ICIC Express Letters
Part B: Applications
Volume **7**, Number **7**, July **2016**

ICIC International ©2016 ISSN 2185-2766
pp. 1591–1598

RESEARCH ON CONSTRUCTION OF INTEGRATED SEMANTIC CRAWLER

Xingwei Gong and Yao Liu*

Institute of Scientific and Technical Information of China
No. 15, Fuxing Road, Beijing 100038, P. R. China
*Corresponding author: liuy@istic.ac.cn

Received December 2015; accepted March 2016

Abstract. *A method of constructing integrated semantic crawler is proposed in this paper to make up for the deficiency in former constructing methods. Semantic crawler developed by this method can generate one kind of semantic structure automatically from one concept. Semantic structure can guide the Web crawler; meanwhile, the system continuously extracts relevant concepts and relations between them and fills them into the semantic structure. The system realizes the integrated operation mechanism with better adaptability of semantic structure evolution and crawling.*

Keywords: Integration, Semantic crawler, Semantic structure

1. **Introduction.** Information explosion leads to a constant raise of cost for obtaining the required information resources. Web crawler is an important tool which can collect information automatically. Focused crawler can consume less system and network resource than common Web crawler and obviously improve the utilization of Web pages at the same time. But when facing polysemy or synonym, it might collect excessive noise pages or miss related pages. To solve this problem, scholars raised a concept of semantic crawler based on focused crawler. They utilize semantic reasoning and semantic relevance calculation to decrease the search time and improve accuracy. Man-made thesaurus, ontologies or other semantic structures are commonly used to construct semantic crawler but this method has obvious defects, such as high cost, low efficiency and worse adaptability. This paper raised a new method to construct integrated semantic crawler borrowing idea from automatic construction of ontology. This kind of crawler has a strong adaptability, which can understand the requirement of users and generate a simple semantic structure automatically from a concept. The structure guides the crawler. Relevant concepts and relations are extracted from Web pages and used to expand the semantic structure when the crawler crawling. An integrated semantic crawler system which contains two circulations, crawler crawling and semantic structure evolution, is constructed.

2. Research Status Both Domestic and Overseas.

2.1. **Construction of topic model.** The relevance between the Web pages and the given topic and the relevance among the links contained in the pages are based on the topic model; therefore, topic model with poor quality may drop a large number of related pages.

Common topic models include Boolean model, vector space model, probability model, etc. However, these models do not include relationships of semantic meaning and concepts of keywords. Some scholars raised several topic models based on semantic structure to solve this problem. For example, Rungsawang and Angkawattanawit [1] designed a topic model based on knowledge base. Xie [2] raised a method to describe topic concept by constructing concept tree based on thesaurus. Liu [3] presented a topic-specific competitive intelligence acquisition system based on LDA and domain ontology, which integrates the content analysis and the link analysis.

Taken together, the method using semantic structure is the future trend of constructing topic model because of its advantage and favorable effect.

2.2. Crawling strategy.

Effective crawling strategy can make the crawler collect related Web pages faster. Batsakis et al. [4] divided the topic crawling strategy into common topic crawling strategy, and improved topic crawling strategy and semantic meaning-based topic crawling strategy.

Common topic crawling strategy mainly refers to the "Fish Search" strategy. Improved topic crawling strategy includes the "Shark Search" strategy and the "Best First" strategy. The advanced topic crawling strategy is based on semantic meaning. Related research work includes that: Ehrig and Maedche [5] proposed an approach for document discovery building on a comprehensive framework for ontology-focused crawling of Web documents which had shown promising results. Du et al. [6] proposed a concept context graph to store the knowledge context based on the user's history of clicked Web pages and to guide a focused crawler for the next crawling.

The scholars, however, do not dig deep into the highly integrated semantic crawler with capabilities of ontology self-built and ontology evolution, but merely into the use of existing ontology or man-made ontology. Since ontology construction is a complex system which needs domain experts to carry out. It has high cost, long time consumption and poor adaptability. Wide spread construction cannot be achieved. Therefore, how to generate semantic structure automatically is an important research trend in the coming period.

