Vector Data Acquisition Methods Based on the Spatial Data Sensitive Crawler

Wang Mingjun, Kang Mengjun, Du Qingyun
School of Resources and Environmental Science, Wuhan University
Wuhan, China
dawnson.wang@263.net

***Abstract*—This paper discussed the expression of vector data under the Web environment, introduced a spatial data sensitive crawler (SDSC), and tested the novel technique to collect vector data in real cases. Furthermore, a defect reduction model has been improved so that the level of confidence of different vector data can be measured more accurately. The results of experiments show that the data collected by SDSC have a relatively higher credibility.**

***Keywords*—Spatial Data Sensitive Crawler; Vector Data; Defect Reduction Model; Measure of Confidence Level**

I. INTRODUCTION

With the rapid development of information technology, a lot of studies based on spatial information acquisition from Web resources have been gained increasing attentions in recent years. Nevertheless, the information acquisition by Web Crawler is becoming more and more difficulty, which can be reflected by the technology transition from traditional Web resources to Deep Web (Hidden Web) resources [1], focused Web resources [2] and Deep Web resources collections based on AJAX [3], etc. As a new member of Web resources, spatial data is an important part of the Deep Web resources and research on spatial data sensitive crawler is very urgent. Some studies such as spatial data mining based on Web resources [4], data mining based on Web Crawler [5] and toponym database updating based on Web Crawler [6] reflect this trend from different aspects.

On the other hand, Web spatial information acquisition has close relationship among user application and post-data production. For spatial information retrieval, expression of information is more accurate and intuitive with the support of spatial data; and for data production, a variety of resources through the Web are available for spatial data updating and maintenance. However, for a research point of view, Web crawler has been very mature, but Web spatial Crawler is still in its infancy.

In this paper, a method of spatial data sensitive crawler is proposed based on the forms of vector data expression. The procedure of vector data collect and the strategy of measure of confidence level based on different types of vector data are discussed. At last, a test is employed to analyze the simulation accuracy of this method and some discussion and conclusions are presented.

II. THE THEORY OF SDSC TO COLLECT VECTOR DATA

A. Spatial Data Sensitive Crawler

Spatial data sensitive Web pages are the Web pages which have a higher correlation with spatial search terms [6], and by using the description of spatial text on this page, the traditional search engine can realize the classification, storage and retrieval of the page.

Spatial data sensitive crawler evolved from Web crawler, and aimed to build a database of spatial data via Web resources. It means that a preset collection of Web pages sensitive to spatial data, for each Web page in the collection, the crawler adopts its own spatial keyword, and finishes the text matching, spatial data exploration, data extraction and the measurement of spatial data correlation according to the page contents. Furthermore, the crawler tracks the links in the Web pages sensitive to spatial data until the end of all links. Yet, in comparison with the ordinary crawlers, the crawler sensitive to spatial data collects the links or resources related to spatial search terms selectively instead of tracking all links in the page.

B. Analysis of Spatial Data Sensitive Web Pages and Exploration of Vector Data

SDSC disposes the URL of each Web page sensitive to spatial data on the basis of the results from commercial engine spatial information matching list, such as Google Search API. The process is shown in Fig.1.

(1) Read the URL of the Web page sensitive to spatial data in the queue, detects the content of the page to ensure the correction of the tag matching, gradation. Then convert the Web page to the DOM mark tree which the root node is “HTML”, and remove the independent styles, scripts, and so on.

(2) Get the linked information of the HTML DOM tree, as well as the text information associated with them, verifying the text-related degree or containment relationship between the text and spatial search term segments (ICTCLAS for Chinese). For easy task, we get the text-related degree by judging the description text whether contains the keywords. If the

keywords are contained, we will have the further procession for the link.

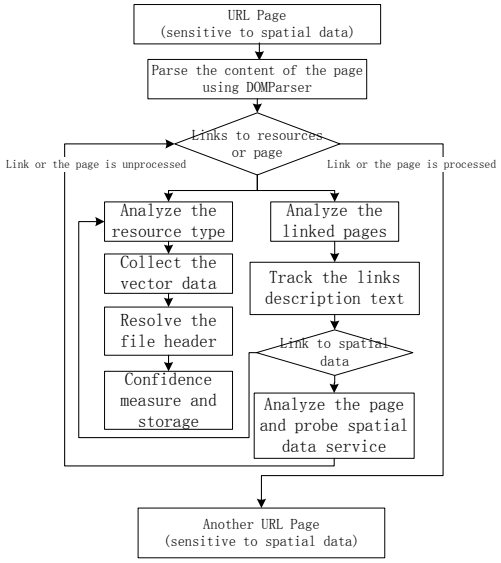

Fig.1. the procedure of the analysis of Web page by SDSC

The linked information contains spatial search term segments divided into links to resources and pages. The former is file resource and the latter is the URL of the Web page.

