Address matching algorithm based on chinese natural language understanding

SONG Zihui

State Key Laboratory of Remote Sensing Science, Jointly Sponsored by the Institute of Remote Sensing and Digital Earth of Chinese Academy of Science and Beijing Normal University, Beijing 100101, *China*

Abstract: Address matching algorithm that has broad application prospects is the core and key technology for location-based services. This paper analyzes the existing three major address matching algorithms which are the level based matching algorithm, the full-text search algorithm and the regular expression algorithm. An address matching algorithm based on Chinese natural l anguage understanding is proposed in this paper. The complete process of this new algorithm includes five parts as pretreatment, address parsing, address elements standardization, reasoning about address matching and matching registration. This paper focuses on address parsing and reasoning matching the two most important parts. The paper establishes a complete Chinese address matching algorithm based on natural language understanding. In the principle of Chinese segmentation and semantic reasoning in natural language understanding, the new algorithm achieves the goal to combine natural language understanding with address matching by processing Chinese address of unstructured format. To check the new algorithm, an address matching experimental system was d eveloped. The matching experiment using 1000 resident addresses of Puyang city, Henan province shows that the matching rate can be 95% or more and the accuracy rate is above 93%.

Key words: natural language understanding, address matching, address element, address parsing, Hidden Markov Model

CLC number: P208 **Document code:** A

Citation format: **Song Z H. 2013. Address matching algorithm based on chinese natural language understanding. Journal of Remote Sensing, 17(4): 788-801 [DOI: 10.11834/jrs.20132164]**

1 INTRODUCTION

Address information is closely related to human activities. Many departments such as statistical departments, the business organizations, the public security departments and so on collect and store a lot of information that contains the addresses. The r esult of US Census Bureau statistics shows that 80% of information systems in the government's management information systems contain address information (O'Reagan, 1987). Address matching technology links text message containing location with geographic coordinates, and integrates spatial information and socio-economic information, and provides supporting technologies for data analysis, positioning, mapping and visualization services. So address matching technology plays an indispensable role in national economic construction and people's life.

Strictly speaking, address matching technology is a process that text address is mapped into geographical coordinates(Daniel, 2007). The process constructs the relation between text position and geographical coordinates. Chinese address matching is a process that text address described by Chinese is mapped into geographical coordinates. Some foreign companies have launched the address matching products or services for example, ESRI's Geocoding, MapInfo's MapMarker, Google's Geocoder for the United States and Europe. Due to differences in national conditions, these products or services cannot meet the needs of China.

From the theoretical point of view, the major Chinese address matching algorithms exist the following three categories. (1) The first category is a level based model matching algorithm (Jiang, 2003; Wang, 2004; Guo, 2009; Sun,2010). (2) The second category is a full-text search matching algorithm (Sun, 2007). (3) The third category is a regular expression matching algorithm (Chen, 2004). The following analysis will discuss the problems of these three types of address matching algorithms. The first type of algorithm has the characteristics that address elements have a unique level of value. Because the level based matching algorithm requires the descending order of address elements, the first type algorithm can only match some special addresses. The second type of algorithm treats addresses as documents, and has the characteristics of high speed, high matching rate and low a ccuracy. The third type of algorithm treats addresses as normal string with the characteristics of low speed, high matching rate and low accuracy.

The use of the natural language understanding, artificial i ntelligence, is the main idea of this paper to create a new Chinese address matching algorithm. The essence of existing three

Received: 2012-05-16; **Accepted:** 2012-06-18; **Version of record first published:** 2012-06-25

Foundation: National High Technology Research and Development Program of China (863 program) (No. 2012AA12A401)

First author biography: SONG Zihui (1980—), male, Ph. D. candidate, he majors in spatial query and address matching. E-mail: szhmvp@gmail.com

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

types of address matching algorithms are based on string, key words or rules address matching technology, rather than based on semantic; so the three types of algorithms cannot achieve semantic matching. The three types of algorithms do not belong to nature-based language understanding of the Chinese address matching algorithm. In addition to these matching algorithm, researches also study the relation between natural language understanding and spatial information, such as the p roposed spatial information retrieval by natural language (Ma & Gong, 2003a, 2003b) , and the matching algorithm based on the agent's address (Hutchinson, 2010) . The characteristic of Hutchinson's algorithm is the introduction of artificial intelligence technology to the address matching process. The Hutchinson's algorithm is based on level based model which is not suitable for Chinese address matching. The combination of address matching with nature language understanding has not yet been found. This paper proposes a new algorithm that the natural language understanding is applied to the Chinese address matching.

2 CHINESE ADDRESSES MATCHING ALGORITHM BASED ON NATURAL LANGUAGE UNDERSTANDING

2.1 The address matching and the natural language understanding

The essence of natural language understanding is a computer processing of natural language parsing, so that the computer has the ability of the natural language understanding. Natural language understanding technology has not reached the level of human intelligence, and is limited to specific areas of shallow natural language understanding. But even the shallow natural language understanding could bring valuable assistance to many application areas.

Because most of the Chinese addresses are described by non-structural text, the essence of address matching is the association between address and spatial information. It is theoretically possible that natural language understanding method is applied to the address matching algorithm.

Every language has a specific syntax. Compared with the English, there is no separator between the words of Chinese, and Chinese belongs to ideographic. Processing Chinese is more d ifficult than English so that many methods for processing English are not suitable for Chinese.

Chinese natural language understanding methods include Chinese segmentation, semantic label, Syntactic analyses, semantic reasoning. Due to the lack of effective knowledge a bstract model, the technology of natural language understanding is generally related to the specific application scenarios. According to the particular scene of address matching, the four major aspects of natural language processing are implemented specially. The Chinese segmentation part solves address segmentation. The semantic label part solves soles labeling. The syntactic analyses part solves address type identification. The s emantic reasoning part solves spatial relation reasoning. Through these targeted methods, this paper achieves the goal to combine natural language understanding and address matching application. As a result this paper designs a Chinese address matching algorithm based on natural language understanding.

2.2 Chinese Matching algorithm based on natural language understanding and process

In this paper, the Chinese address matching algorithm based on natural language understanding process is shown in Fig. 1. The complete algorithm processes include five parts as pretreatment, address parsing, address element standardization, reasoning about address matching, and matching registration. First, the a ddress data to be matched is preprocessed to simple code and filter information. The second step of processing the address p erforms address parsing to convert unstructured address into a structured representation. The third step does address elements standardization to convert non-standard address elements into the standard address elements. The fourth step does reasoning about address matching to get geograhpical coordinates by using a knowledge base. The last step is matching registration for matching results to quantify results matching.

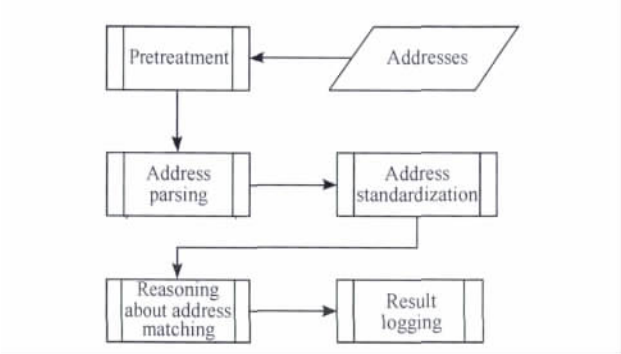

Fig. 1 Chinese address matching based on nuture language understanding

In the process of the algorithm, the pretreatment is ordinary procedure to process string. The standardization procedure address element does query dictionary to standard elements names. The matching registration procedure writes matching result into XML file. The above three procedures are simple and are not the main point. In Section 2.4 and Section 2.5, this paper will respectively describe two important procedures address p arsing and reasoning about address matching in the alogrithm of Chinese address matching algorithm based on natural language understanding.

