THE DESIGN OF A UNIFIED ADDRESSING SCHEMA AND THE MATCHING MODE OF CHINA

Li FANG[1] *Zhuo-yuan YU*[2] *Xiang ZHAO*[1]
(1 College of resources science and technology, Beijing Normal University, Beijing, 100875)
（2 Institute of geographic sciences and natural resources research, Chinese Academy of Sciences, Beijing, 100101）

1. INTRODUCTION

Address databases are an important part in the construction of the geographic information infrastructure of Chinese cities. The existing address databases in China are mainly applied in cities and rarely cover countryside areas. And these databases usually include addresses which have regular formats, but addresses of other forms existed in some areas are not considered. In addition, the address database standards and database construction programs in different cities are different, and most of them are constructed repeatedly[1]. So it's urgent to propose the construction of the address databases in country level, which can help to standardize the work of address databases construction in Chinese cities, and help to data sharing and reduce duplicated construction of these databases. Therefore, to find out a unified address data schema, which can provide standard address data for GIS application in different cities, is significant to the address databases construction and application in China.

2. THE DESIGN OF A UNIFED ADDRESS SCHEMA

An unified address data schema is proposed on the basis of analyzing the addresses frequently used in China. The model describes and stores all types of address elements, such as cities, towns and villages, etc., and their relationships. (1) It describes and stores address elements of cities and countryside areas. (2) It establishes correlation between administrative regions and streets or resident areas, satisfying the address description/storage format of "administrative region + street" or "administrative region + resident area". (3) It establishes the line-point data structure of roads and buildings, which can satisfy the address description/storage requirements that house numbers are along the two sides of streets according to odd-even numbers or along the streets in sequence. Also this structure can meet the need of locating to street numbers by data interpolation. (4) It establishes the polygon-point structure of resident areas and buildings, satisfying the address description/storage format of "resident area + building". (5) It includes independent buildings' information, which does not have regular numbers. (6) It includes landmarks' information and stores symbolized point/polygon feature objects. (7) It includes Gazetteer and establishes the relationship between place name and other types of address elements, so as to realize the place-name-based address description/storage. (8) It includes address alias information and establishes relationship between address alias and its relative address elements, so as to realize the alias-based address description/storage.

3. THE ADDRESS MATCHING MODE

Based on the addressing schema, the rules and algorithms of the address matching mode are proposed. Fig. 1 shows the procedures of address matching. Firstly, split the input address into structured address phrases according to address element template; then match the address phrases with corresponding records in address databases according to certain rules and algorithms; and finally, get matching results.

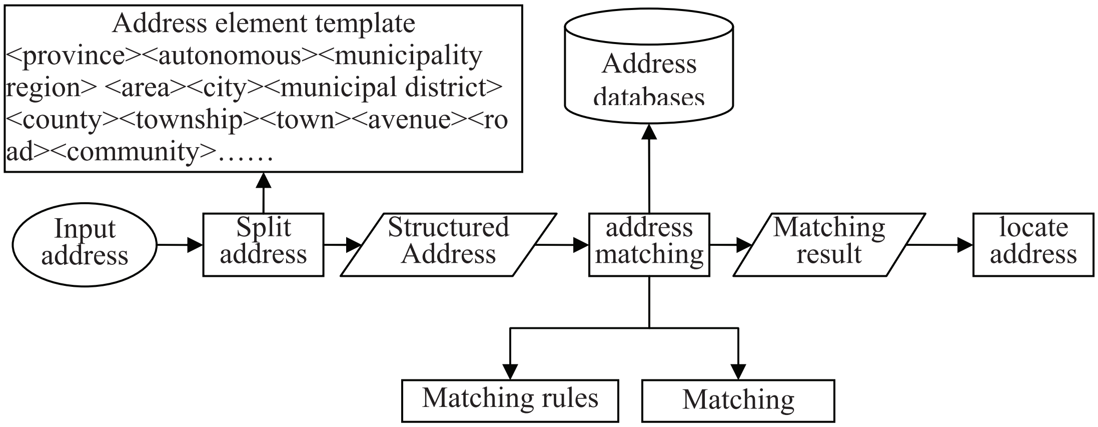

Fig.1 Address matching procedures

Address matching rules are shown in Fig. 2. According to spatial region precision, classify address elements into administrative region level, basic region level and point level. As the different administrate elements in administrative region level are hierarchical, match every address element with administrate element through the sequence from step 1.1 to step 1.4 as shown in Fig.2 until the minimum administrative region element is matched successfully. And the address elements in basic region level have the same matching sequence. When an address element is matched successfully in this level, it's not necessary to match the next address elements in this level, but go to the point level to go on matching. Street and resident area address elements are divided into two paratactic elements, i.e., city and countryside. If the minimum successfully-matched element in administrative region class is "town" or above it, it is city address and goes to the matching sequence of "city streets" —> "city resident area" —> "Gazetteer" —> "alias" —> "city building" —> "landmark"; if the minimum successfully-matched element in administrative region level is "town", then it is village address and goes to the matching sequence of "village" —> "Village roads" —> "Gazetteer" —> "alias" —> "village building" —> "landmark".