(3) The file resource (left branch) of the Web links involves various types of data, such as images, compressed files, PDF files, and so on. Among those data, including the vector data, which is distinguished according to the suffix, such as MIF, SHP, AI etc., then we should save the data to spatial data database, and resolve the file header, extract the spatial information, e.g. coordinate system, coverage area, etc. And at last, we should measure the confidence level by the resolve information, resource description and file format, and record the relevant information to database for later use.

Spatial data service is another type of vector data hidden in a Web page, which is represented by the OGC. OGC combined with ISO/TC211 standard, put forward the Interoperability Specification of spatial data based on Web services (OWS, OGC Web Service), which push the spatial data sharing to new heights [7]. OWSs (WMS, WFS, WCS, and WPS) have the vector data to realize the network sharing and processing based on XML and HTTP technology.

(4) Analysis on the URL of Web page (right branch) focuses on separating normal hyperlinks and spatial data services. Given trace level, there maybe a spatial data services if the page pointed to the same address repeatedly. Using SDSC to collect vector data service is relatively difficult by automatic program, so on the basis of the known data services, such as WFS，WMS，WCS，etc., it is still more desirable way to collect vector data by adjusting the strategy of SDSC. When the SDSC collects the feedback GML data purposefully and also can converts GML to known data format to show.

(5) Within the preset track level, the SDSC collects all the links in the Web pages sensitive to spatial data, saves related vector data, and records the relevant information to database until the end of track.

C. Measure of Confidence Level to the Vector Data

A good quality of spatial data is the premise of rational utilization. The quality evaluation model to traditional vector data, such as Weighted Defect Reduction Model, Weighted Average Model, Valuation Models based on the View of Fuzzy Set Theory, etc., relate to all aspects of vector data and there are still some problems in the quantization and measurement. As a main goal, the measure of confidence level of the vector data in Web environment is to identifying the correct format of vector data, the integrity of coordinated system, coordinate extends and non-spatial information, etc.

The Weighted Defect Reduction Model is a kind of quantitative methods to quantify vector data [8, 9]. However, it doesn't suite for the Web spatial data, because of the missing many metadata of collected data. On this basis, a defect reduction model has been improved so that the level of confidence of different vector data can be measured more accurately. The main principles are as follows:

(1) Based on the quality element of the Web spatial data, taking the subtraction score mechanism to set different degrees, such as the coordinated system, coordinate extends, other properties, and so on.

(2) Set different proportion for different quality elements.

(3) Measure the whole correlation with formula:

$$P(D) = T - \sum_{i=1}^{k}(W_i * F_i)$$

Where $P(D)$ and T are the final and initial value (1.0) of vector data quality confidence level. W_i and F_i are the weight and proportion of each quality element, k is the count of quality elements.

In order to facilitate the quantitative embodiment, considering the relevant elements of spatial data, give different elements different weighting coefficients, as shown in Table I. In order to reflect the difference in quality between the different data, set a larger difference in proportion between different types of quality element, as shown in Table II.

TABLE I. QUALITY ELEMENTS AND WEIGHTS (WI)

Index	**CS[a]**	**CE[b]**	**PR[c]**	**FC[d]**	**IO[e]**
Weight	0.3	0.2	0.2	0.15	0.15

a. Coordinated system, b. Coordinate extends, c. Properties relationship,

d. Format classification, e. Other information

TABLE II. ELEMENTS OF CLASSIFICATION AND PROPORTION (FI)

Quality element	**Element type**	**Proportion**
Coordinated system	Have coordinate system	0
	No coordinate system	0.6
	Unable to extract coordinated system	1.0
Coordinate extends	Extract directly	0
	Extract the contents	1.0
Properties	Have properties file	0

relationship	Have relationship but the properties not exist	0.5
	No properties file	1.0
Format classification	GIS data format	0
	Map data format	0.4
	Exchange data format	0.7
	Other data format	1.0
Other information	Suffix matching and readable	0
	Suffix matching but unreadable	0.5
	Suffix not matching	1.0

Since the measurement of confidence level to vector data involves the data production, distribution, the data elements and many aspects, so it is just some rough metric based on data file.

