$.

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

A containment relationship exists between the elements of administrative division, that is outside the provincial division and other divisions at all levels belong to 1 or n superior administrative divisions. By this reason, provincial administrative elements are calculated first, then computes the non-provincial division elements.

a) The intersection of provincial division elements

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

ρ is uncertain, it means that an administrative division in the provincial division of administrative division element is $\varnothing$. Administrative divisions of provincial zoning elements are empty at this time needs to use the dictionary to get the Administrative division Administrative division at the provincial division element is non-empty.

There is an administrative division $D=\{d_1,d_2,d_3,d_4,d_5\}$, in this Eq. , $d_1=\varnothing, \exists d_k \neq \varnothing, k=2,\ldots,5$. A division element d_k in D is selected, and it represented by Eq. (3).

$$d_k = \arg\min_k \left\{ d_k \,\middle|_{d_k \neq \varnothing} \right\} \quad (3)$$

$DS\left(d_k\right)$ is a set consisted by m administrative divisions which are queried from administrative divisions dictionary.

$$query\left(d_k\right)=DS\left(d_k\right)=\left\{\left\{d_{11},\ldots,d_{1k}\right\},\ldots,\left\{d_{m1},\ldots,d_{mk}\right\}\right\} \tag{4}$$

If $D=D_1$,

$$\mathrm{d}_{11}\cap d_{21}=\left\{wd_{11}\cap d_{21}\cup\ldots\cup wd_{m1}\cap d_{21}\right\} \tag{5}$$

b) The intersection operation of one administrative division set

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

When elements are equal, the intersection results are the division elements;

When elements are not equal, and elements are not null, the result is null;

When elements are not equal and one of them is null, as well as if the parent element is not empty $\left(\exists dI_j\neq\varnothing\right)$, the result is the non-null element values.

（2） The intersection operation of one administrative division set

There is administrative division set $DS=\left(D_1,D_2,\ldots,D_m\right)$, and the provincial division elements are not null. The intersection of DS is $DI\left(D_1,D_2,\ldots,D_m\right)$ which represented by Eq. (7).

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

（3） The intersection operation of multiple administrative division sets

The intersection of multiple administrative division sets is every two sets of multiple administrative divisions respectively.

$$DSI=\left(DS_1,DS_2,\ldots,DS_n\right)=\left\{\begin{array}{c} \left\{DS_1\cap DS_2\right\},\left\{DS_1\cap DS_3\right\}, \\ \cdots,\left\{DS_1\cap DS_n\right\}, \\ \left\{DS_2\cap DS_3\right\},\cdots\left\{DS_2\cap DS_n\right\}, \\ \cdots,\left\{DS_{n-1}\cap DS_n\right\} \end{array}\right\} \tag{8}$$

Conditional set operation. Due to the confusion and disorder in Chinese address, the result sets of multiple set operations will not have any associated probably, that will cause the result of the set operation is null set. If the results of multiple administrative division intersection operation is null in the Eq. (8), that is $DSI\left(DS_1,DS_2,\ldots,DS_n\right)=\varnothing$， it will be in loss of administrative information. In order to avoid the loss of administrative information, the article proposes a conditional set operation.

When $DSI\left(DS_1,DS_2,\ldots,DS_n\right)=\varnothing$, the intersection operation of administrative divisions will turn into union operation. That is $DSI\left(DS_1,DS_2,\ldots,DS_n\right)\rightarrow DSU\left(DS_1,DS_2,\ldots,DS_n\right)$ and represented as Eq. (9).

$$DSI\left(DS_1,DS_2,\ldots,DS_n\right)\rightarrow\cup DSU\left(DS_1,DS_2,\ldots,DS_n\right)=DS_1\cup DS_2\cup,\ldots,\cup DS_n=\cup\begin{pmatrix} D_{11},D_{12},\ldots,D_{1p} \\ D_{21},D_{22},\ldots,D_{2q} \\ \cdots \\ D_{n2},\ldots,D_{nm} \end{pmatrix}=\begin{array}{c}\left\{D_{11},D_{12},\ldots,D_{1p}\right\}\cup\left\{D_{21},D_{22},\ldots,D_{2q}\right\}\cup \\ ,\ldots,\cup\left\{D_{n1},D_{n2},\ldots,D_{nm}\right\}\end{array} \tag{9}$$

Experiments

In this paper, we used about 250,000 address data extracted from the Internet by the web crawler to carry out the Chinese address administrative division matching experiment.

"Road" feature word grouping processing. In order to enhance the accuracy in the administrative division, in this paper, we filter the road name of the address. The general naming rules of streets name are "name + road feature words". Chinese address is grouped according to the characteristics of the road, and we take the first group. Then intercepting the first half of the first group as the address string which is used to calculate the address administrative division and matching the element words of the administrative divisions.

Therefore, in this paper, the preprocessing of Chinese address is divided into direct address processing and "Road" feature word grouping address processing. In this paper, direct address, group address and perfectly matching query (F), perfectly matching query + partial matching query (P) are combined to carry out the experiment. Experimental results are shown in Table 2.

Table 2. The "Road" feature word grouping comparison table

Address	The total number	Perfectly administrative Division match			Perfectly + Partially administrative Division matches		
		Number of correct	Accuracy /%	Time consumption /ms	Number of correct	Accuracy /%	Time consumption /ms
The original	254459	210977	82.91%	12,579	187423	73.66%	57,347
“Road” Feature grouping	254459	213604	83.94%	12,001	237221	93.51%	10,872

On the accuracy rate, it can be seen from Table 2 that because perfectly + partially matching query is to match the keywords , for the original data, the correct rate of choosing perfectly administrative division matching is higher than perfectly + partially administrative division matching. On the time consumption, the temporality of choosing perfectly administrative division matching query is much higher than perfectly + partially administrative division matching query, the reason is perfectly + partially matching query is keywords matching and the number of queries is much more than perfectly matching query, so it leads to more time consumption.

