International Cartographic Association
Association Cartographique Internationale

Research on Deep Web POI Acquisition based on Retrieving Word Optimization and Spatial Adaption

Wang yong *, Luo an, Jiping Liu, Yuanhui Cao
Chinese Academy of Surveying and Mapping; No. 28 Lianhuachi West Road, Haidian District, Beijing100830, China;
e-mail: wangyong@casm.ac.cn;
*Corresponding author

Abstract: POI data is a geographical information resource that is the most closely related to the public life, and has been successfully applied in various fields such as urban planning, urban logistics, and car navigation. With the development of technologies such as mobile networks and Internet of Things, the network contains a large number of high-value POI information resources. How to effectively acquire and utilize data resources has become a research hotspot in the field of spatial information. In this paper, a deep-web POI information search method based on independent coverage ranking and spatial adaptive partition is proposed to solve the problems of difficult construction of retrieval word base and limited data request. By constructing candidate search terms, searching greedily, optimizing dimensionality reduction of search terms, and crawling spatially adaptive partitioning, the maximum coverage optimal solution of POI search is approached step by step, and the full POI information of deep web is obtained. It is of great significance to improve the recall rate and collection efficiency of POI data for enriching geographic information resources and improving the ability of spatial information service and content management.

Keywords: Deep web POI, data collection, retrieving word optimization, spatial adaptive subdivision

1. Introduction

In recent years, with the rapid development of information technology, mobile communication technology and so on, the demand of location-based service rises sharply, and the data containing location information presents explosive growth in personal information service, scientific research and other fields. As an important expression element in topographical map and navigation maps, POI is a kind of point data representing real geographical entities. Its data volume, accuracy and current situation largely determine the quality of spatial location services [1]. With the rapid development of the social economy and the acceleration of the urbanization process at the present stage, the POI data in close relation to the position of the city is constantly increased and changed, so the network provides large number of POI data with high potentials, fast propagation and abundant information, and the scale and quality are improved constantly. The extraction of POI data from the network becomes an important means for obtaining data in the multiple areas such as application of urban planning.

Researchers at home and abroad have carried out relevant research around the issues of network information acquisition and so on[2-5]. Liu Wei et al. proposed a graph database-based web database sampling method. Approximate random samples returned from Web database are obtained incrementally through query interface, and query iteration is carried out by using original localized sample records. This method obtains relatively high correlation results at a small cost, but it is difficult to realize deep web data. Maximum search of Library and manual setting of sampling parameters[6]. George V et al. proposed a Rank-aware Hidden Web crawling method, which can achieve the maximum relevance of the results of deep web queries and has high performance, but can not achieve the maximum coverage crawling of deep web databases[7]. L Barbosa et al. proposed a query traversal method based on high-frequency words. By counting the frequency of occurrence of query words in the returned results, high-frequency words were selected as query words for subsequent queries. The disadvantage is that high-frequency words cannot ensure acquisition Multiple query results[8]. From the angle of reducing duplicate document extraction, Lu Jianguo et al. crawled the Web database by using deep-net database sampling method. First, they acquired part of the sample data, then selected a set of query words with lower duplicate documents in the sample data, and used the set of query words to crawl the content of the target Web database. the disadvantage is that the query words obtained only by sampling may not completely cover all the records in the Web database [9].

Comprehensive research at home and abroad can find that for most POI data in deep network, it is difficult to obtain effective POI data for general search engine and ordinary deep network crawling method. The main reason is that there are many kinds of POI data, and different websites use different classification systems [10-11]. As the only access to deep-web POI database, the retrieval interface needs to Submit search terms to access deep-web POI data resources. It is an important challenge to construct a set of candidate search terms with the largest query coverage. In addition, for deep-net POI data query, spatial range is a necessary input condition. In the actual implementation

Proceedings of the International Cartographic Association, 2, 2019.
29th International Cartographic Conference (ICC 2019), 15–20 July 2019, Tokyo, Japan. This contribution underwent single-blind peer review based on submitted abstracts. https://doi.org/10.5194/ica-proc-2-142-2019

process of deep-net POI acquisition, if the spatial range submitted by crawlers is too large, the vast majority of background data services often can not return all the results list that meet the requirements, How to achieve "full access" to all geographic range data is also an urgent problem to be solved.