2.3. Extraction algorithm for Web information.

Information extraction is to process the Web pages and extract the main body according to some rules.

Related research includes NoDoSE [7] and DEByE [8] system which utilize the character of the same structure in some similar Web pages. An and Xu [9] raised a vision-based information extraction method. Wang and Xu [10] proposed an approach based on CURE algorithm of Web pages segmentation and text extraction rules. Yang et al. [11] proposed an improved DOM based content extraction method by using point density to replace text density which obtains better effect.

3. Key Issues to be Solved.

3.1. Semantic structure of automatic generation.

For the semantic structure of some concepts which we cannot get from existing knowledge base, the crawler needs to auto-construct semantic structure by using the information from the results of search engine.

3.2. Appraisal of semantic structure.

Appraisal mechanism should be introduced into the crawler system to decide when to stop the semantic structure evolution iteration to avoid lower association caused by endless iteration.

4. Ideas and Basic Methods.

The process and structure of the integrated semantic crawler can be shown in Figure 1.

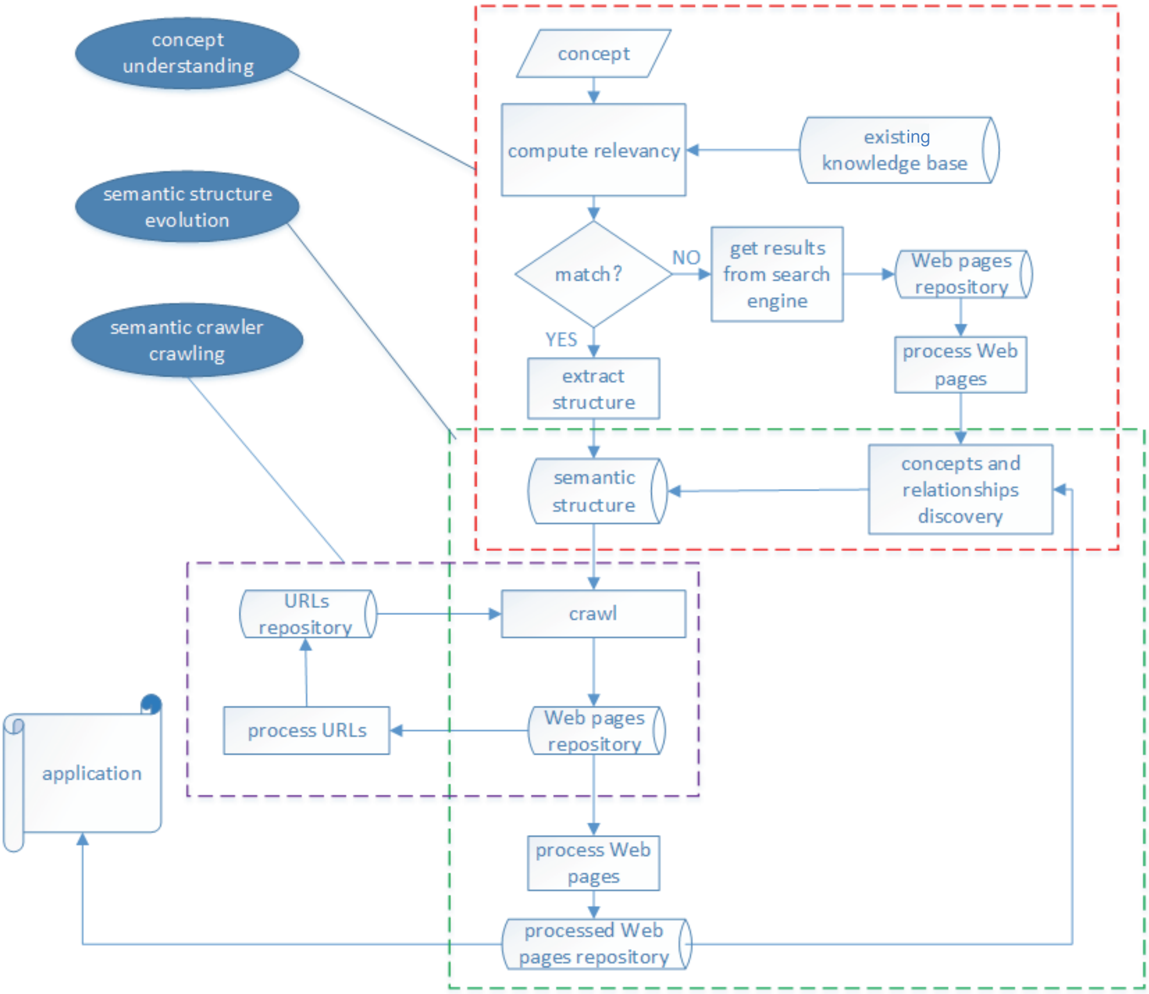

Figure 1. Process and structure

4.1. **Concept understanding.** Firstly, understand the given concept, and compute relevance with those concepts in existing knowledge base. If matched concept founded, extract related structure; otherwise, get some results of search engine by using search engine APIs. Then we extract related concepts by computing mutual information and $tf * idf$ from those results. At last, the crawler extracts the relationships of those concepts and constructs semantic structure automatically by using the hierarchical agglomerative clustering algorithm.

4.2. **Semantic crawler crawling.** Use the semantic structure to direct crawler crawling, extract the text information from each crawled page and compute relevance with the semantic structure. Then save the related Web pages, meanwhile, extract the links in those pages and compute their relevance by comparing the anchor texts or Web page titles with the concepts in the semantic structure. Only those links who have higher relevance than the threshold can be used to crawl.

4.3. **Semantic structure evolution.** Extract the concepts and relationships from the processed Web pages repository and expand the semantic structure by computing relevance with the concepts of the semantic structure while the crawler crawling. In this way, the semantic structure evolves automatically. Then it can be used to direct crawler constantly.

