THE DESIGN AND IMPLEMENTATION OF ADDRESS MATCHING ENGINE

CHEN Xu [a, b,*], LI Qing-yuan [a], WANG Yong [a]

[a] Government GIS Research Centre, Chinese Academy of Surveying and Mapping, No.16, Beitaiping Road, Haidian District, Beijing, China
[b] School of Geomatics, Liaoning Technical University, Fuxin, Liaoning, China

KEY WORDS: address matching, geocoding, word segmentation, GIS, non-spatial information, spatial information, information resources

ABSTRACT:

In real life there are various types of information, not all information contains a clear geographical coordinates of the spatial information, a large portion of non-spatial information is relevant to space location, such as address, ID card number, telephone number, postcode, and so on, such information does not directly contain spatial coordinates, but implicit location information. In this paper, a engine, which natural language matches address, based on the word segmentation technology has been put forward. The engine can extract address name, ID card number, telephone number information from natural language and convert them to a clear geographical coordinates or the bounds of geographical coordinates. It will be the vital link between non-spatial information and spatial information.

1. INTRODUCTION

1.1 The background of research

In recent years, with the development of computer networks, information technology, digital earth construction, the sharing of geographic information is becoming more and more popular. The users' distribution in space extent has become an increasingly widespread and gradually penetrated into all fields of life. Many industries have accumulated a large number of data relative to geographic position but not clear geographical coordinates. These data cannot be used as GIS data source directly, let alone be analyzed by GIS software. In fact, most of this information, such as postcode, telephone number, ID card number, is closely related with location. At present, most sectors of data are isolated from each other; Information resources cannot be exchanged easily, without a "bridge" to connect the various business systems. With the Geographic Information System widely used in digital earth construction, the need of information resources integration and sharing is getting higher and higher, It is an important component that converting non-spatial data to spatial data.

Address matching is an effective solution to this problem. Address matching, also known as geocoding, refers to the establishment of geographical coordinates with the consistency of a given process address. It is a space-based positioning technology as a coding method. It provides the way that matching the location information which described by word to clear geographical coordinates.

1.2 The status of theory research

Many scholars have done a lot of research on address matching. Currently the main way of address matching is divided into two categories. One is based on the geographical grid; another is based on the geographical entity.

Grid-based address matching is the method that matches the address properties of geographical object through the grid-related methods to establish the relative between geographical object and address. This method has the advantage of positioning accuracy, precision can be freedom controlled according to the needs. The shortcomings of the object are difficult to establish the geographical relationship between the topology of space. It also requires the establishment of a set of strict grid system about the geodetic datum, reference ellipsoid, projection mode, the provisions of the preferred grid, grid origin and so on. This method also needs a certain period of time to be accepted and used by people.

The address matching method based on geographical entity is that match the address properties of geographical entity through the direct link between geographical entity and address. The advantages of this method are that it is easy to be acceptable to all parties and in accordance with the customary way of thinking, to promote easy. The disadvantages are that the irregularity and repeatability of address makes the address matching difficult.

Through analysis to the existing address matching engines, most of these engines have the following questions:

(1)The engine often cannot work correctly if users input fuzzy place names which contain non-standard address.

(2)The engine often lacks for semantic analysis function. If the address is contained in a sentences described by natural language, the engine often cannot exact the address from the sentence.

(3)The engine often cannot match postcode, telephone number, and ID card number to coordinates.

* Corresponding Author: Chen Xu, Major in GIS, Tel:(010)88217807, E-mail:newgis@163.com .

2. THE DESIGN OF ADDRESS MATCHING ENGINE

2.1 The address model of address matching engine

The original address will be splitted and standarded to one of common standard address model using address standardization methods. The common standard address model can be divided into seven classifies based on practical application situation and in reference to "The rules of coding for address in the common platform for geospatial information service of digital city" (CH / Z 9002-2007) guiding technical documents.