2.3 The address model of the spatial relationships and the logic model of address database

Address model is abstract to the address to express the relationship between the address elements. Address database model is a storage model of address elements to record the relationship of the address elements. Address model and address database model are the basis and foundation for the address parsing and reasoning match. This paper proposes the spatial relationship a ddress model to solve abstract problems of the Chinese address by a new perspective, and builds a new address database model to to solve the storage problem of the address information for knowledge representation and address elements of the inner c ontact.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.3.1 A spatial relationship address model instead of the hierarchical address model

The Chinese address is composed by place name, orgnization name, house numbers, etc. These ingredients are known as the address elements. Address model is abstract to the relationship of address elements. Different language and culture cause the differences of address expression. As a result the a ddress model is not the same. Common address model is hierarchical address model. The hierarchical address model is proposed according to the U. S. situation. The hierarchical model assumes that affiliation exists between adjacent elements of the address. This model is only suitable to describe simple street address for China. Usually Chinese address is composed by zoning and local location name or number elements. These elements exists spatial relation. By analyzing the Chinese address this algorithm finds that address elements have the five kinds of spatial relations: containing, adjacency, adjoining, direction, and distance. This p aper proposes the spatial relationship address model to take place of the hierarchical address model for China. Fig. 2 shows the relationship between address elements.

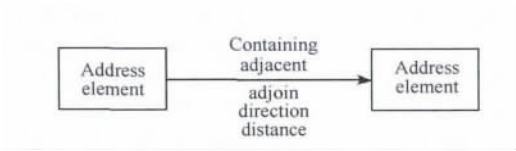

Fig. 2 The spatial relationships between address elements

The address elements of the spatial relationships and the type of address elements are closely linked. Containing relation: the region contains region, roads, Points of Interest (POI); road contains the POI. Adjacent relation: region adjacents to region and road, and road adjacents to road and POI, and POI adjacents to the POI adjacent. Adjoining relation: road adjoins road. Direction relation: direction between the POI points. Distance relation: distance between the POI points. In order to more clearly explain the relationship between the address elements, The following three addresses, The department of transportation, Longhua district, Puyang city 200 m west, Building 5, Fenglinshuian, Lidu road, Longhua district, Puyang city and No. 161, Daqing road, Shengli east road, Longhua district, Puyang city, are resolved, As shown in Fig. 3.

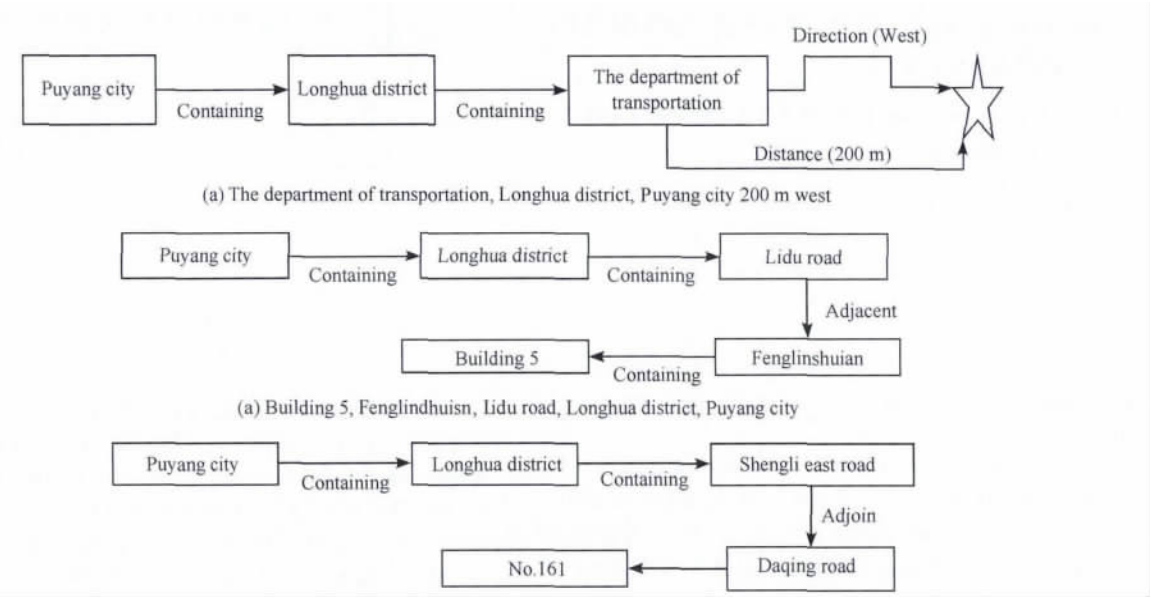

Fig. 3 Instance of address element relationship

2.3.2 Address database logical model

The semantic understanding is based on knowledge base. In the paper the knowledge base is built based on address database. The knowledge is the relationship among address elements in the knowledge base. Address database, also known as the address reference library, is the address matching engine data base to meet the demand of address match and exchange.

Because address reference library is a very important foundation library, reasonable design is essential. As most of the needs of business systems are the residential address and enterprises address matching, the address database designed by this paper is focused on these two types of application. The address database design is also covered by these special data records to ensure a number of important locations (bus stops, bridges, roads, ports, parks, stadiums, squares, etc.) are matched.

The logical model design of the address database refereces the names address data standard (2010). Address records contain structured and unstructured data to provide sufficient data basis for a variety of different types of address matching algorithms. The address elements are the basic unit of the address in the a ddress database. The address is stored in the structure of the a ddress elements as the basic unit. The spatial relationship among address elements is the core of the address database model design. The address database consists of three parts: zoning, spatial relationships, and place names. The affiliation of Zoning (provinces, cities, counties, streets, etc.) is the main structure in the logical model. The spatial relationships of address e lements are recorded. The alias table of the model is the base t able, which records the most detailed information of names. The alias table is refrenced by other tables to avoid redundancy and to

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

keep consistency. The auxiliary table of the address record is the gate address table, which records the actual collection of local point.

The address database model requires data support. The data collection is a heavy task of work, for which the main method is to use high-resolution remote sensing images (aerial imagery) to manully plot, and some areas' surveys are supplemented by artificial to reduce the cost of data collection.

2.4 Address parsing

The address parsing refers to the unstructured Chinese address into address elements and determines the type of address elements. The address parsing is considered as the second important procedure. Usually the Chinese address is described by unstructured text, and there is no separator between the address elements, and not easy to split. Address segmentation that solves the problem of unstructured addresses split is the basis of semantic processing. No separator between Chinese words, so the Chinese address segmentation is a more complex process than the English segmentation.

The parsing algorithm consists of two components: the segmentation and the label of the address elements. The Chinese address parsing is similarity with Chinese segmentation, which is a process of splitting Chinese statement into words and labeling the words.

The algorithm process's framework references the Chinese word segmentation process, and introduces some special processing techniques according to the characteristics of the Chinese address. Address segmentation aims to recognize place names and organization names, so address segmentation is application-specific named entity recognition. Over time, a large number of new place names and organization names appear, the forms of which are not stable.

The address segmentation faces the following challenges.

(1) Place names coverage: the gazetteer library cannot keep up the changes with the society, which leads to difficulty to recognize unregistered named entity.

(2) Boundary identified: need to determine the boundaries of the place name or organization. Because of the complexity of the law of the place names and organization names, it is not easy to use simple rules to match and to determine the boundary.

Address elements label faces the following problem.

(1) Duplicate named entities: The problem of Duplicate names causes that some names have many address element types.

(2) The uncertain type of unregistered named entities: as the entity name does not exist in the address database, the type of unregistered named entities is uncertain.