5. **Research Progress.** Our team did some researches on related technologies and developed a crawler system, which is capable of semantic crawling to some extent. The

system is based on C/S architecture, its background is realized by Java language and the interface of the system is developed by GUI. The crawler system uses MySQL5.2 and runs on Windows platforms. The computer needs Java runtime environment and must have a speed of the net not lower than 20kb/s.

This system offers five modules, i.e., common crawler, smart crawler (semantic crawler), special crawler, task scheduling and optional components. The core module is smart crawler (semantic crawler), which includes the global network smart crawler (see Figure 2) and keywords deep-crawling smart crawler (see Figure 3). The former function can crawl Web pages from the Users' keywords. It can get the results URLs list of the search engine as a crawling entrance and compute the relevance between the anchor texts or Web page titles with the concepts in the semantic structure. We can control the crawling depth, use regular expression to filter Web pages and decide whether to extract content and generate index or not. Simultaneously, the semantic structure evolution mechanism [12] is also integrated into this function. The latter function can crawl the global website upon seed pages and filter the URLs by using the given keywords. The crawling depth, number of threads and regular expression are also controllable. Further, we also developed crawlers oriented at special domain, such as WEIBO, thesis, news and patent (see Figure 4).

Moreover, we developed an ontology construction platform [13] in pre-work and constructed a complete ontology structure in chemical industry field which can be used to develop the crawler.

We also proposed a new extraction algorithm for Web information. It determines the beginning and the end of the main body by analyzing character density.

We raised the following equation to compute the relevance:

$$score(q,d) = coord(q,d) \times \sum_{t\,in\,q} (tf(t\,in\,d) \times idf(t)^2 \times t.getBoost() \times norm(t,d))$$

In the equation, tf is frequency of the concept, *idf* is inverse document frequency, *Boost* is a score which represents the weight of the concepts and its attribute, *norm* is a normalization value and *coord* is a cooperative factor which is based on the number of the search term in the document.