When we focus on the measurement of confidence level to spatial elements, we should select specific evaluation factors to specific application, such as Difference Curve methods to linear object [10], density, area, direction and point group similarity to point object [11]. So, the measurement of confidence to spatial elements reflects the quality of spatial data service and it's the overall merit to later use.

III. CASES AND ANALYSIS

The experiment is built under Nutch1.2 and MyEclipse9. In the latter platform, Nutch1.2 was improved to support spatial data acquisition. Here SDSC selects a city internal spatial vector data service interface (WFS) as its seeds, collects the hidden vector data, and assembles independent data into a whole data.

Standard experimental data is the center lines of the main road network in the city in 2009. There are 147 roads which width is greater than 30 meters, and their attribute information includes name, left connecting roads, right connecting roads and other information. The data of the experimental area is shown in Fig. 2 and Table III.

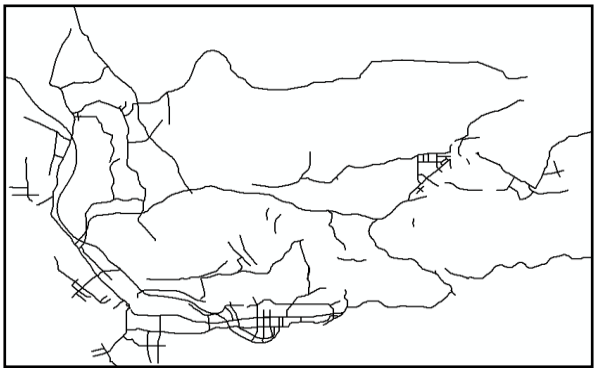

Fig.2. Roads in Experimental Area

TABLE III. INFORMATION OF STANDARD ROADS (UNIT: M)

Item	Information
Count	147
Min Width	30
Max Width	60
Avg Width	40
Min Length	21
Max Length	54896
Avg Length	5006

SDSC takes each road name (147) as WFS retrieval entrance, collects returned GML data, then gains a total of 280 road information and converted into common format (SHP). For those 280 vector roads, there are many problems, such as same roads different name, several roads for one name, etc. Then after merging and filtering these roads by ArcGIS, we finally get 134 vector roads with unique and complete information.

Comparison between standard roads and collected roads by SDSC is shown in Table IV. The number of collected roads is 13 less than that of the standard roads. There are possible reasons: ①some of road information is not published on Internet; ②do not match with the keywords of standard road, such as road name, so cannot build collect relationship; ③ SDSC only considers roads of the same name, and ignores the road whose name is changed.

TABLE IV. COMPARE THE STANDARD AND COLLECTED ROADS (UNIT: M)

Item	Standard Roads	Collected Roads
Count	147	134
Min Width	30	30
Max Width	60	68
Avg Width	40	37.5
Min Length	21	13
Max Length	54896	45906
Avg Length	5006	4730

Although those roads (134) came from updating by PC software, the vector data (280) collected from Web resources, and the measurement of confidence level to spatial data also has its value. The measure result of each road (GML) is 0.895, and updating file (SHP) is 0.90 by using the improved defect reduction model in this cases. Furthermore, Using Optimized Difference Curve [12] with $A(M) = 2e^{-1/M}$, the similarity of specific roads is shown in Table V, and the distribution of the road geometric similarity is shown in Fig. 3. Counting the road geometric similarity in the 7 ranges, 77.6% of roads are roads whose similarity is more than 70%, it means that the collected data can match with standard data well. In general, although there are inconsistent with WFS, those collected data also have higher confidence level after been updated, so the spatial data service has higher credibility.