The artificial address segmentation algorithm is different from the traditional segmentation algorithm. Word is the smallest semantic unit in linguistics, and is also the composition of unit of address element. Usually the word is less semantic unit than the address elements. Address elements may be composed by one or more words. The address elements can be viewed as compound words. The core of Chinese address intelligent segmentation makes use of the address context, and automatically optimizes the segmentation results.

The intelligent segmentation algorithm combines rules and statistics, and simulates human to process address segmentation process. The rules not only contain the framework rules, but also contain the constraint rules of the address elements. The statistical segmentation algorithm makes use of the word probability to identify word.

The intelligent address matching algorithm determines the action based on the type of address elements. Therefore, to ensure the matching algorithm correctly, the type of address elements must be correct. The purpose of address labeling is to determine the type of address elements.

The type of place names is determined by two aspects: the context and expression habits. The context has a strong binding with the occurrence of address elements and determines the type of some address elements. Habits of expression belong to statistical areas, and the inter-occurrence probability of the address elements type determines the address type. Taking into account these two factors on the address type, this algorithm combines rules and statistics. The statistical model used here is the Hidden Markov Model (HMM) (Rabiner, 1989).

HMM is widely applied to recognize states by observing the sequence values. For this application, the states' values of HMM correspond to place name type, and observing values correspond to place names. Solving the type of address elements is converted into finding the optimal state sequence. The brute-force strategy is huge computational, and cannot be applied in practice. The algorithm finds a local optimal solution by the HMM decoding algorithm (Viterbi algorithm). The type of address elements relates with the relationship among the address elements. Such as, building's name cannot follow the road name. Therefore, when using the Viterbi algorithm constraints should be considered. These constraints are obtained by summing up the combination law of the address elements.

2.5 Reasoning about address matching

The reasoning about address matching is a process of reasoning and location based on nature language understanding, and is one of the two important aspects of this algorithm. From the perspective of artificial intelligence, the reasoning is used to infer another. Because of varieties of linguistic expressions, a position can be described by many different addresses. The following example shows that a simple position of Puyang city, Henan Province has at least five forms. (1) No. 17, Jinrong road, Hualong district, Puyang city; (2) the residential area of the industrial development bureau, Jinrong road, Hualong district, Puyang city; (3) the residential area of the industrial development Bureau, Hualong district, Puyang city; (4) No. 17, Jinrong road, hualong district, Puyang city; (5) No. 17, the residential area of the industrial development bureau, jinrong road, Hualong district, Puyang city. For this same location a variety of descriptions, it is impossible that matching result is correct.

The reasoning about address matching makes use of natural language understanding technology to reason and determine the spatial relationships among the address elements. According to the characteristics expressed by the Chinese address, the major components of the Chinese address are the regional names (zoning areas, functional areas, the natural area) and a local point. The relationships include adjacent, containing, adjoin, d

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

irection, and distance. Reasoning is based on decision tree e stablished by address database to select the appropriate action. The description of the address may contain the redundant information, such as there is equivalence relationship of the address elements. For example, house number with the corresponding o rganization can be used to represent the local point, and is an e quivalence relation. Reasoning process contains the diagnosis and treatment of multiple relationships. The reasoning of the main action contains finding, relationship recognition, the relative position of the calculation, and interpolation calculation. The purpose of finding collection is to judge whether the address elements is in the address database records. Relations determine spatial relations or equivalence relations between the current address elements and the previous address elements. When the address is described by the reference position, the a lgorithm needs reference sites as a benchmark to calculate of the offset position. The accuracy of the offset location of the reference point is determined by the offset position description. When the address point does not exist in the address database and the a ddress contains doorplate, interpolation is used to solve two types of the address position. The first class is the e ncoding method, and the second is sequential coding method.

The reasoning process uses heuristic strategies, and knowledge (the relationship of address elements) is stored in implicit. The method of implicit graph search solves problems. In order to reduce the size, the search uses backward reasoning strategy. The core of the intelligent matching is knowledge-based reasoning to find the optimal solution in the knowledge space. In this algorithm, the knowledge base is also the address database. The final result of knowledge reasoning is a path from the root node to a leaf node with depth-first strategy. Because of the record pointer from the child node to parent node (parent node identifier) in the address database, the complexity of depth-first search reduces significantly. The depth-first search recursively finds the current child node as a starting point and finds parent node for the end of local path. If the path exists, it is successful that moving from the parent node to the current node. Otherwise, the path does not exist.

The reasoning about address matching algorithm is shown in Fig. 4 and the process is described as below.

Step 1 The input of standized address and matching object (building, doorplate, road, etc.) is the premise. The procedure judges the validity of the address. If the address is valid, the p rocedure goes to Step 2, else goes to Step 9.

Step 2 The step determines wheather the address contains d irectional elements. If the directional elements do not exist, then goes to Step 3, otherwise goes to Step 7.

Step 3 The step recognizes the address type. If the address is simple, than goes to Step 4, otherwise goes to step 5.

Step 4 The step queries the positon by full-text search method, and than goes to Step 9.

Step 5 The step computes position with finite state machine, and than goes to Step 6.

Step 6 The step locates the position by interpolation, and than goes to Step 9.

Step 7 The step matchs partial address, and than goes to Step 8.

Step 8 The step computes the relative position, and than goes to Step 9.

Step 9 The step logs ther matching results.

The reasoning algorithm simulates human behavior. The process is similar to post in accordance with zoning extent. The matching success or failure is determined by the path from start to finish node.

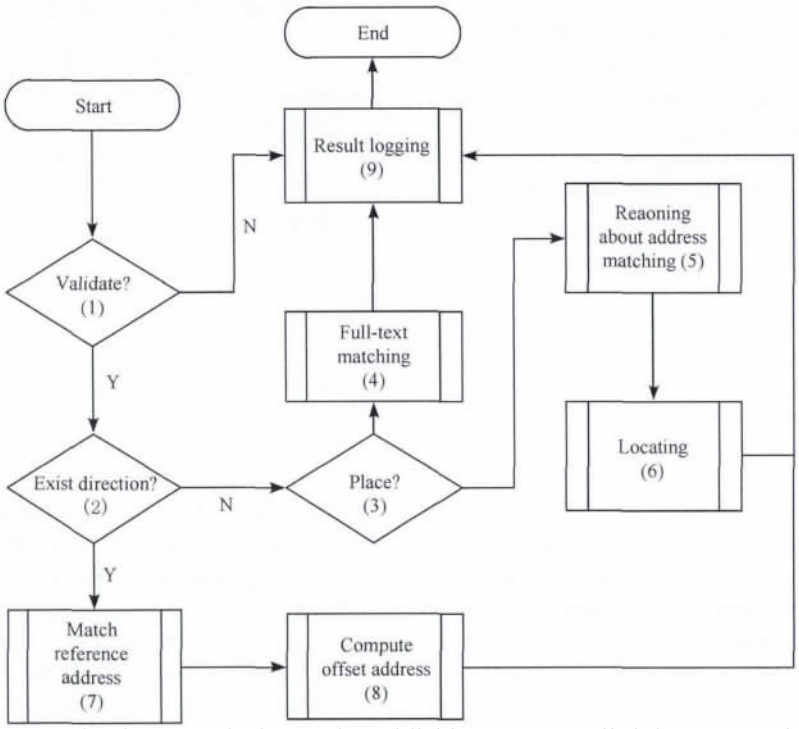

Fig. 4 Reasoning about address matching

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3 EXPERIMENTAL RESULTS AND ANALYSIS

In order to verify this algorithm, an experimental prototype system of Chinese address matching is developed. The experimental system uses 15336 items of records from population l ibrary of Puyang city, Henan Province. 1000 items of addresses are used as the test data. Then this article analyzes experiment's results. In order to measure the effectiveness of the algorithm, the paper uses the matching rate and accuracy to measure the pros and cons of the algorithm. These two indicators are defined as follows. Where M is as total of matching addresses, and N is as total of successful matching addresses, and K is as total of c orrect matching addresses, matching rate is defined by $(N/M) \cdot 100\%$; correct rate is defined by $(K/M) \cdot 100\%$.