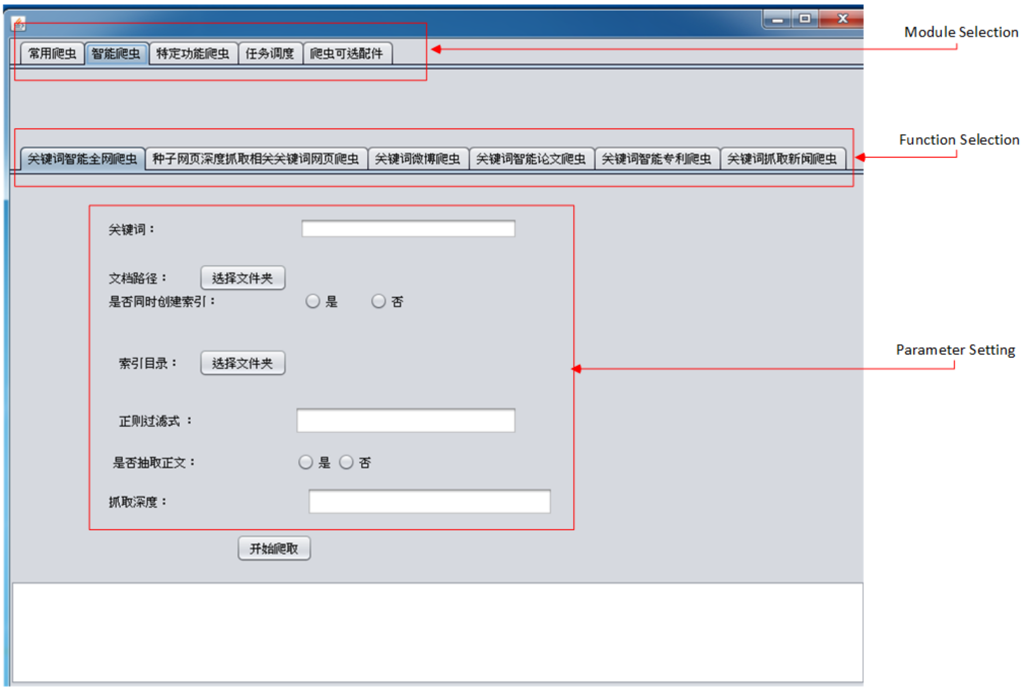

FIGURE 2. Global network smart crawler

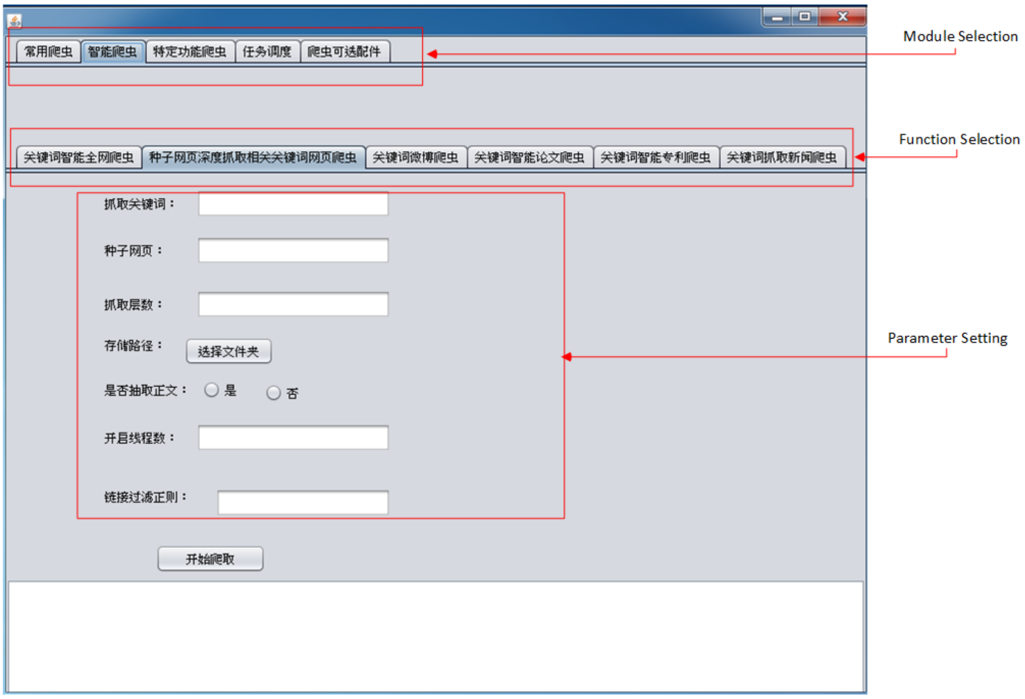

Figure 3. Keywords deep-crawling smart crawler

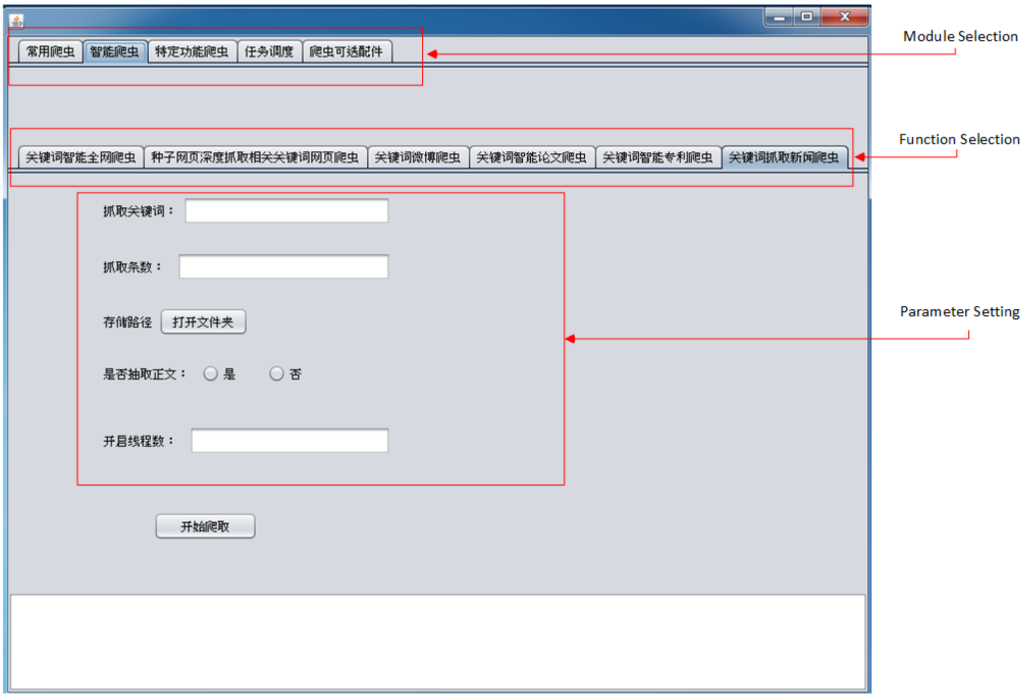

Figure 4. WEIBO smart crawler

6. **Experiment.** To test and verify the effect of semantic crawler we developed in prework, a comparison experiment is designed. We choose "sulphuric acid" as the subject and compare the effect of the global network smart crawler in the above system with a famous focused crawler in China.

Experiment environment: Windows 7+Tomcat 7.0+JDK7

Hardware configuration: CPU: AMD Athlon$^{\mathrm{TM}}$ x4 750 3.4GHz, Memory: 4G, Network bandwidth: 10M

Our semantic crawler uses the top-ten result pages of Baidu, Sogou, Bing, HaoSou and Google as the seed pages and finishes crawling when it reaches the closed-loop. The number of threads is 50. The crawling speed is about 20,000 pages per hour. The threshold of the conceptual relevance is 0.9 and the threshold of the attribute relevance is 0.4. The crawler obtains 5138 pages and the number of unfiltered pages is 61000. The effective rate is about 8%. For the famous focused crawler, we choose chnlsw.com, china.guidechem.com/15373, liusuan.100ppi.com and the top-ten result pages of Baidu as the seed pages. The crawling speed is about 10,000 pages per hour. The crawler crawled 3070 pages. The result is as shown in the following Figure 5 and Figure 6.