TABLE V. THE SIMILARITY BETWEEN STANDARD AND COLLECTED ROADS

Similarity（SIM）	Count
100%	20
90%	26
80%	38
70%	32
60%	12
50%	4
30%	14
Total	134

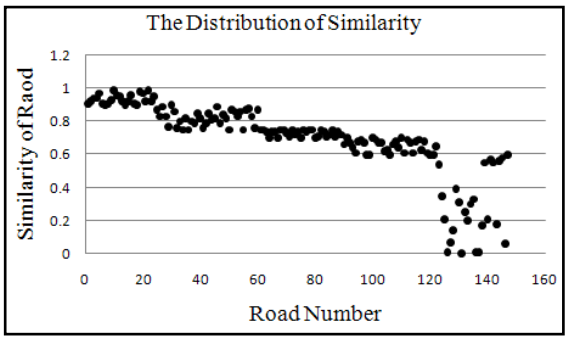

Fig.3. Distribution of Similarity of Roads

In order to show the difference between collected data and standard data, a slight shift is made to the collected data. And the effect is shown in Fig. 4, collected roads are red, and standard roads are black. By analyzing the result, there are excess parts in collected roads, such as part ①、④、⑤、⑥, and defective parts, such as part ②、③. It shows that the former are the new roads compare to the standard data, so they can update the parts in standard roads, and the latter are missing data because of road name changed.

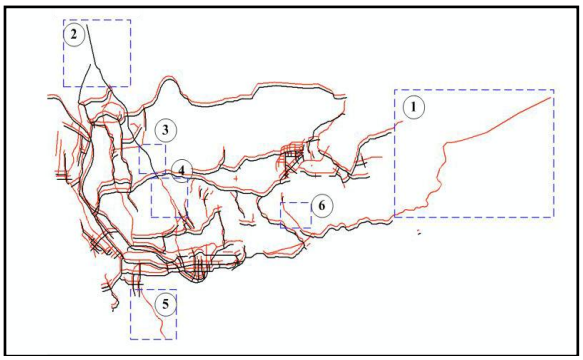

Fig.4. Collected Roads (Red) and Standard Roads (Black)

IV. CONCLUSIONS

Using SDSC to collect the vector data in the Web page is an effective means of establishing regional spatial database, and can provide a reference for the later data updating. Separating the vector data from the Web page, analyzing the data through the crawler internal method and evaluating the quality of the vector data to a certain extent are several key problems to the vector data acquisition. The experiment collects the road data by SDSC, and verified the effectiveness of vector data collection and evaluation. However, due to the complexity of Web spatial data and the difficulty of establishing a rational and efficient evaluation system to spatial data, more theory should be established on the acquisition of spatial data by SDSC.

V. ACKNOWLEDGMENT

I want to thank Cai Zhongliang, Weng Min, Yang Yan and Liu Xiao for their helpful comments on an earlier version of this paper and improving the style of this paper.

REFERENCES

[1] Bergman, Michael K., "White paper: the deep web surfacing hidden value," J. The Journal of Electronic Publishing, vol. 7, 2001.

[2] H. Y. Liu, J. Janssen and E. Milios, "Using hmm to learn user browsing patterns for focused web crawling," J. Data & Knowledge Engineering, vol. 59, pp. 270-291, 2006.

[3] A. Mesbah and A. Deursen, "A componet- and push-based architectural style for ajax applications," J. Journal of Systems and Software, vol. 81, pp.2194-2209, 2008.

[4] M. Fu, "The research of web-based spatial data mining," D. Central South University, 2004.

[5] M. Thelwall, "A web crawler design for data mining," J. Journal of Information Science, vol. 27, pp.319-325, 2001.

[6] C. J. Zhang, X. Y. Zhang, S. A. Zhu and X. T. Xu, "Method of toponym database updating based on web crawler," J. Journal of Geo-Information Science, vol. 4, pp.492-499, 2011.

[7] http://www.opengeospatial.org/standards/

[8] Q. Zhu, S. L. Chen and D. Huang, "Key issues on quality standardization of geospatial data," J. Geomatics and Information Science of Whuhan University, vol. 10, pp.863-866, 2004.

[9]F. F. Wang, "The evaluation of spatial data quality and quality control system," D. Sichuan University, 2005.

[10] E. M. Knorr, R. T. Ng, and D. L. Shilvock, "Finding boundary shape matching relationships in spatial data," in The 5th International Symposium (SSD'97). vol. 1262, pp. 29-46, 1997, unpublished.

[11] Y. Y. Mei, H. W. Yan and Q. Li, "Research on similarity relations of point features in multi-scale geographic space," J. GeoMatics & Spatial Information Technology, vol. 2, pp. 18-21, 2010.

[12] L. L. Tang, B. S. Yang and K. M. Xu, "The road data change detection based on linear shape similarity," J. Geomatics and Information Science of Whuhan University, vol. 4, pp.367-370, 2008.