Therefore, in order to solve these problems, this paper proposes a deep-web POI information search method based on independent coverage ranking and spatial adaptive partitioning. Through the main processes of preliminary construction of candidate search terms, searching greedily, optimizing dimensionality reduction of search terms, and crawling spatially adaptive partitioning, this method uses gradually approaching POI search. the maximum coverage optimal solution of POI search is approached step by step, and the full POI information of deep web is obtained., which is of great significance for enriching geographic information resources, improving spatial information services and content management capabilities.

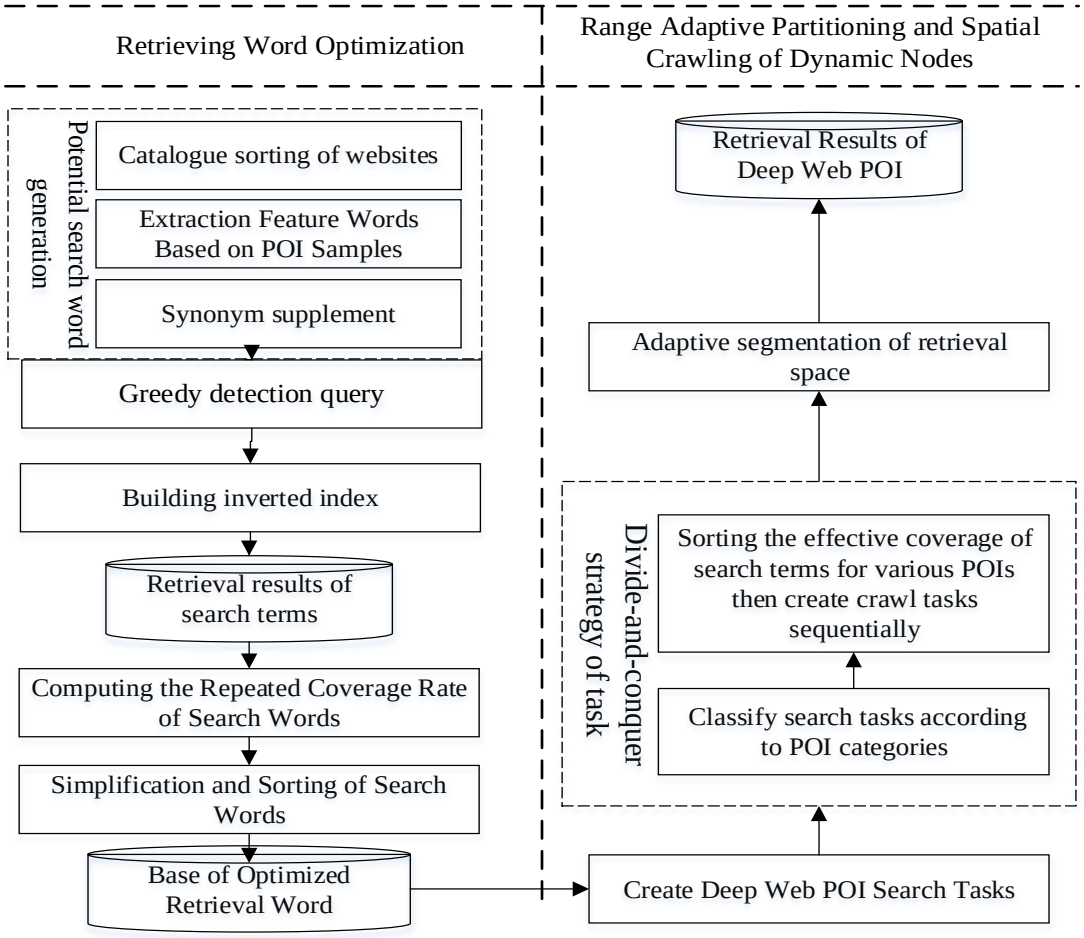


Fig.1 Retrieval word optimization and spatially adaptive deep web POI acquisition process