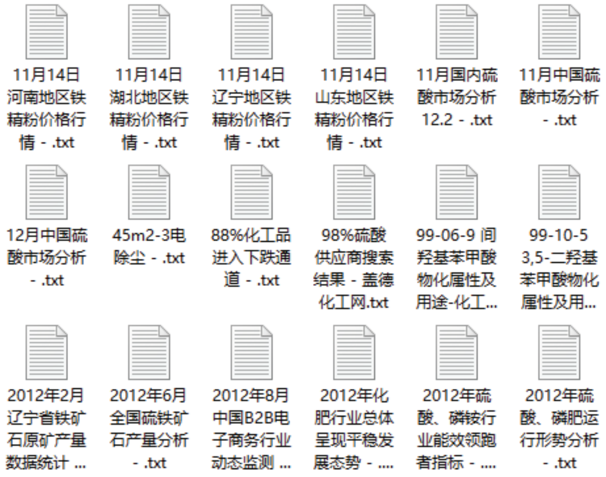
11月14日河南地区铁精粉价格行情 - .txt
11月14日湖北地区铁精粉价格行情 - .txt
11月14日辽宁地区铁精粉价格行情 - .txt
11月14日山东地区铁精粉价格行情 - .txt
11月国内硫酸市场分析12.2 - .txt
11月中国硫酸市场分析 - .txt
12月中国硫酸市场分析 - .txt
45m2-3电除尘 - .txt
88%化工品进入下跌通道 - .txt
98%硫酸供应商搜索结果 - 盖德化工网.txt
99-06-9 间羟基苯甲酸物化属性及用途-化工...
99-10-5 3,5-二羟基苯甲酸物化属性及用...
2012年2月辽宁省铁矿石原矿产量数据统计 ...
2012年6月全国硫铁矿石产量分析 - .txt
2012年8月中国B2B电子商务行业动态监测 ...
2012年化肥行业总体呈现平稳发展态势 - ....
2012年硫酸、磷铵行业能效领跑者指标 - ....
2012年硫酸、磷肥运行形势分析 - .txt

FIGURE 5. Pages crawled by the famous crawler

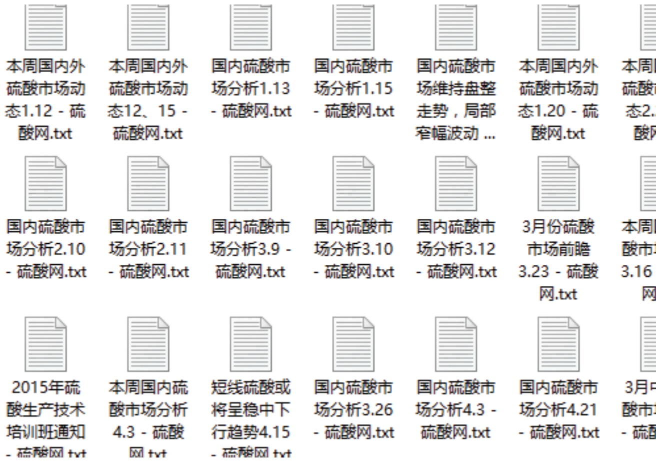
本周国内外硫酸市场动态1.12 - 硫酸网.txt
本周国内外硫酸市场动态12、15 - 硫酸网.txt
国内硫酸市场分析1.13 - 硫酸网.txt
国内硫酸市场分析1.15 - 硫酸网.txt
国内硫酸市场维持盘整走势，局部窄幅波动 ...
本周国内外硫酸市场动态1.20 - 硫酸网.txt
国内硫酸市场分析2.10 - 硫酸网.txt
国内硫酸市场分析2.11 - 硫酸网.txt
国内硫酸市场分析3.9 - 硫酸网.txt
国内硫酸市场分析3.10 - 硫酸网.txt
国内硫酸市场分析3.12 - 硫酸网.txt
3月份硫酸市场前瞻3.23 - 硫酸网.txt
2015年硫酸生产技术培训班通知 - 硫酸网.txt
本周国内硫酸市场分析4.3 - 硫酸网.txt
短线硫酸或将呈稳中下行趋势4.15 - 硫酸网.txt
国内硫酸市场分析3.26 - 硫酸网.txt
国内硫酸市场分析4.3 - 硫酸网.txt
国内硫酸市场分析4.21 - 硫酸网.txt