Address model	Examples
Door(House)Address|Street name | Administrative region name	No.16 Beitaiping Road ,Haidian District, Beijing, China
Markers name | Street name |Administrative region name	Xidian Hotial ,Yongding Road, Haidian District, Beijing, China
POI name| Street name |Administrative region name	Tian'an men square Chang'an Street, Haidian District, Beijing, China
Door(House)Address|Community name| Administrative region name	No.2 building, Unit 3,No.301, Xiaoyue Community, Haidian District, Beijing, China
Markers name |Community name |Administrative region name	Carrefour Supermarket, Xiaoyue Community, Haidian District, Beijing, China
POI name |Community name |Administrative region name	Xiangqing Restaurant, Xiaoyue Community, Haidian District, Beijing, China
POI name	Tian'an men square

Table 1. The classifies of Standard address model

2.2 The reference database in address matching engine

The reference database is the basis for the matching. It plays a decisive role in the accuracy matching of the nature language to geographic addresses. So establishing reference database is the essential work. The reference database, including the basic address database, the rules database of ID card number, postcode database, the rules database of mobile phone number, POI database (including enterprises and institutions, important markers, etc.) and administrative region entity object database.
The following principles should be followed in the process of establishment the reference database:

(1)The principle of uniqueness: Each of the geographic entities in the database has a unique id.(2)The principle of science: The affiliation of geographical entity can be identified from the coding.(3)The principle of scalability and sustainability: Should be changed to adapt to the development of object.(4)The principle of standards: It must be adapted to the national standard coding system in order to achieve the data sharing.(5)These principles play an important role underlying function in the system planning, design and implementation guidance.

2.3 The architecture of address matching engine

The engine is divided into four parts, operation and interaction, address standardization, coordinates matching and evaluation, reference database.

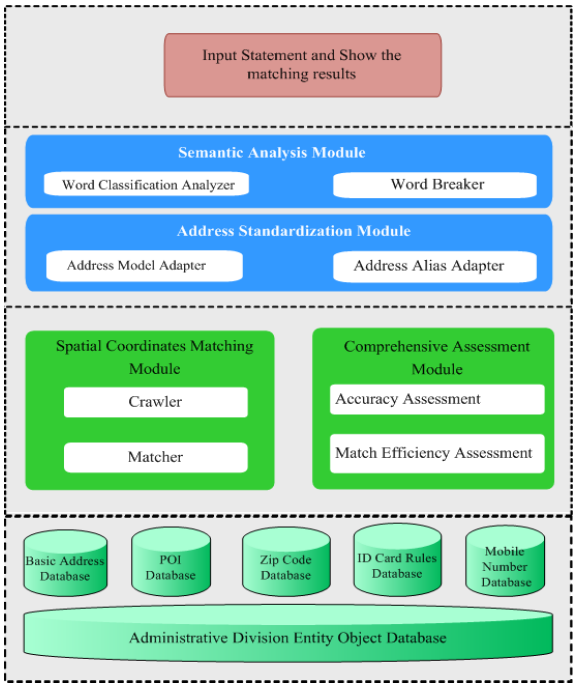

Fig 1. The framework chart of address matching engine

The part of operation and interaction provides the entrance of inputting sentence and the windows of showing the results;
The part of address standardization makes the data standard which inputted by users, and converts the "non-standard" data or "fuzzy" data into one of the standard address model;

The part of coordinates matching and evaluation realizes the identification on place name, postcode, ID card number, phone number from sentence and match them using segmentation technology, search technology. And then evaluate the accuracy and efficiency.

The part of reference database is the basis for coordinates-matching; it provides data to support to coordinates matching part.

3. THE REALIZATION OF ADDRESS MATCHING ENGINE

3.1 The workflow of address matching engine

The engine can use word classification analyzer and word breaker to extract place name, address, postcode, ID card and mobile phone number from the statement which users inputted. Place name and address can be standarded for one of the standard address model by address alias adapter and address alias dictionary. The most suitable location coordinates can be obtain by spatial coordinates matching module and comprehensive evaluation module based on the data from reference database.

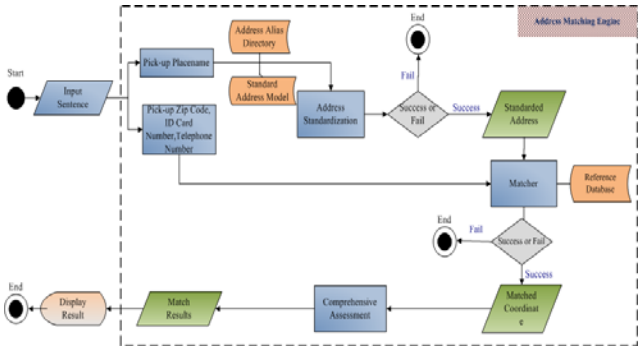

Fig 2. The workflow chart of address matching engine

3.2 Pick up address, Postcode, ID card, telephone number from the sentence

The word breaker uses segmentation technology to split the sentence which the users inputted. Segmentation is a part of Chinese information processing. It's not the aim, but the necessary stage of follow up processing. It's the basic technology of Chinese information processing. The Maximum Matching method is a method based on dictionary and can scan Chinese document effectively and decompose the document into collections of words. Thus, Chinese structured text is achieved. The engine splits words into collections words using the maximum matching method. For example, the sentence is "My address is No.16 beitaiping road, haidian district, Beijing, China ".The segmentation result is "My /r address /n is /v No.16 /m beitaiping road /ns , haidian district /ns ,Beijing /ns ,China /ns".
The engine picks up address using address alias directory. The directory is designed by the situation of a country. In China, the directory contains lots of tables provinces, cities, districts(counties), roads(streets), communities, houses, important buildings.