2. Deep Web POI Acquisition based on Retrieving Word Optimization and Spatial Adaptive

In theory，the core problem of deep web POI is to find two optimal solution。Firstly，the most compact search word sequence which can take over all kinds of conditions should be developed as compact as possible. Secondly，the number of space units should develop as small as possible on condition that the space traversal can obtain as much POI as possible. Therefore, this paper proposes a deep-web POI information search method based on independent coverage ranking and spatial adaptive partitioning. The method can get the full amount of POI information of deep network through the main processes of preliminary construction of candidate search terms, greedy search, optimization of search terms, dimensionality reduction, spatial adaptive partitioning and crawling, the maximum coverage optimal solution of POI search is approached step by step. The main flow is shown in Figure 1 below.

2.1 Potential search word generation

Retrieval interface is the only access to deep-web POI database, and search term is the most important input condition for obtaining deep-web POI data. In order to achieve the maximum coverage of POI crawler search results on Deep Web POI information resources, we first need to construct a set of candidate search terms with sufficient coverage. The construction methods of potential search terms include: sorting out the classification catalogue of websites and obtaining the classification names; extracting classification feature words based on POI samples; and supplement potential search term by synonyms. Experience shows that classification names and

Proceedings of the International Cartographic Association, 2, 2019.
29th International Cartographic Conference (ICC 2019), 15–20 July 2019, Tokyo, Japan. This contribution underwent single-blind peer review based on submitted abstracts. https://doi.org/10.5194/ica-proc-2-142-2019 | © Authors 2019. CC BY 4.0 License.

classification feature words have the greatest contribution to POI information acquisition, and also constitute the main content of the potential search term set.

2.2 Greedy detect and query

By exploring and querying specific deep-web POI databases with sufficient coverage search terms, we can accurately evaluate the contribution of each search term to the crawling results. The main idea is to select the test area，construct query requests using existing potential search terms, and greedy search the deep-web POI databases. The POI data of each query request and its response are recorded completely, and the data is prepared for the subsequent optimization of search terms.

Using the idea of inverted index to form a retrieval record table by recording the retrieval history of each search term. As shown in Table 1 below, the first column is POI returned in the process of greedy detection query, the second column is its corresponding type, and the third column is the search term that retrieves the corresponding POI.

Tab.1 Retrieval record table

Objects（P）	Type（T）	Search term（ST）
POI_1	T_1	Q_1、Q_2
POI_2	T_2	Q_1、Q_2、Q_3、Q_4
POI_3	T_3	Q_2、Q_4
POI_4	T_2	Q_1、Q_3

In order to facilitate the calculation of the retrieval effect of a certain search term on a certain type of POI, a retrieval inverted index table is constructed based on the retrieval record table. The association between the POI type and the search term is formed by merging the records of the search record table according to the type field T, and the inverted index table is as shown in Table 2 below.

Tab.2 Inverted index table

Type（T）	Search term and hit number
T_1	Q_1（1）、Q_2（1）
T_2	Q_1（2）、Q_2（1）、Q_3（2）、Q_4（1）
T_3	Q_2（2）、Q_4（1）

Based on the inverted index table constructed, the retrieval performance index of the search term can be calculated by database retrieval technology, and the indexes such as hit number, hit rate and repeated hit rate of the search words can be quickly obtained, which is convenient for the subsequent optimization of search words.