FIGURE 6. Pages crawled by the semantic crawler

We choose the first one hundred pages, judge the relevance manually, tot up the number of relevant pages for every ten pages and plot the graph of the accuracy rate. It is worth mentioning that the relevance does not rely on whether the keyword appears, but is based on the content of the document which can describe the topic. The result is as shown in Figure 7.

We can find that, for the first one hundred pages, the accuracy rate of our semantic crawler is slightly better than the famous focused crawler, but the disparity is not huge. To compare two crawlers at a large scale, we choose the one thousand pages and compute the accuracy rate for every 50 pages. The result is as shown in Figure 8.

It is clear that, from about 150 pages, the accuracy rate of the famous focused crawler has a significant drop and it keeps falling to about 0.4. The most likely reason may be the appearance of a big amount of advertising information. But because of the match

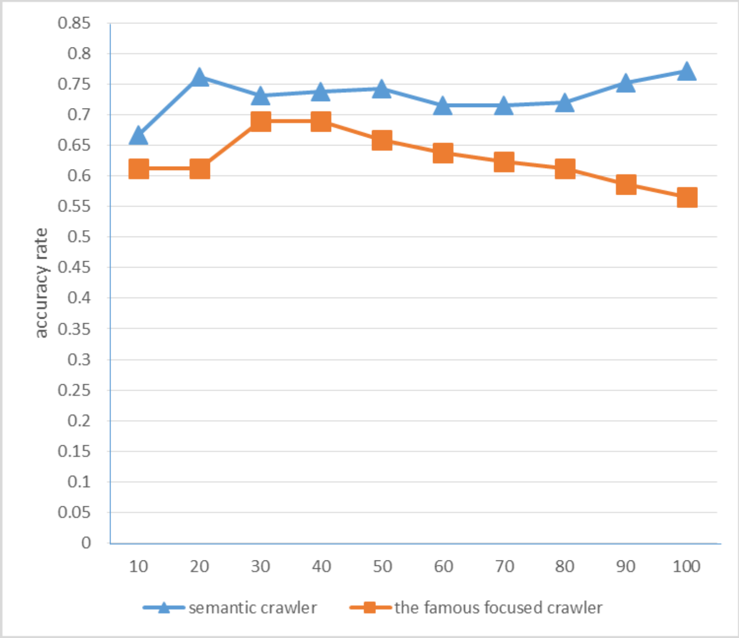

Figure 7. Accuracy rate of the first one hundred pages

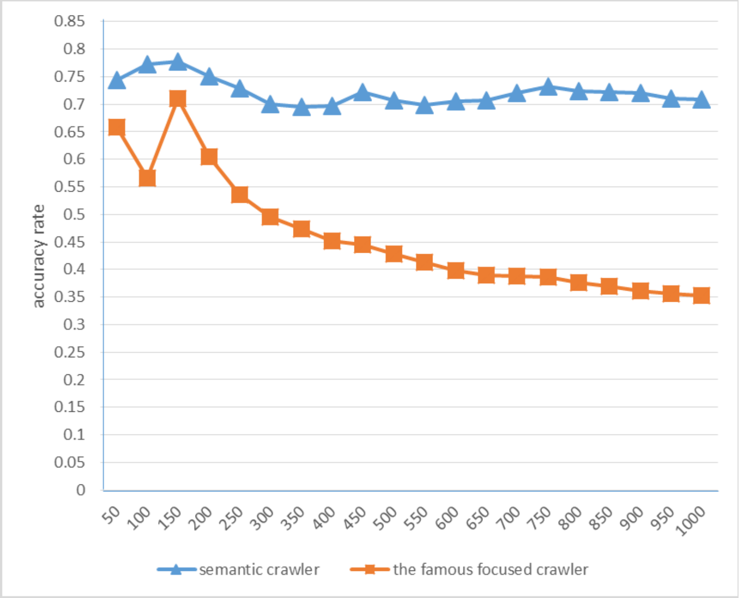

Figure 8. Accuracy rate of the first one thousand pages

mechanism based on semantic meaning, the accuracy rate of our semantic crawler still remains around 0.7 which is far more stable than the famous focused crawler.

7. **Conclusion.**

This paper proposed a method of constructing integrated semantic crawler to make up for the deficiency in common construction ways, explained the basic idea and process and described some key issues to be solved. The pre-work of our team is also introduced and the effect of our system has been validated by comparison experiment.

The future work includes exploring more accurate method to compute the relevance and a better way to extract the relationships between the concepts, and finding out the experience value of how many times should the semantic structure evolution iteration be by experiment or introducing the artificial judgment mechanism into the system.