3.1 The artificial, experimental prototype system of Chinese address matching

Based on domestic geographic information system platform GeoBeans, the SDK 8. 1, the three-tier software framework is used to develop the Chinese address smart match the experimental system with the algorithm proposed by this paper. The experimental system includes mainly the gazetteer library module, the address training samples module, the matching m odule and the visualization module. The gazetteer library module managers place names. The address training samples module is used to the address model (HMM) parameter training. The matching module and the visualization module are used to interactive matching. The back-end database of the experimental system is the ORACLE 10. 2 g with ordinary computer desktop (2 G of memory, the CPU Intel E6550, hard drive 7200 rpm).

3.2 Experimental results and validation

Using the above developed experimental system with Population address data of Puyang City, Henan Province, we selected some compliant address data to build address database with 15336 items. In addition, we chose the Puyang City planning and administrative data with zoning name, road name, district name, village name, and the organization name for the gazetteer to cover most place names.

In order to calculate the matching rate of the algorithm, we randomly selected 1000 items from Population address library to do experiments. Depending on the application requirements, we set match level to the building. In order to test the accuracy of the spatial understanding semantic for the address matching algorithm, we artificially constructed the same location described with more than one addresses. 100 items from the address library that can be completely accurate matched is as considered as basic addresses. We added or removed part of the address elements to construct new address. These new addressed have the same positions with the former to do semantic experiment.

The matching visualization of this algorithm is shown in Fig. 5. The figure shows the matching result of the address experimental kindergarten block, Longhua district, Puyang city, 102 victory road, Building 1 with red diamond symbol.

The statistical results of the experiment are shown in Table 1.

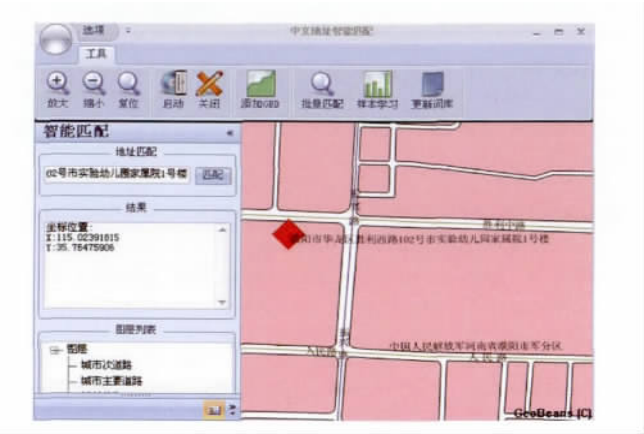

Fig. 5 Matching visulization

The matching results are manually checked to test the accuracy. The average time consuming of each address match takes about 0. 2 s.

Table 1 Matching result

Source	Number	Matching rate/%	Accuracy/%	Speed (Every Item)/s
Random residence address	1000	98.7	93.5	0.2
Artificial address	100	98	96	0.2

3.3 Result analysis

The experimental results showed that matching rate and a ccuracy rate were over 93% with matching to building level. By analyzing the raw experimental data, we found that the reason for the failure of partial addresses matching is incomplete gazetteer library. The matching rate of artificial addresses is very high, and this shows that the matching algorithm based on natural language understanding can understand spatial semantic. The reason for matching failure of the artificial addresses is lack of spatial relationship between address elements. The matching algorithm depends on the quality of the address database.

The matching efficiency of this algorithm in the common desktop is about 5/s, so the algorithm achieves a practical level. The recognition of the spatial relationship is the most time-c onsume procedure. The algorithm needs a large number of read operations in the procedure. External storage devices reading and writing (I/O) speed is far behind the CPU operation speed, so the file I/O is the main factor affecting the algorithm; In order to r educe the frequency of reading the file, the buffer is very useful for high-end equipment to improve the matching speed.

4 CONCLUSIONS

This paper analyzes the three main categories matching a lgorithms for Chinese address, elements of level matching, full-text search method, regular expression method. We found that the relative limitations of these algorithms. A Chinese address matching algorithm is proposed based on natural language u nderstanding with verification experiment using 15336 items of addresses from population database, Puyang city, Henan. The r esults showed that the matching rate and accuracy of the algo-

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

rithm can achieve a high level, indicating that this algorithm is an effective method of address matching with good adaptability. This algorithm not only can be used to match the specification of the Chinese address, but also can be used to match non-standard address.

REFERENCES

Chen X Q, Chi Z X and Jin N. 2004. Application and study of city Geocoding system. Computer Engineering, 30(23): 50 - 52

Goldberg D W, Wilson J P and Knoblock C A. 2007. From text to geographic coordinates: the current state of geocoding. Journal of the Urban and Regional Information Systems Association, 19(1): 33 - 46

Guo H, Song G F, Ma L Q and Wang S H. 2009. Design and implementation of address geocoding system. Computer Engineering, 23(1): 250 - 252

Hutchinson M. 2010. Developing an Agent-Based Framework for Intelligent Geocoding. Perth: Curtin University of Technology

Jiang Z, Li X L and Liu B S. 2007. Geocoding technical standardization of geographic information system. World Standardization and Quality Management, (5): 22 - 25

Ma L B and Gong J Y. 2003a. Application of spatial information natural language query interface. Geomatics and Information Science of Wuhan University, 28(3): 301 - 305

Ma L B and Gong J Y. 2003b. Research on spatial database query oriented natural language. Computer Engineering and Applications, 39(22): 16 - 19

O'Reagan R T and Saalfeld A. 1987. Geocoding Theory and Practice at the Bureau of the Census. Statistical Research Report Census/SRD/RR-87/29. Washington, D. C. U. S. Census Bureau

Rabiner L R. 1989. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2): 257 - 286 [DOI: 10.1109/5.18626]

Sun C Q Zhou S P and Yang L. 2010. Chinese geo-coding based on classification database of geographical names. Journal of Computer Applications, 30(7): 1953 - 1955, 1958

Sun Y F and Chen W B. 2007. Address matching technology based on Chinese segmentation//Geographic Information System Association Member Congress and 11th Annual Conference Papers. Beijing: China Association for Geographic Information Service: 1 - 13

Wang L Y, Li Q, Jiang Z. 2004. Study and development of geocoding database oriented China. Computer Engineering and Applications, 40(21): 170 - 171, 215

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

自然语言理解的中文地址匹配算法

宋子辉

遥感国家重点实验室 中国科学院遥感与数字地球应用研究所, 北京 100101

摘 要: 在分析现有 3 类主要的中文地址匹配算法: 要素层级匹配法、全文检索法、正则表达式法的基础上,提出了基于自然语言理解的中文地址匹配算法。新算法中建立了空间关系地址模型以解决中文地址抽象问题、地址库逻辑模型以解决地址信息的空间知识表达问题。新算法的完整流程包括预处理、地址解析、地址要素标准化、推理匹配和匹配登记等 5 个环节,本文重点阐述了地址解析和推理匹配这两个重要环节,分别依据"自然语言理解"中的中文分词和语义推理原理,对用非结构化的中文自然语言描述的中文地址进行处理,实现自然语言理解方法与地址匹配之间的结合,从而建立完整的基于自然语言理解的中文地址匹配算法。为验证该算法,开发了中文地址智能匹配实验系统,对河南省濮阳市人口库 1000 条居民地址数据进行匹配,匹配率达到了 95%,准确率高于 93%。

关键词: 自然语言理解,地址匹配,地址要素,地址解析,隐马尔科夫模型

中图分类号: P208 文献标志码: A

引用格式: 宋子辉. 2013. 自然语言理解的中文地址匹配算法. 遥感学报,17(4): 788 -801

Song Z H. 2013. Address matching algorithm based on chinese natural language understanding. Journal of Remote Sensing, 17(4): 788 -801 [DOI: 10.11834/jrs.20132164]