2.3 The retrieval word sequence is generated based on repeated overlay iterative computation

Due to the limitation of network bandwidth and other conditions in the process of web crawler, in order to reduce the query cost of search words and obtain acceptable crawling coverage rate, the set of potential search words must be reduced and optimized. According to the existing POI classification, under a certain threshold condition of repeated coverage, the collection of crawling search terms corresponding to various POIs is calculated through multiple iterations to realize the "simplification" and "sort" of the candidate search terms. The flow is shown in figure 2 below. Constructing optimized retrieval thesaurus and abandon retrieval thesaurus, extracting the most significant search terms in the potential search term sequence, and comparing the repeated coverage rate with the repeated coverage rate of the optimized search lexicon. If it is greater than it, it will be put into the discard search. On the other hand, it will be placed in the optimized search term. Repeat the above steps until the potential search term is empty. Calculate the overall coverage rate of the optimized search term database and compare it with the threshold which has been set. If it is greater than the threshold, the optimized search term database will be constructed successfully; otherwise, the threshold shall be adjusted and all the above operations shall be repeated.

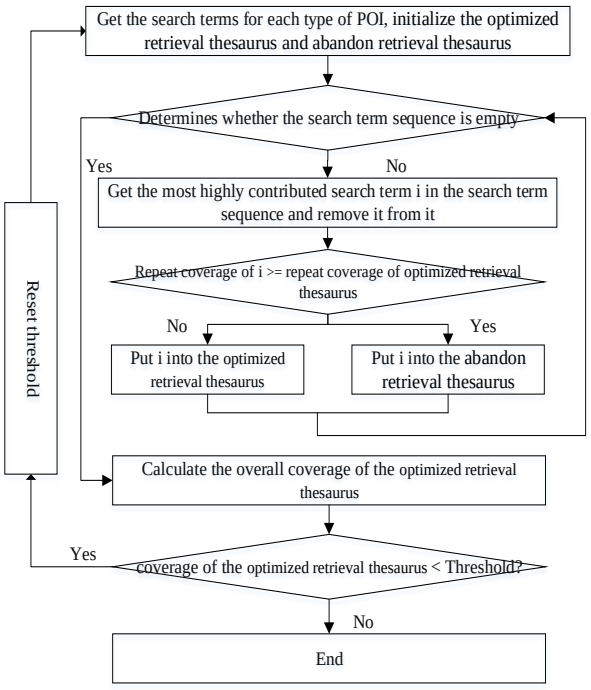


Fig.2 Retrieval word generation process based on repeated coverage iterative computation

2.4 Space range adaptive segmentation for space creep

The crawling strategy is the core mechanism of deep web POI crawler operation, which determines the ability and efficiency of crawler to run. Due to the variety of POIs, task divide-and-conquer strategy is adopted to ensure data acquisition efficiency and crawling coverage. Search tasks are divided according to POI categories, and the corresponding crawl tasks are sequentially created after sorting the effective coverage rate of various POI search words. On the basis of spatial segmentation (split) according to a specific rule in the specified geospatial range, the process of information retrieval and acquisition

Proceedings of the International Cartographic Association, 2, 2019.
29th International Cartographic Conference (ICC 2019), 15–20 July 2019, Tokyo, Japan. This contribution underwent single-blind peer review based on submitted abstracts. https://doi.org/10.5194/ica-proc-2-142-2019 | © Authors 2019. CC BY 4.0 License.

according to a traversal strategy follows the definition of the formal crawling task as shown in the following formula 1:

$$CrawTask = (ST, Bounds, Limit) \quad \text{formula (1)}$$

Where ST represents the search term, Bounds represents the spatial search range corresponding to the task, and Limit represents the maximum number of records that the deep net POI database can return for a single search term. For specific spatial crawling tasks, ST and Limit are constants, and the algorithm of information traversal for a given spatial range, Bounds, determines the ability and efficiency of POI data acquisition. The primary goal of spatial crawling is to achieve maximum access to POI information. Therefore, this paper proposes a spatial crawling method for spatially adaptive segmentation, the flow of which is shown in Figure 3. For the search terms of each subtask obtained by the task division strategy, record the total number of pieces obtained in a given space, and compare it with Limit. If it is less than Limit, go to the next subtask, on the other hand, indicating that the current geographical scope is too large, it will be divided by 2×2, and generate four subtasks, at the same time the current task is removed.