Analyzed from the aspect of choosing or not choosing the "Road" feature word grouping processing, It can be seen from Table 2 that the accuracy and time efficiency are not affected by the "Road" feature word grouping process when using the perfectly administrative division matching method, basically had no variation. And after the address string has tested by selecting the perfectly matching query + partially matching query being the "Road" feature word grouping processed, the accuracy has improved significantly, about 20%, reached 93.51%.

Algorithm comparison. By analyzing the application of Chinese address resolution in a variety of algorithms, the algorithm in the paper compared with Chinese geo-coding based on classification database of geographical names[5], Address matching technology based on segmentation[6], Rule-based approach to semantic resolution of Chinese addresses[7]. In this paper, the accuracy and efficiency of the four algorithms are compared, and algorithm comparison table is in table 3.

Table 3. Algorithm comparison table

Algorithm	Efficiency (ms/Piece)	Accuracy
Algorithm 1: Address matching based on classification database of geographical names	0.039	89.21%
Algorithm 2: Address matching technology based on word segmentation	0.038	85.63%
Algorithm 3: Rule-based approach to semantic resolution of Chinese addresses	0.22	83.23%
This algorithm	0.043	93.51%

According to the experimental results, we can see that the accuracy of algorithm 3 and algorithm 2 are not quite different. The reason is that algorithm 2 is based on the dictionary of address elements for segmentation, and algorithm 3 designed rules and algorithm based on the characteristics word library. Algorithm 1 can be simultaneously fuzzy matching and perfectly matching. But the final matching result may be more than one. In this paper, the algorithm not only be able to match and query a complete set of administrative divisions for the non-normalized address, but also when the results are more than one, the algorithm can use the set operation to calculate the most accurate address, which increase the accuracy rate. In terms of efficiency, these three algorithms and the algorithm of this paper are building the hierarchy dictionary by using hierarchical structure features of Chinese addresses to match and query, so the efficiency is fast. Through experimental comparison, we can see that the algorithm has a great advantage in the accuracy rate, and it has high efficiency, which proves the effectiveness of the algorithm.

Summary

Current, it is unable to get correct administrative divisions by using general word segmentation matching algorithms, an administrative divisions extracting algorithm for non-normalized Chinese addresses is proposed. This algorithm uses the address matching algorithm based on moving window algorithm, and takes address semantics into account. Based on the characteristics of address expression, the algorithm established set operations rules of administrative divisions, it improved accuracy and timeliness of the analysis of Chinese address administrative divisions. This algorithm proposed a method for pretreatment of the address data--the "road" feature word grouping processing, which filtered out interference address street information in Chinese address administrative division analysis and improved the analytical efficiency of Chinese address administrative divisions. The algorithm also raised conditional set operation of administrative divisions, which can easily handle multiple administrative division sets and parse the most complete and the most accurate information, moreover, the algorithm can avoid the loss of address information. This algorithm does not depend on the source of address, it can extract the information of administrative divisions from non-normalized Chinese address, and it has obvious advantages in performance. Therefore, the algorithm is practical in location service.

Acknowledgements

This work is partially supported by Special plan of Surveying and Mapping Geographic Information Public Welfare Scientific Research Special Industry: The Key Technology of Development of Fundamental Geographical Information Ontology Databases and Its Demonstrations (201412014), the National High Technology Research and Development Program (863 Program) (2013AA12A202), the Natural Science Foundation of Hubei Province of China (2013CFA125), the grant of China Scholarship Council, the National Natural Science Foundation of China (61502354), the Natural Science Foundation of Hubei Province of China(2012FFA099, 2012FFA134, 2013CFA125, 2014CFA130, 2015CFB451), Youths Science Foundation of Wuhan Institute of Technology under Grant No. k201622, and the 7th Graduate Student Innovation Fund Projects for Wuhan University (CX2015053).

References

[1] GUO Hui, SONG Guan-fu, MA Liu-qing, et al. Design and Implementation of Address Geocoding System. Computer Engineering. 2009(01): 250-252.

[2] GUO Wenlong. Cleaning Approach to Large Amounts of Chinese Address Based on SNM Algorithm. Computer Engineering and Applications, 2014, 50(5): 108-111.

[3] Xu Juan, Cao Ye, Zhang Qi. Chinese Address Standardization for Plain Text. Computer Applications and Software. 2015(08): 22-24+93.

[4] CHEN Xiqian, CHI Zhongxian, JIN Ni. Application and Study of City Geocoding System. Computer Engineering, 2004(23): 50-52.

[5] SUN Cun-qun, ZHOU Shun-ping, YANG Lin. Chinese Geo-coding Based on Classification Database of Geographical Names. Journal of Computer Applications. 2010(07): 1953-1955+1958.

[6] Sun yafu, Chen Wenbin. Address Matching Technology Based on Segmentation, China association for geographic information system for the fourth time the member representative assembly and the 11th annual meeting. 2007: Beijing, China. 12th.

[7] ZHNAG Xueying, LV Guonian, LI Boqiu, et al. Rule-based Approach to Semantic Resolution of Chinese Addresses. Journal of Geo-Information Science, 2010. 12(1): 9-16.