1 引 言

地址信息与人类社会活动紧密相关。在中国,统计、工商、公安等部门都需要记录和保存大量的地址信息。美国人口普查局统计,政府的管理信息系统中有 80% 的信息和空间位置相关,而这些数据大部分都包含地址信息(O'Reagan 和 Saalfeld,1987)。地址匹配技术能把含有位置的文字信息与空间信息关联起来,整合空间信息和社会经济信息,提供数据分析、定位、制图和可视化等服务,在国民经济建设和人们生活中发挥着不可或缺的作用。

严格地说,地址匹配是指把文字描述的地址信息映射成地理坐标的过程(Daniel,2007),这个过程建立了文字描述具有空间位置的地址信息与地理坐标的联系,并实现了这种联系的定量转换。中文地址匹配则是指把中文文字描述的地址信息映射成地理坐标的过程。目前一些公司已经推出了针对欧美国家的地址匹配软件产品或服务,比如 ESRI 的 Geocoding,MapInfo 的 MapMarker,Google 的 Geocoder。但由于国情差异,这些软件产品或服务均无法满足中国日益增长的中文地址匹配的需求。

从理论方法的角度看,迄今为止现有的中文地址匹配算法主要有以下 3 类: (1) 以地址要素层级模型为核心的地址匹配算法(江洲等,2007; 王凌云等,2004; 郭会等,2009; 孙存群等,2010),(2) 以全文检索模型为核心的地址匹配算法(孙亚夫和陈文斌,2007),(3) 以正则表达式匹配为核心的地址匹配算法(陈细谦,2004)。以地址要素层级模型为核心的地址匹配算法的特点是地址要素都有级别属性,每一类地址要素的属性都有唯一的级别值。这类方法对地址的描述要符合等级规则,地址要素对应的级别按照降序排列,这使得该方法只能匹配一些特殊的地址。以全文检索模型为核心的地址匹配算法是把地址库作为文本库,待匹配的地址作为检索条件。这个算法的特点是只考虑关键词匹配,匹配速度快,匹配率高,但准确率不高。以正则表达式匹配为核心的地址匹配算法完全是建立在字

收稿日期: 2012-05-16; 修订日期: 2012-06-18; 优先数字出版日期: 2012-06-25

基金项目: 国家高技术研究发展计划(863 计划)(编号: 2012AA12A401)

第一作者简介: 宋子辉(1980—),男,博士研究生,主要从事空间信息搜索与地址匹配技术研究。E-mail: szhmvp@gmail.com

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

符串比较的基础之上,匹配速度慢,匹配率高,但匹配的准确率低。

利用自然语言理解这一人工智能领域新技术来建立新的中文地址匹配算法是本文的主要思路。上述现有的3类地址匹配算法本质上是采用字符串、关键词、规则的地址匹配技术,而不是从理解地址语义的方向入手,无法实现语义匹配,因此这3类算法也不属于基于自然语言理解的中文地址匹配算法。目前,除了上述3类中文地址匹配算法有关文献做了大量研究工作外,研究自然语言理解与空间信息关系的文献还有马林兵和龚健雅(2003a,2003b)提出的空间信息自然语言查询。Hutchinson(2010)提出了基于智能体的地址匹配算法,其特点是把人工智能技术引入到地址匹配过程,但是该算法的处理过程是基于地址要素层级模型(LBM)设计,这种模型不适合用于抽象的中国的地址,因此该算法解决不了中文地址匹配。目前尚未发现在地址匹配算法与自然语言理解结合方面进行研究的相关文献。本文正是把自然语言理解应用到中文地址匹配这个特定领域的探索。

2 基于自然语言理解的中文地址匹配算法

2.1 地址匹配与自然语言理解

自然语言理解实质是对自然语言进行计算机处理,使计算机具有理解和运用自然语言的功能。当前自然语言理解技术还没有达到人类的智能水平,对人脑知识库表达还没有研究透彻,仍然是对特定领域的浅层自然语言理解。但是即使浅层自然语言理解也会给很多领域的应用带来有价值的帮助。

由于中文地址基本上是采用非结构化的中文自然语言来描述,地址匹配的本质是把含有位置信息的文字信息与空间信息关联起来,因此,对用非结构化的中文自然语言来描述的中文地址进行空间语义理解,即自然语言理解,把自然语言理解中的方法应用到地址匹配算法中,这在理论上是可行的。

每一种语言都有特定的语法,因而其计算机处理方式存在特殊性。中文属于表意文字,中文的词语之间没有分割符,而英文的词语之间有分割符,这导致中文计算机处理比英文要复杂很多,使得一些适合英文的自然语言处理方法直接应用在中文上会有很大的障碍。

中文的自然语言理解方法包括中文分词、语义标注、句法分析和语义推理等环节。由于缺少有效的知识抽象模型,当前自然语言理解技术一般都与具体应用场景相关。所以本文根据地址匹配的特殊场景,对自然语言处理的这4个主要环节要做有针对性处理实现。中文分词环节实现地址分割,语义标注环节实现地址分割单元的角色标注,句法分析实现地址类型标识,语义推理实现空间关系推理。通过这些有针对性的处理,实现了自然语言理解方法与地址匹配应用之间的结合,从而建立基于自然语言理解的中文地址匹配算法。

2.2 基于自然语言理解的中文地址匹配算法组成及流程

本文提出的基于自然语言理解的中文地址匹配算法流程如图1所示,完整的算法流程包括预处理、地址解析、地址要素标准化、推理匹配和匹配登记等5个环节。第1步针对待匹配的地址数据做预处理,进行一些简单的编码处理和信息过滤。第2步对处理后的地址做地址解析,把非结构化的地址转换成结构化表示形式。第3步对结构化的地址做地址要素标准化,把地址要素转换为标准地址要素。第4步对标准化后的地址做推理匹配,这是利用地址要素之间的关系和知识库(地址库)做语义推理,确定待匹配地址对应的地理坐标。第5步是对匹配的结果做匹配登记,即对匹配的结果做量化处理。

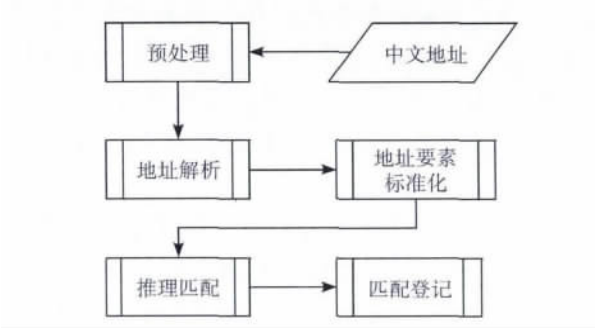

图1 基于自然语言理解的中文地址匹配算法流程

在本文算法流程中,预处理环节只是普通的字符处理,地址要素标准化是使用词典库查找要素名称对应的标准名称,匹配登记是把匹配的结果使用XML描述,并写入外部文件;这3个环节处理过程相对简单,不是也不应是本文的重点。在第2.4节和第2.5节,将分别阐述基于自然语言理解的中文地址匹配算法中两个重要的环节:地址解析与推理

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

匹配所用的方法。

2.3 空间关系地址模型与地址库逻辑模型

地址模型是对地址的抽象,用于表达地址要素之间的关系。地址库模型是地址要素的存储模型,用于记录地址要素之间的关系。地址模型和地址库模型是地址解析和推理匹配的依据与基础。本文针对地址模型建立了空间关系地址模型,以新的视角解决中文地址抽象问题; 针对地址库模型建立了地址库逻辑模型,以新的视角解决地址信息的空间知识表达和地址要素内在联系的存储问题。