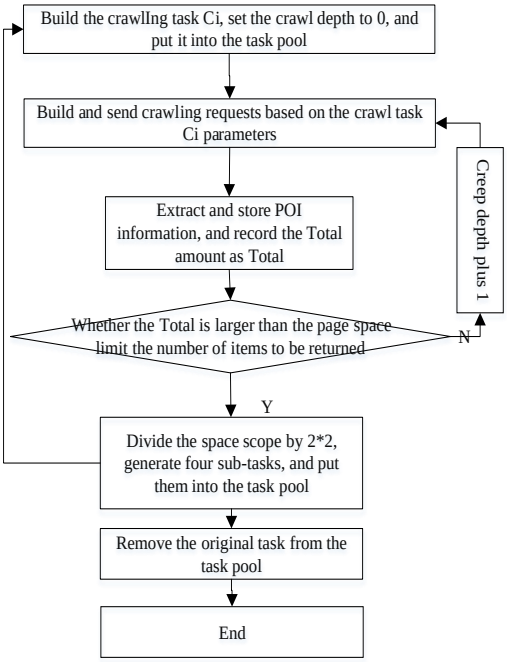


Fig.3 Spatial crawling process based on adaptive partition of spatial scope

3. Result and Analysis

The paper carried out the experiment of search term generation and POI network acquisition in the case of Education project in two large map service websites, and then assess the crawling result. As shown in the figure 4 below, the paper take the geographical area of 30Km * 30Km within the boundary of Beijing fifth ring road as the climbing target area (the solid blue lines in the figure demonstrate the circumference), and three 1.5km x 1.5km rectangular regions (shown in the red solid line) within the target region were selected as the training areas to Generate potential search words and optimize search sequences.

Fig.4 Test area

3.1 Optimize search word generation

First, by adding the category names related to the "education" type from the two website classification directories, they are added to the candidate search term queue; Then, in the three training areas, the "education" keywords are input to the two websites based on the manual query method, and the keyword extraction algorithm is used to generate 21 keywords from the names of the POI sample sets; Based on the synonym acquisition method, four synonymous (near sense) words are obtained and included in the candidate keyword set. In the end, a total of 32 candidate keywords were obtained. The results are shown in Table 3 below.

Tab.3 Educational candidate search words set and its source

Website / Source	Website A	Website B
Category Name	Education, kindergarten, primary school, middle school, university, training (6)	Universities, secondary schools, primary schools,kindergartens, training, vocational and technical schools, adult education (7)
Feature word extraction result	College, headquarters, branch, branch school, campus, college, technical secondary school, junior college, vocational high school, vocational high school, technical school, technical school, party school, evening school, adult education, driving school, continuing	

Proceedings of the International Cartographic Association, 2, 2019.
29th International Cartographic Conference (ICC 2019), 15–20 July 2019, Tokyo, Japan. This contribution underwent single-blind peer review based on submitted abstracts. https://doi.org/10.5194/ica-proc-2-142-2019 | © Authors 2019. CC BY 4.0 License.

	education, private school, education group, network school, nursery school（21）
Synonym acquisition result	Universities, colleges, adult education, kindergartens (4)
Total	**32**

The above-mentioned search terms were used to obtain automatic information acquisition based on greedy algorithm for the three training areas. 476 and 423 return results were obtained on the two websites respectively, and the search hit rate of each candidate search word of the two websites was obtained.

