Acquisition and Analysis Methods of Geographic Space Data from the Internet

Su Deguo, Sun Lijian
Research Center of Government GIS
Chinese Academy of Surveying and Mapping
Beijing, China
E-mail: sdg@casm.ac.cn,slj@casm.ac.cn

Cui Ximin, Yuan Debao
College Of Geoscience and Surveying Engineering
China University of Mining & Technology
Beijing, China
E-mail: cxm@cumtb.edu.cn, ydb@cumtb.edu.cn

Abstract—How to analyze the spatial location relevant data from the Internet efficiently is one of the most important work in the software system field, which is one of the important directions. For many years, researchers have provided lots of analysis methods, but most of them are for ordinary data, and are not related to spatial location data. Due to the spatial location dependent data particularity, the corresponding special method can be studied. This paper introduces the using self-mapping visualization method to analysis from the spatial location data of Internet, and map expression can enhance understanding of the data visualization, easily to find rules behind the data. Through a large number of experimental applications，it is proved that the acquisition and analysis methods by self-mapping visualization are feasible and useful.

Keywords- Acquisition Method; Geographic; Space Data; Internet; Self-Mapping

I. INTRODUCTION

The content of this paper belongs to the category of data analysis. In particular, it is to study how to analyze the spatial data from the Internet, and provide a feasible method for high efficiency analysis. At present, there is no mature analysis methods for the spatial data from the Internet, mostly using the manual method to obtain data, and the method of data analysis is mainly directed against the common data, and the obvious characteristics of spatial data is position related, therefore, it is necessary to the study methods for the analysis of the spatial data from the Internet to find rules behind the data.

With leapfrog development in information technology and the popularization of the portable mobile terminals, the need for spatial cognition also becomes increasingly urgent. Information produced by human activities is ubiquitous, and there will be a lot of various types of events on the Internet every day. How to extract useful information from the complex and relevant data source? Cartographic visualization to show the spatial distribution of useful information can help users to find the patterns behind those events more easily. This will form useful knowledge to help human spatial cognition faster and more fully.

Web crawler is mainly based on the target data model. The target data is the data on the network that generally conforms to a certain pattern. Another way of description is to establish ontology or dictionary of a target domain. It can be used to analyze the importance degree of different characteristics in a topic from a semantic perspective. The comprehensive characteristics of web crawler can completely be combined with dynamic message driven in space geographic data mapping services to be an important mechanism for finding spatial position related web events. This paper focused on using the web crawler technology to obtain geographic spatial data from the Web and the method of utilizing Self-Mapping technology to analyze the spatial position related data.

The following parts of the article are divided into four parts. Firstly, this paper gives the classification and building methods of spatial position related web events, and studies the method to build the web events model by driven mechanism. The second part is the information extraction method on spatial position related web event elements, which gives out the process of collecting web event information by web crawler. Then, this paper describes how to build self-mapping model, which is helpful for analyzing online data. In the last part, this paper gives out the conclusion.

II. THE CLASSIFICATION AND BUILDING METHODS OF SPATIAL POSITION RELATED WEB EVENTS

According to the standardization theory and method, classifying web events is to distinguish and sort the web events by a certain principle and method according to their attribute or characteristics. Then, there are some classification systems and some sort order to build. Three methods were used to the classification of web events in the article as followed: line classified method, surface classified method and blending classified method. There are two key elements during the web events classification: one is the classified objects, and the other one is the basis for the classification. The determination of classified objects is the committed step for the classification of web events. It mainly includes some steps that the collection and arrangement of news groups, the judgment of web events, the selection of attributes of web events. The determination of the classification basis mainly contains some questions, such as the study on whether web events and the higher concept system are compatible or not, how to combine the classification and the application of the actual mapping and so on. Finally, making unified coding for the classified web events.

The method to build the web events ontology is the key technology to the study of driven mechanism. The ontology is consistent in understanding, sharing and reusable. This paper uses the construction methods of basic events, such as skeletal methodology, pair work method, and combines them with the thought of software engineering. On the base of web events and the logical structure definition of the ontology, this paper abstractly outlines the concepts of web events and the link relations between the concepts, and builds an ontology that focuses on hot events, disasters, key events, and public emergency finally. It can guide the extractions of multi-source event. The method of building web event ontology is showed as Figure 1

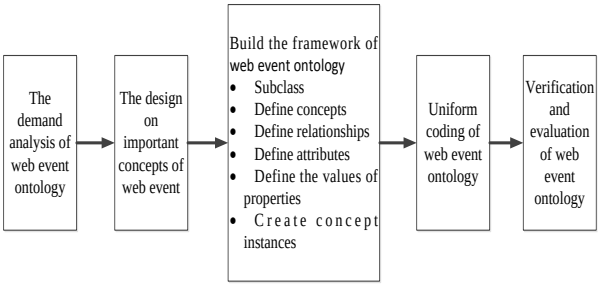

Figure 1. The Method to Build Web Event Ontology

The storage of web events is used RDBMS (Relational Database Management System) as commonly. The structure that describes event ontology is mapping to it. The mapping method is table-based mapping and object-relationship mapping.