2.3.1 代替层级地址模型的空间关系地址模型

中文地址是由地名、单位名、门牌编号等组成,这些成分称为地址要素; 地址模型是对地址要素之间关系的抽象。不同的语言文化,地址的表达方式存在差异,地址模型也不相同。目前常用的地址模型是层级地址模型,该模型是针对美国的情况提出,该模型假设相邻的地址要素之间存在隶属关系; 这个模型用于描述中国各地的地址有明显的局限性,只能用来描述普通的街道地址。通常中文地址是由以区划地名和局部地点名或者编号组成,各个要素在空间关系上构成约束的形式,空间范围有大到小逐级递减,这是中文地址的典型特征。通过分析中文地址发现,地址要素之间存在 5 种空间关系: 空间包含、相邻、邻接、方位、距离关系,为此建立一种代替层级地址模型的空间关系地址模型来抽象中文地址。图 2 反映了地址要素之间的关系。

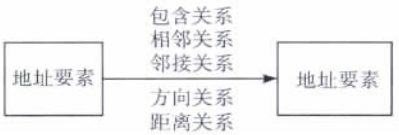

图 2 地址要素之间的空间关系

地址要素之间的空间关系与地址要素的类型有密切联系。包含关系: 区域包含区域、路、兴趣点(POI)或者路包含 POI。相邻关系: 区域相邻,区域与路相邻,路与 POI 相邻,POI 相邻。邻接关系: 路之间相连关系。方位关系: POI 点之间的方位。距离关系: POI 点之间的距离量算。为了更明确解释地址要素之间的关系,下面对 3 条地址濮阳市龙华区交通局西 200 m、濮阳市龙华区丽都路枫林水岸 5 号楼和濮阳市龙华区胜利东路大庆路 161 号解析并用图 3 表示。

2.3.2 地址库逻辑模型

语义理解是以知识库为基础,本文算法采用地址库构建知识库。知识库中的知识为地址要素信息及要素之间的内在联系。地址库又称地址参考库(Referencing Address Database)是地址匹配引擎的数据基础,满足地址匹配交换的需求。

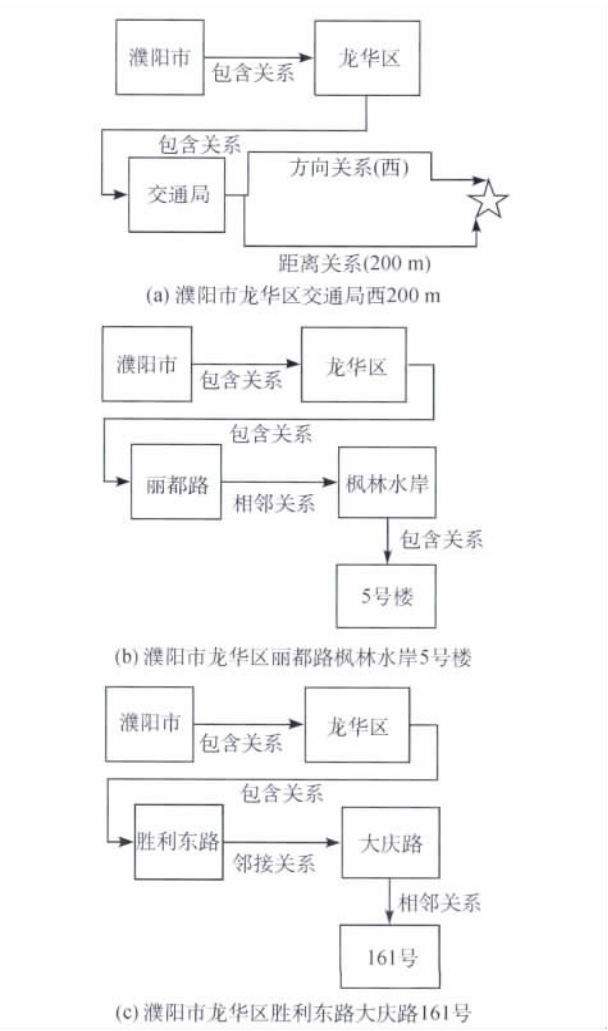

图 3 地址要素关系实例

由于地址参考库是非常重要的基础库,要满足知识的获取,合理的设计是必不可少的。大部分业务系统主要的需求是居民地址和企事业单位地址的匹配,所以本算法设计的地址库也是围绕这两类应用。为保证一些重要地点(公交站点、桥、道路口、公园、体育场、广场等)的地址能够匹配,地址库的设计也涵盖这些特殊数据的记录。

地址库逻辑模型,参照了《地名地址数据规范》(2010),在地址记录方式上采用结构化与非结构化相结合的方式,为多种不同类型的地址匹配算法提供充分的数据依据。地址要素是地址的基本组成单

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

位,在地址库中,地址的存储是以结构化的地址要素为基本单位,地址要素之间的空间关系是地址库模型设计的基础。地址库由3部分组成:区划、空间关系、地名。以区划(省、市、县、街道等)隶属关系为核心,记录地址要素之间空间包含关系、相邻、邻接等关系。模型的别名表是地名基表,记录最详细的地名信息,这是多类地名要素表的名称参考,减少地名的存储并保证地名的逻辑完整性。地址记录的辅助表是门址表,记录实际采集的局部点数据。

地址库模型需要数据支撑,采集数据是一项任务繁重的工作,数据采集方法主要以在高分辨遥感影像(或航空影像)为底图的人工标绘采集为主,并对部分地区辅以人工调查,降低数据采集的成本。

2.4 地址解析

地址解析是指把非结构化的中文地址拆分为地址要素并确定地址要素所属类型的过程,也是本文算法中第2个环节。通常中文地址采用非结构化文本表示,地址要素之间没有分隔符,不便于地址拆分。地址分割算法解决了非结构化地址的拆分问题,是地址语义处理的基础。中文词语之间没有分割符,所以中文地址分割过程要比英文表述的地址更加复杂。

解析算法包括两个组成部分:地址要素分割和地址要素标注。中文地址解析与中文分词有很大的相似性,中文分词也是把非结构化的中文语句分割为词条单元并标注相应的词性。

本文算法的解析流程的框架参考了中文分词流程,并根据中文地址的特征引入了一些特殊处理技术。地址分割主要解决地名和组织机构名称识别的问题,所以地址分割属于命名实体识别领域的特定应用。由于随着时间的推移,会出现大量新的地名和组织机构名,名称构成不够稳定,规律性差。特别是一些大城市,每年都会出现大量新的组织机构名称,而这些机构名称命名随意。

地址分割面临以下难题:

(1)地名库覆盖的问题:地名词典库收录的地名和单位名称跟不上社会的变化,面临着未登录命名实体识别的问题。

(2)边界确定的问题:由于地名库的限制,需要确定未登录地名或组织机构的边界。由于地名和组织机构名称生成的规律复杂,结构多样,提取构成规则困难,无法使用简单的规则匹配确定边界。

地址要素标注面临以下难题:

(1)命名实体重名:由于重名问题,地址要素有多种类型对应,名称同类型之间不是一对一的关系。

(2)未登录实体名类型不确定:由于实体名在地址库中不存在,未登录实体名的要素类型不确定。

智能地址分割算法不同于传统的分词算法,词是语言学中最小的语义单位,词也是地址要素的构成单位,词所代表的粒度小于地址要素。地址要素由一个或多个词组成,地址要素可以看作是复合词。中文地址智能分割的核心是利用地址的上下文结构,自动优化分割结果。

智能分割算法是规则分割与统计分割相结合的算法,模拟人类处理地址分割的过程。这里的规则既包括地址要素之间的框架规则,也包含地址要素之间的约束规则。而统计分割算法则利用字与字之间的成词概率,识别为登录的词。