Tab.4 Statistics on hit percentage of search words in test area

Index	A Website	B Website
Retrieve the total number of hits	476	423
Candidate search term hit rate ordering	School 346, Training 226, Education 308, Kindergarten 108, Primary School 97, Adult Education 86, Kindergarten 84, Middle School 78, University 71, Continuing Education 68, University 55, Nursery 28, Party School 19, Driving School 12, Adult Education 3	Education|School 354, Training 196, University|University134, University 80, Kindergarten|Kindergarten 49, Middle School 46, Elementary School 35, Continuing Education 29, Adult Education| Adult Education 26, Nursery School 24, Driving School 14, Party School 11

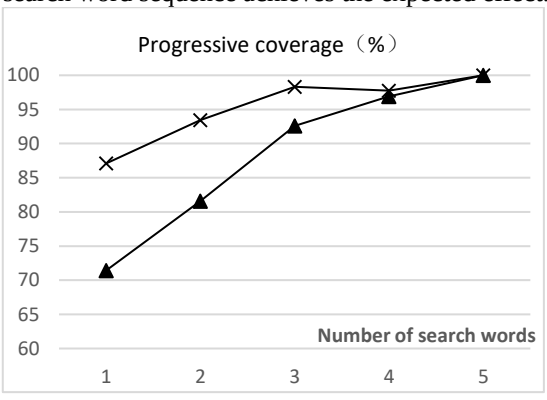
The crawl coverage rate is set to four thresholds of 0.6, 0.7, 0.8, and 0.9. The loop iteration algorithm is used to select the crawling keywords, and the keyword sequences corresponding to the thresholds are obtained respectively, as shown in Table 5 below. It can be seen from the results that: (1) With the increase of coverage requirement, the number of search terms included in crawling is more; (2) The search terms with higher effective hit rate may have larger results with the pre-order search terms. Repeat coverage (even if its search result is a subset of pre-search terms), so it is not necessarily included in the crawl search term sequence. For example, the “training” in the Sky Map website is not included in the coverage of no more than 0.9.

Tab.5 Crawling retrieval word set generated under different coverage rates

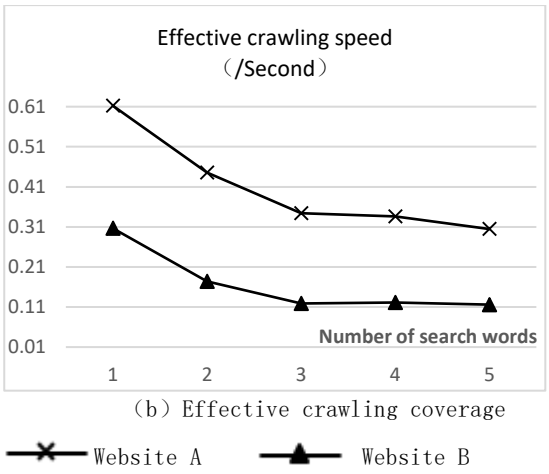
Coverage threshold	**A Website**	**B Website**
0.6	School	Education|School
0.7	School, Adult Education	Education|School
0.8	School, Adult Education, Colleges	Education|School, Training
0.9	School, Adult Education, Colleges, Driving School	Education|School, Training, Nursery
0.95	School, Adult Education, Colleges, Driving School, Continuing Education, Training	Education|School, Training, Nursery, Continuing Education

3.2 Crawling Efficiency

The top five search words in the optimized search words list were chose to test the deep web crawling software, which would collect information from the two websites. In order to avoid abnormal access pressure on the target website, a random sleep time of 10-15 seconds was set between every two requests during crawling. Since the raw data files of the backend database were not available, this paper used the progressive coverage rate to estimate the coverage of the entire crawling process approximatively. As can be seen from Figure 5(a), the growth rate of progressive coverage gradually decreases with the search words being added one by one. This means we can assume that the crawling results were gradually approaching the backend database while the number of new POIs was gradually decreasing. Meanwhile, through the method of manual search and crawler retrieval, this paper took the school as an example to sample and compare in Wuhan, Nanjing, Kunming, etc., and found that the actual coverage rate can reach 87%~93%. As can be seen from figure 5 (b), the effective crawling speed decreases gradually with the addition of search terms, indicating that the number of newly retrieved POI decreases gradually in the subsequent crawling process. From the trend line, the speed curve corresponding to the website B is more stable than the website A, indicating that the new object in the website B is more "uniformly" detected, and thus the corresponding search word sequence achieves the expected effect.