III. THE INFORMATION EXTRACTION METHOD ON SPATIAL POSITION RELATED WEB EVENT ELEMENTS

The key technology to build the ontology base of web event is that the event's characteristic information must be obtained quickly, efficiently and completely. The extraction of web event elements includes two levels. The first step is to obtain dynamic event information and the second step is to extract the relatively static event information.

Firstly, dynamic event information is obtained by using the web crawler. Original information of all kinds of themes (such as disasters, public security, social economy, resource, etc.) can be got from multi-source (such as blog, webpage, and Twitter, etc.) and multi-format (text, html, xml, etc.) data source. Then the uniform document format is obtained by removing the duplicated web pages and noise filtering; all documents are stored in the original document repositories. In the next step, the index module builds inverted index stored in inverted index library according to the original document repositories. Users can query and manage documents by query module. At the same time, these original documents (their main content is text format) can be relatively static information source for the extraction of web event information. The process of building normalized document library by web crawler is shown in Figure 2. The process used by the web crawler to collect web event information is shown in Figure 3.

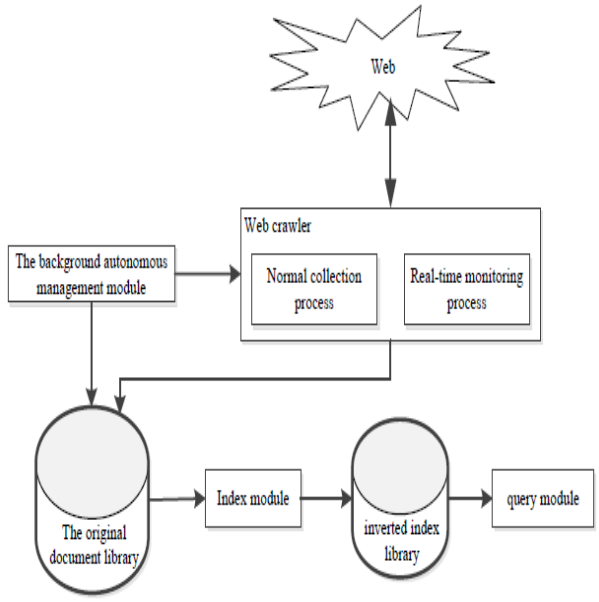

Figure 2. The Building Process of Normalized Document Library by Web Crawler

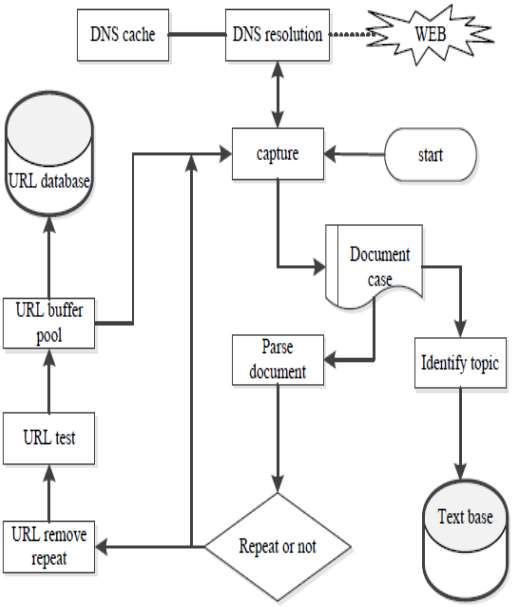

Figure 3. The Process that Web Crawler Collects Web Event Information

The process analyses the concrete forms of web event (text carrier), based on the web event normalized library that is updated constantly and dynamically. The exact extraction of web event is the base of understanding the drawing demand correctly. The general text information extraction includes event's location information, time information and web event's space-time attribute information and so on, and completing the confirmation of address by matching the database of base geographical name.

Then, we use the dynamic update of the network event specification information base, for the specific performance of network events (text carrier) for analysis and processing, the information of these events in the event of accurate information extraction is a correct understanding of the incident. Among them, the general

text information extraction, including the geographic location information (name, location description, etc.), temporal information and network events of non-temporal and spatial attributes (event nature, causing disaster factors, casualties, economic losses, social and economic impact, personnel input, fund input, etc.) and other information extraction. The annotation system based on temporal and spatial information and the semantic relationship of the network event, and the address confirmation was carried out with the base place name database. The recognition of place names is based on the machine learning method based on conditional random fields, and the spatial relation is realized by the rule model.

IV. THE METHOD ON BUILDING AND ANALYZING SELF-MAPPING MODEL

The right visualization representation of location-related data obtained from web is the important content of analyzing data. Firstly, it is building self-mapping model. The model can support the visualization representation pattern that is based on event driven web geographic information. It has the intelligent auto-mapping function. It can complete self-visualization mapping in different environment or for different use and the link from web event to self-mapping model.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. The Visualization Model of Time and Space of Web Events

Cartographic visualization can express the time and space attributes of web event. It can reflect the space-time dispersing or the interrelationships between continuous variation trend and events. It is the difficulty of self-mapping. Under the self-mapping mode, web events have unique, difficult to extract, difficult to understand and possibly missing case. The base of cartographic visualization is making space-time visualization for web events. If there are many time nodes, it will involve cartographic visualization problem in the process of events. The paper designs different time points to self-map repeatedly according to the method that conducts mapping experiments by the main time interval. Thus, it can dynamically reflect the occurrence, development and change to the end of events, and the spatiotemporal variation of complex event.