智能地址匹配算法依据地址要素的类型确定处理动作,因此要保证匹配算法正确运行,地址要素的类型必须正确。地址要素标注的目的是确定地址要素的类型。

地名类型由两方面确定:上下文结构和表达习惯。上下文有很强的约束性,明确了部分地址要素类型出现位置。表达习惯是属于统计学范畴,由地址要素类型的互现概率决定地址类型。考虑到这两种因素对地址类型的影响,本文算法使用规则与统计相结合的方式,判定地址要素的类型。这里使用的统计学模型为隐马尔科夫模型(HMM)(Rabiner,1989)。

隐马尔科夫模型在解决由观察序列值确定对应的状态序列问题上有广泛的应用。针对本文应用问题,隐马尔科夫模型的状态值对应地名类型,观察值对应地名。求解地址要素的类型,转换为求最优的状态序列。使用穷举策略,计算每一种标注结果,能够筛选出最优的标注序列;但当状态比较多时,计算量非常庞大,无法在实际中应用。本文算法借助HMM的解码算法(Viterbi算法),求局部最优解。获得地址要素所属类型由于地址要素之间的关系有很强的约束性,比如楼房名后不能再跟随道路名,所以在使用Viterbi算法时要引入约束条件。这些约束是通过总结地址要素结合的规律获得。

2.5 推理匹配

推理匹配是指基于语义理解,实现推理和位置定位的过程,也是本文算法中第4个环节。从人工

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

智能的角度定义，推理是指由已知判断推断出另一种判断。由于语言表达形式多样，表达同一地点的地址，可以使用多条不同的中文语句表达。比如表达濮阳市工业局家属院址的至少有 5 种形式：（1）濮阳市华龙区金融路 17 号；（2）濮阳市华龙区金融路市工业局家属院；（3）濮阳市华龙区市工业局家属院；（4）濮阳市华龙区金融路 17 号濮阳市工业局家属院；（5）濮阳市华龙区金融路 17 号市工业局家属院。这些地址表达方式中第 1 条符合《地名地址数据规范》（2010）标准，第 5 条符合邮政行业《邮政地址信息数据结构》（2006）标准，而其余 3 条则不符合任何地址规范。对于这种同一位置多种表述的问题，不用自然语言理解角度处理，是无法实现正确匹配的。

推理匹配利用自然语言理解技术，推理判断地址要素之间的空间关系。根据中文地址表达的特征——中文地址的主要组成部分是由区域地名（区划区域、功能区域、自然片区）和局部点组成，描述的地理实体之间是包含、相邻、邻接、方位、距离关系。推理基于地址库建立决策树，根据不同的地址要素选择相应的动作。在地址描述中，经常遇到冗余描述，比如地址要素之间还可能存在等价关系，比如门牌号码与相应的单位都可用于表示局部的点，是一种等价关系。所以推理过程包含了多种关系的判断与处理，所以推理的动作主要有集合查找、关系确认、相对位置计算、插值计算。集合查找的目的是判断地址要素是否在地址库中有记录。关系确认是确定当前地址要素与前一个地址要素之间是空间关系还是等价关系。当地址是使用参考位置来表达时，则需要以参考地点为基准计算偏移位置，偏移位置的精度有参考点以及偏移位置描述的精度确定。当描述的地址点在地址库中不存在，而地址点是用门牌号表示，可以使用插值计算用于解决两类地址点位置计算，一类是门牌号是采用距离编码方法，另一类是模糊位置计算指定使用插值方法确定。

推理的过程采用启发式策略，知识（地址要素关系）存储采用隐式存储。采用隐式图搜索方式，求解问题。为了减少搜索的规模采用逆向推理策略。智能匹配的核心是基于知识推理，在知识空间内寻找最优解。知识库为地址库。知识推理的过程的最终结果得到一条从根节点到叶子节点的路径，知识推理过程采用深度优先策略。由于在地址库中记录了从子节点到父节点的指针（父节点标识），所以深度优先搜索的复杂度显著减少，只需采用递归寻找以当前子节点为起点，寻找以父节点为终点的局部路径，如果路径存在，则说明从父节点转移到当前节点是成功的，则可以继续做深度搜索；如果路径不存在则需要做异常处理。

推理匹配的算法如图 4 所示。推理匹配的过程描述如下：

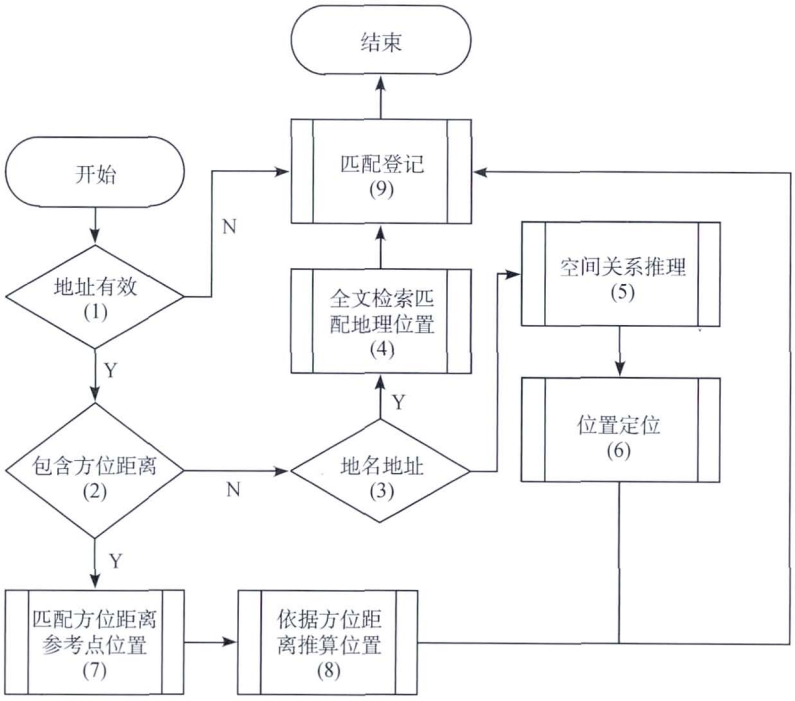

图 4　推理匹配算法流程

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

步骤 1 以地址解析的标准化后的结果和匹配的最终目标(楼房、门牌号、道路等)作为输入条件,判断地址的有效性。如果地址有效则转入步骤 2,否则转入步骤 9。

步骤 2 判断地址要素中是否存在方位描述。如果不存在则转入步骤 3,否则转入步骤 7。

步骤 3 对地址的类型判断,如果是普通名称类地址则转入步骤 4,否则转入步骤 5。

步骤 4 基于全文检索实现匹配定位,转入步骤 9。

步骤 5 利用有限状态机,实现要素关系推导,计算空间位置,转入步骤 6。

步骤 6 定位计算,转入步骤 9。

步骤 7 对地址划分,匹配无方位部分参照地址,转入步骤 8。

步骤 8 通过方位和距离估算目标位置。转入步骤 9。

步骤 9 对匹配的结果登记。

本文算法所采用的推理技术是模拟人类在寻址行为,因此地址匹配的过程类似与邮件邮寄,处理按照有大到小的过程,每一次匹配完成了路径查找。匹配的成功与否是由从起点到终点的路径决定。

3 实验结果与分析

为了验证及应用本文算法,结合上述研究的成果,开发了中文地址智能匹配实验系统,并结合河南省濮阳市人口库 15336 条地址数据进行试验验证,对其中 1000 条居民地址数据进行匹配,并对实验结果进行了分析。

为了衡量本文算法的有效性,使用匹配率和准确率这两个指标来衡量算法的优劣,这两个指标的定义如下: 匹配率定义为(N/M) · 100%。其中 M 为用作匹配的地址条数,N 为能够匹配获得地理坐标的地址条数。准确率定义为(K/M) · 100%,K 为正确匹配的地址条数。匹配率反应了成功匹配的概率,准确率反应了匹配结果的准确性。