The engine uses regular expression to pick up postcode, ID card, telephone number. A regular expression, often called a pattern, is an expression that describes a set of strings. They are usually used to give a concise description of a set, without having to list all elements. For example, the expression, " ^[1-9]\d{5}$", can match to postcode in China.

3.3 Convert the "non-standard" data or "fuzzy" data into the standard address model

The common names and abbreviations alias of standard address are stored in address alias directory. The engine can modify "non-standard" data or "fuzzy" data into the standard address model based on reference database and address alias directory. For example, the "non-standard" sentence, "yongdinglu, haidian Beijing", will be converted to standard address model, "yongdinglu road, haidian district , Beijing , China".

3.4 Calculate the optimal result through spatial operation

The engine can calculate the optimal result through spatial operation in accordance with each matched coordinate scope. This operation will return the closest place to the users described information. For example, the phrase entered by the user contains the two matched results, "China" and "Beijing", "Beijing," the scope of the corresponding space is included in "China" region, and the engine will return "Beijing" as the optimal matched result.

4. DISCUSSION

The engine can pick up address information, ID card number, postcode, telephone number from the inputted sentence which described in natural language, and then match them to coordinates or spatial extent. It has the following characteristics: The engine can match most of "non-standard" or "fuzzy" address. The ideal situation is that the submitted sentence exactly matches to record in the reference database, but in most cases it's in a small probability. Short names or alias names of address are widely used. These will undoubtedly increase the difficulty of matching addresses. The engine can match the address contained in natural language. The engine can detect whether the statement contains postcode, ID number, phone number, if the information contained them, engine can match and return to the scope of latitude and longitude coordinates or coordinates scope. The engine can undertake a comprehensive assessment to the matched results of address, postcode, ID card number, phone number, get the optimum result. But this engine cannot match all of the "non-standard" or "fuzzy" data. If some alias or short name which the engine cannot be matched, you should add the alias name or short name to address alias directory, next time the engine will match them successfully. With the development of digital city, more and more attribute data need to be matched to geography coordinate. Address matching technology will be widely used. It will be the vital link between non-spatial information and spatial information.

REFERENCES

Jiang Zhou, Li Qi. *Research on the Applications of Geocoding*. Geography and Geo-Information Science, 2003, 19(003), pp. 22-25.

Xue Ming, Xiao Xuenian. *Considering on some questions of Geocoding*. Surveying and Mapping of Beijing,2007, (2), pp. 54-56.

Li Qi, Luo Zhiqing, and etc. *Research on Urban Grid System and Geocoding*. Editorial Board of Geomatiics and Information Science of Wuhan University, 2005, 30(5), pp. 408-410.

Gao Zhaoliang. *City Geospatial Dictory-Geocode*. Urban Geotechnical Investigation & Surveying, 2008, 2(2), pp. 20-22.
Zhang He, Kong Lingyan. *Research on Development and Application of Urban Geocoding*. Bulletin of Surveying and Mapping, 2008, (007), pp. 58-60.

Zhang Jiaqing. *The Construction and Perspectives in Space Information-supported E-government*. Science of Surveying and Mapping, 2003, 28(1), pp. 21-23.

Wu Xiuqin, Zhang Hongyan, and etc. *the Application and Practice of GIS (ArcGIS9)* . TSinghua University Press , 2008.
Zhu Qianfei. *Geocode and its Application in MapInfo*. Surveying and Mapping of Sichuan, 2001, 24(003), pp. 117-119.

Zhang Yifeng, Wu Jianping. *The Improvement of Geocoding in ArcGIS*. Geomatics & Spatial Information Technology, 2007, 30(3), pp. 710-713.

Liu qun, Zhang Huaping. *Chinese Lexical Analysis Using Cascaded Hidden Markov Model*. Journal of Computer Research and Development, 2004, 41(008), pp. 1421-1429.
Crosier, S. *Geocoding in ArcGIS* , ESRI. 2004.