B. The Method of Personalized Mapping

First, based on the existing research, the existing mapping technology is integrated into the autonomous mapping model, including the establishment of user model (including user background information, user intention and behavior and user perception of the three parts), equipment (hardware and communication devices) and environment model (user's physical state and external natural environment and physical conditions, including time, lighting, weather, temperature, etc.), which is based on the realization of personalized mapping, which is a typical context factors. Based on these models, the formal description method of the model is established based on ontology. The ontology based technology is built to be easy to be shared, and it supports ontology reasoning and model reuse. In addition, because the information is an open set, it is not possible to take all the attribute elements into account. Therefore, this study uses the ontology to model the key individual elements.

C. The Design of Self Mapping Model

Self-mapping system can convert web events to standard mapping demand description language and properly understand the mapping demand of events according to the research on understanding and representation of the mapping demand and the standardized description of web events. The system designs the interpretation algorithms on the base of trial and error. It builds the whole structure and the application interface of mapping scheme by the object ideas. And it establishes an automatic combination mechanism about mapping scheme by researching the method on storage and management of mapping scheme. Then the most optimized mapping scheme can complete combination automatically according to the standardized mapping demand and the personal model and form a standardized mapping scheme.

D. Experimental Verification of Self-Mapping Model

The experimental verification is divided into two stages. Firstly, based on the existing work, the use of spatial data services, data services, and the existing spatial database, the database, and the existing spatial database, the database, in the local area network simulation of network events to establish a good service process validation, and then placed in a real network environment for autonomous mapping process validation. In the two process, at any time, the model is not reasonable to modify the mapping knowledge, and the semantic coupling between the network event and the model of the self-mapping model may need to be modified repeatedly.

Under the Internet environment, test the correctness of the network event response capability and mapping knowledge, and to test whether the desired effect of the expert level mapping can be achieved through the investigation of various users.

V. CONCLUSION AND FUTURE WORK

Through the integrated mapping process, it provides a powerful tool for emergency, emergency handling, and provides a powerful tool for the application of the knowledge discovery in the field of emergency and emergency, and provides a powerful tool for expanding the application domain of the map, and providing a powerful tool for the knowledge discovery in the field. At the same time, we can solve the problem of the user's personalized drawing (user adaptability), and meet the needs of the intelligent drawing of non-manual intervention, and meet the requirements of the expert level.

In future, the author will study the huge data about geography from web and the internal relationship of the data. Through the research, it is possible to find interesting unknown things.

ACKNOWLEDGMENTS

This research was funded by the Central-level Public Welfare Research Institutes for Basic Research and Development Operations under grant No. 7771528. In the project, the research task of network mapping is one of the important directions.

REFERENCE

[1] A. Buccella, A. Cechich, and I. Fillbottran, "Ontology-Driven Geographic Information Integration: a Survey of Current Approaches," Computers & Geosciences, 2009, pp. 710-723.

[2] C. Bizer, T. Heath and T. Berners-Lee, "Linked data-Thestory so far," Journal, International Journal on Semantic Web and Information Systems, 5(3), 2009, pp. 1-22.

[3] J. Park, "Developing a knowledge system for storing and using the design knowledge acquired in the process of a user-centered design of the next generation information appliances," Design Studies, 32(5), 2011, pp. 482-513.

[4] Q. Du, F. Ren, and Z. Cai, "Research on key technology of autonomous mapping service middleware based on Ontology," China science and technology achievements, 2013, 21, pp. 73-78.

[5] E. Grabska, G. lusarczyk, "Knowledge and reasoning in design systems," Journal, Automation in Construction, 20(7), pp.927-934, 2011.

[6] Mackinlay, Jock D., 1987. Automatic design of graphical presentations. Ph.D. thesis, Stanford University, Stanford, CA, USA.

[7] Michael Friendly, October 16, 2008. Milestones in the history of thematic cartography, statistical graphics, and data visualization. [3] Robinson, Arthur H., 1982. Early Thematic Mapping in the History of Cartography. Chicago:University of Chicago Press. ISBN 0-226-72285- 6.

[8] Tukey, John Wilder, 1962. The future of data analysis. Annals of Mathematical Statistics, 33:1–67 and 81.

[9] Bachi, Roberto, 1968. Graphical Rational Patterns, A New Approach to Graphical Presentation of Statistics. Jerusalem: Israel Universities Press.

[10] Monmonier, Mark, 1989. An alternative isomorphism for mapping correlation. International Yearbook of Cartography, 19:77–89.

[11] Zhu Guo-Rui, 1996.9. Map Design and Drawing. Wuhan University of Mapping Science and Technology, pp.219-228.

[12] Alessandro Cecconi, Christopher Shenton and Robert Weibel, 1998, Tools for Cartographic Visualization of Statistical Data on the Internet, University of Zurich, Winterthurerstrasse 190, 8057.