3.1 中文地址智能匹配实验系统

中文地址智能匹配实验系统主要包括地名词典库管理、地址样本训练、匹配可视化等功能模块。名词典库管理模块用于管理地名词典管理,解决地址解析词库的编辑; 地址样本训练模块用于地址模型参数训练,解决地址类型标注模型(HMM)的训练; 匹配可视化模块用于地址交互式匹配,使用基于自然语言理解的中文地址匹配算法获得地址对应的地理坐标,并在地图中可视化显示。实验系统后台数据库选择 Oracle 10.2 g。实验所用计算机普通台式机(2 G 内存,CPU Intel E6550,硬盘 7200 转)。

3.2 实验结果及验证

使用中文智能地址匹配实验系统,以河南省濮阳市人口地址数据为基础,从中筛选部分符合规范的地址数据,创建以地址库逻辑模型为逻辑结构的地址库。使用的基础地址共计 15336 条,这些原始中文地址数据是采用非结构化的中文自然语言来描述的,如濮阳市龙华区胜利西路 102 号市实验幼儿园家属院 1 号楼这样的地址数据。另外再选用濮阳市规划及行政区划地理数据中含有区划地名、道路名、小区名、村名以及单位名称的地址数据为地名词典项,这使得地名词典库能够覆盖大部分地址中出现的地名要素。选择部分居民地址数据作为测试数据。

为了计算本文算法的匹配率,采用多条地址匹配实验,随机抽取濮阳市 1000 条户籍地址做实验。根据实际应用要求设置户籍地址匹配到楼房。为了检验地址匹配算法对空间语义理解的准确性,人为构造同一位置多条地址进行匹配。这里从户籍库中抽取 100 条能够完全准确匹配的地址作为基础,从中增加或去掉部分地址要素。增加或去掉部分地址要素后的地址数据与原地址数据表述的空间位置相同。用这些构造的地址数据作为实验数据。

地址匹配效果的可视化结果如图 5 所示,展示的是地址濮阳市龙华区胜利西路 102 号市实验幼儿园家属院 1 号楼的匹配结果,使用红色钻石符号标绘。

地址匹配实验的统计结果如表 1 所示。表中地址来源表示实验所用地址的数据源,条数表示参与匹配的地址数,匹配率表示匹配成功率,速度(每条)是匹配一条地址所用时间。为了检验匹配的准确性,对匹配后的结果做人工检查,检查的匹配的位置是否正确。对于 1000 条随机抽取的户籍地址数据,地址匹配实验结果是匹配率 98.7%、准确率 93.5%,对于 100 条人工构造的地址数据,地址匹配实验结果是匹配率 98%、准确率 96%。平均每条地址匹配耗时约 0.2 s。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

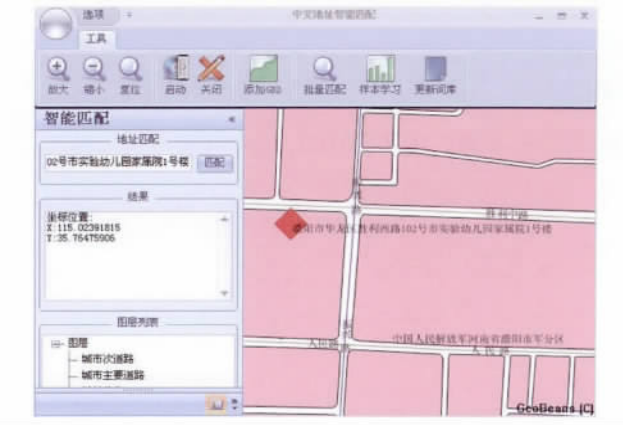

图 5 地址匹配可视化展示

表 1 地址匹配实验结果

地址来源	条数	匹配率/%	准确率/%	速度(每条)/s
随机抽取户籍地址	1000	98.7	93.5	0.2
人工构造的地址	100	98	96	0.2

3.3 结果分析

上述实验结果表明,户籍地址匹配到楼房级别,匹配率和准确率均超过 93%。通过分析原始实验数据发现,导致户籍地址匹配失败的原因是由于地名词典库覆盖的地名不全。对人工构造的地址,实验结果表明匹配率和准确率均很高,这说明本文提出的基于自然语言理解的中文地址匹配算法能够很好地理解空间语义,能够实现由多种表达方式的人工构造地址都能正确匹配,匹配失败的地址,均是在地址库中没有完整记录地址要素之间关系。如果地址库质量高,覆盖的数据广,意味着知识库丰富,则能很好的处理同一地点多种地址描述的匹配要求; 如果地址库的覆盖度不足,导致匹配率低,精度差; 匹配算法对地址库的质量有较强的依赖性。

本文算法在普通台式机上匹配的效率大约是每 5 条/s,达到了实用水平。匹配过程中需要推断地址要素之间的关系,这导致在算法需要大量的文件读取,由于当前外部存储设备读写(I/O)速度远落后于 CPU 的运算速度,因此文件 I/O 是影响本文算法的主要因素; 为了降低读取文件的频率,对与配置较高的设备采用缓存方式,把推理过程中对最近处理的地址要素做缓存,可以提高匹配速度。

4 结 论

本文分析了现有 3 类主要的中文地址匹配算法—要素层级匹配法、全文检索法、正则表达式法,指出了这些算法的相对局限性。在此基础上,提出了基于自然语言理解的中文地址匹配算法,并结合河南濮阳市人口库 15336 条地址数据进行试验验证,结果显示算法的匹配率和准确率都达到很高的水平。这说明,本文建立的空间关系模型能够用于抽象中文地址、基于自然语言理解的地址匹配算法能够更好地理解地址表述的空间语义,可以更好地实现地址匹配。本算法是一种有效的地址匹配方法,具有很好的适应性,不但能够用于匹配规范的中文地址,也能够用于匹配非规范的地址。

参考文献(References)

陈细谦, 迟忠先, 金妮. 2004. 城市地理编码系统应用与研究. 计算机工程, 30(23): 50 – 52

Goldberg D W, Wilson J P and Knoblock C A. 2007. From text to geographic coordinates: the current state of geocoding. Journal of the Urban and Regional Information Systems Association, 19(1): 33 – 46

郭会, 宋关福, 马柳青, 王少华. 2009. 地理编码系统设计与实现. 计算机工程, 23(1): 250 – 252

Hutchinson M. 2010. Developing an Agent-Based Framework for Intelligent Geocoding. Perth: Curtin University of Technology

江洲, 李小林, 刘碧松. 2007. 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, (5): 22 – 25

马林兵, 龚健雅. 2003a. 空间信息自然语言查询接口的研究与应用. 武汉大学学报(信息科学版), 28(3): 301 – 305

马林兵, 龚健雅. 2003b. 面向自然语言的空间数据库查询研究. 计算机工程与应用, 39(22): 16 – 19

O'Reagan R T and Saalfeld A. 1987. Geocoding Theory and Practice at the Bureau of the Census. Statistical Research Report Census/SRD/RR-87/29. Washington, D. C. U. S. Census Bureau

Rabiner L R. 1989. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2): 257 – 286 [DOI: 10.1109/5.18626]

孙存群, 周顺平, 杨林. 2010. 基于分级地名库的中文地理编码. 计算机应用, 30(7): 1953 – 1955, 1958

孙亚夫, 陈文斌. 2007. 基于分词的地址匹配技术//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. 北京: 中国地理信息系统协会: 1 – 13

王凌云, 李琦, 江洲. 2004. 国内地理编码数据库系统开发与研究. 计算机工程与应用, 40(21): 170 – 171, 215

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net