（a）Progressive coverage



（b）Effective crawling coverage

Fig.5 Coverage and speed of crawlers on different websites

4. Conclusion

In order to solve the problems of difficulty in constructing the search words database and limited data request in the process of crawling deep network POI data, this paper proposes a deep network POI information search method based on independent coverage sorting and spatial adaptive partitioning. Through the preliminary construction of candidate search words, greedy detection search, search word optimization dimension reduction, spatial adaptive segmentation crawling and other major processes, the full coverage of deep network POI information is realized by gradually approximating the maximum coverage optimal solution of POI search. Taking the typical deep network POI service as an example to carry out technical experiments, the effect of crawling coverage was 87%~93%, and the recall rate and collection efficiency of deep network POI data were greatly improved, which has a great significance for enriching geographic information resources and improving spatial information services and content management capabilities. The next step is to improve the data acquisition efficiency of the method in terms of the inefficiency in the greedy detection query process and the spatial distribution characteristics of the POI in the non-urban area.

In order to solve the problems of difficulty in constructing the search words database and limited data request in the process of crawling deep network POI data, this paper proposes a deep network POI information search method based on independent coverage sorting and spatial adaptive partitioning. Through the preliminary construction of candidate search words, greedy detection search, search word optimization dimension reduction, spatial adaptive segmentation crawling and other major processes, the full coverage of deep network POI information is realized by gradually approximating the maximum coverage optimal solution of POI search. Taking the typical deep

5. Acknowledgement

This work was supported by the National Key Research and Development Program of China（NO.2017YFB0503502,2017YFB0503601）and Basic scientific research operating expenses of the Chinese academy of surveying and mapping(7771817).

6. References

[1] JIANG Rui. Research on POI information acquisition method in web page text[D]. Nanjing Normal University,2012.

[2] TIAN Jianwei, LI Shijun. Retrieving deep web data based on hierarchy tree model[J]. Journal of Computer Research and Development, 2011, 48(1):94-102.

[3] HOU Dongyang，WU Han，WANG Junfeng.A Web Map Service Discovery Method Based on Deep Web Crawler[J]. Geography and Geo-Information Science, 2015, 31(5):10-13.

[4] Cao X, Klusch M. Advanced Semantic Deep Search for 3D Scenes[C]// IEEE Seventh International Conference on Semantic Computing. IEEE Computer Society, 2013:236-243.

[5] LI Yanni. Modeling and Algorithm Research of key issues in Deep Web data integration and mining[D]. Xi'an Electronic and Science University,2013.

[6] LIU Wei, MENG Xiaofeng, MENG Weiyi. A survey of deep web data integration[J]. Chinese Journal of Computers, 2007, 30(9):1475-1489.

[7] George V, Alexandros N, Dimitrios G. 2011. Rank-aware crawling of hidden Web sites [C]. Proceedings of the international workshop on the web and databases. Athens, Greece.

[8] L Barbosa, J Freire. 2004. Siphoning hidden-Web data through keyword-based interfaees, In:XIX Simposio Brasileiro de Bancosde Dados, Distrito Federal, Brazil, Anais, 309-321.

[9] Lu J, Wang Y, Liang J, et al. An Approach to Deep Web Crawling by Sampling[C]// Ieee/wic/acm International Conference on Web Intelligence and Intelligent Agent Technology. IEEE Computer Society, 2008:718-724.

[10] ZHANG Ling. Research on POI classification standard[J]. Bulletin of Surveying and Mapping, 2012(10):82-84.

[11] HOU Dongyang . Method of Land Cover Web Information Discovery[J] . Acta Geodaetica et Cartographica Sinica,2017,46(1):133.

[12] ZHENG Dongdong, ZHAO Pengpeng, CUI Zhiming. On the research and design of deep web crawler[J]. Journal of Tsinghua University(Science and Technology) , 2005, 45(9):1896-1902.

[13] WANG Yong . Research on Crawling and Consistency Processing of POIs form Deep Web[J]. Acta Geodaetica et Cartographica Sinica, 2017，46(3)：399.