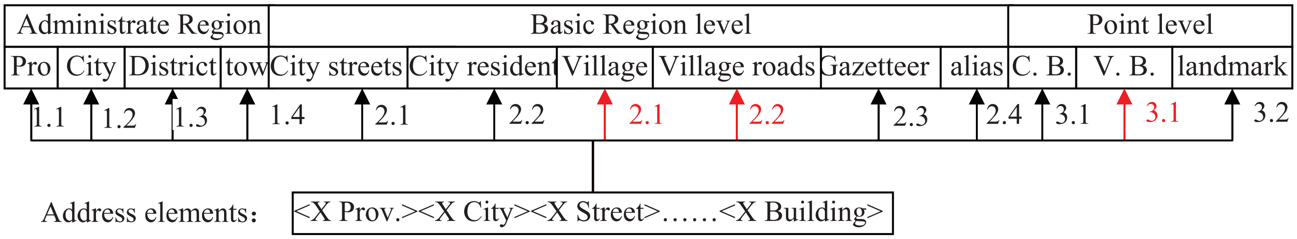

Fig.2 Address matching rules

Abbreviate Province and write `Prov.'.
Abbreviate City buildings and write `C. B.'.
Abbreviate village buildings and write `V.B.'.

In the process of address matching, matching algorithms are used to obtain matching results according to address matching rules. As there may have many problems in input addresses, such as abbreviation, misspelling, and character leakage, etc., various matching technologies should be comprehensively adopted to enhance address matching rate. For example, fuzzy logic algorithm can be used to get the address results which have the same substring with queried address; match algorithm based on split phrase can obtain results which have discontinuous matching string with queried address; establishing a pinyin index of an address element can correct misspelling of the same pronunciation, etc. The address matching rate of above algorithms may not high because it depends on whether the input addresses are standard. Another algorithm, the dynamic programming algorithm, which is widely applied to biology and information search field, can be applied in address matching. This algorithm finds out whether two strings are matching or not by calculating their similarity, thus the address input problems of miss characters, characters leakage, abbreviation, and non-standard format, etc. can be solved, and the address matching rate can be greatly enhanced. The algorithm can be improved by set filter conditions according to matching rules to increase efficiency.

The result of address matching returns some spatial reference object, then locating the given address by getting point coordinate according to the reference object. It is probably to return one, several or zero spatial reference object. If the result includes several objects, these objects are to be listed and sorted by similarity, and other additional information can be used to decide the most accurate coordinate for the given address. If the result returns null, then the address matching is failed. If the result returns only one object, then locate the address by getting the object’s spatial coordinate. If the object is point, directly locate the address according to point coordinate; if the object is line, precise coordinate of the given address can be gotten by interpolate the coordinate along the line object. If the object is polygon, then its centroid or any point in the polygon can be chosen for locating. Choosing any point of the polygon can avoid many addresses overlap on one point[2].

4. EXPERIMENTS AND RESULTS

Finally an address database and a prototype system are implemented based on the addressing schema and the address matching mode. An Oracle-based address database is established, which include Beijing basic geographic map scaled at 1:10000 as reference map and more than 50000 pieces of addresses records collected from Chaoyang District, Beijing City. And the address matching rules and algorithms are also realized. The system applies B/S structure, accessing and visiting address data and Web map data through J2EE development environment and ArcSDE. The experiment shows the system has high address matching rate and fast address matching speed.

5. REFERENCES

[1] Z. JIANG and Q. LI, " Research on the applications of geocoding," *Geography and Geo-Information Science*, vol. 19, no.3, pp. 22-25, 2003.

[2] DAVIS. C.A. Jr., FONSECA F.T, "Assessing the certainty of locations produced by an address geocoding system," *Geoinformatica*, vol.11, no.11, pp.103-129, 2007.

[2] ESRI, Geocoding in ArcGIS , ESRI Press, Redlands, CA, US. 2004

[3] ESRI. Census 2000 TIGER/Line Data Description. http://www.esri.com/data/download/census2000_tigerline/description.html

[4] U.S. Census Bureau. TIGER, TIGER/Line and TIGER-Related Products. 2009-9-8 http://www.census.gov/geo/www/tiger/

[5] GILBOA S. M., MENDOLA P., OLSHAN A.F. etc. "Comparison of residential geocoding methods in population-based study of air quality and birth defects," *Environmental Research*, vol. 101, pp. 256-262, 2006.

[6] M. Morad. "British Standard 7666 as a framework for geocoding land and property information the UK," *Computers, Environment and Urban Systems*, vol.26, pp. 483-492, 2002.

[7] P. Eichelberger. "The importance of addresses—the locus of GIS," in URISA 1993 Annual Conference, URISA: Atlanta, Georgia, 1993.