Acknowledgment. This work is partially supported by National Key Project of Scientific and Technical Supporting Programs No. 2013BAH21B02. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] A. Rungsawang and N. Angkawattanawit, Learnable topic-specific web crawler, *Journal of Network and Computer Applications*, vol.28, no.2, pp.97-114, 2005.

[2] Z. Xie, A new subject-based Web crawler with concept tree, *Computer and Modernization*, no.4, pp.103-106, 2010.

[3] Q. Liu, A study of competitive intelligence acquisition system based on LDA and domain ontology, *Information Science*, no.4, pp.51-55, 2013.

[4] S. Batsakis, E. G. M. Petrakis and E. Milios, Improving the performance of focused Web crawlers, *Data & Knowledge Engineering*, vol.68, no.10, pp.1001-1013, 2009.

[5] M. Ehrig and A. Maedche, Ontology-focused crawling of Web documents, *Proc. of the 2003 ACM Symposium on Applied Computing*, Melbourne, FL, pp.1174-1178, 2003.

[6] Y. Du, Q. Pen and Z. Gao, A topic-specific crawling strategy based on semantics similarity, *Data & Knowledge Engineering*, vol.88, no.6, pp.75-93, 2013.

[7] B. Adelberg, NoDoSE – A tool for semi-automatically extracting semi-structured data from text documents, *Proc. of ACM SIGMOD*, Seattle, Washington, pp.283-294, 1998.

[8] A. H. F. Laender, B. Ribeiro-Neto and A. S. D. Silva, DEByE – Data extraction by example, *Data & Knowledge Engineering*, vol.40, no.1, pp.121-154, 2002.

[9] Z. An and J. Xu, The research on vision-based Web page information extraction algorithm, *Microcomputer & Its Applications*, vol.29, no.3, pp.38-41, 2010.

[10] C. Wang and J. Xu, Approach based on CURE algorithm of Web page segmentation and information extraction, *Microcomputer & Its Applications*, vol.31, no.12, pp.11-14, 2012.

[11] Q. Yang and M. Yang, A method of webpage content extraction based on point density, *Intelligent Computer and Applications*, vol.5, no.4, pp.42-44, 2015.

[12] Y. Liu, X. Chen and Z. Sui, Intelligent information systems and data mining study on evolution of domain ontology, *The 2nd International Conference on Innovative Computing, Information and Control*, Kumamoto, Japan, 2007.

[13] Y. Liu, Z. Sui, Y. Hu and Q. Zhao, Research on automatic construction of medical ontology, *International Conference on Biomedical Engineering and Computer Science*, Wuhan, China, pp.1-4, 2010.