http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (17): 4033-4039, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.4033.4039

Research and Realization of E-commerce Monitor System Based on Focused Web Crawler

XueGang Chen
Department of Computer science, Xiangnan University, Chenzhou, China

Abstract: With the development of E-commerce, the E-commerce monitor system is very important and traditional web crawler can't support real-time monitoring for E-commerce web. To solve the problems, this study proposes the theme crawler and monitoring technology and research and realization of E-commerce monitor system. User can customize his own tasks, including search the website and search the theme. The system may monitor a website according to monitoring period that the user defined and the results are returned to the user immediately. And it uses ontology to describe the topic representation model and it can accurately describe the user requirements. This study presents a new evaluation model about the topic correlation based on information quality. The attribute keywords of the represented topic webpage are viewed as a sequence in the model and it indicates the difference between the theme webpage using information quality. The simulation results show that the method can not only analyze similarity of the web page for crawling fast and effectively, but also improve the efficiency of the theme crawler for crawling and it greatly improves the monitoring efficiency using the E-commerce monitor system.

Key words: Focused web crawler, e-commerce, monitor system, information quality, ontology

INTRODUCTION

With the popularity of the Internet, people have entered the E-commerce era. According to the ministry of commerce, they released the "Business Logistics Development Plan" and expected that the E-commerce transaction volume will maintain a steady growth rate of 20% annual in the next five years and it will up to 12 trillion in 2015. Facing to the huge market of E-commerce, On the one hand, some unscrupulous businesses or enterprise eager to conduct E-commerce through the network, on the other hand, they make some counterfeit products and fraud in order to reap greater benefits. At present, the government manage E-commerce market through traditional market supervision mechanism, obviously it doesn't work, to strengthen supervision the market, we use web crawler technology (Chakrabarti *et al*., 1999; De Bra *et al*., 1994). To research and develop a E-commerce monitoring system, web crawler is the first part of its application and it is the key technology and E-commerce WebPages are collected to the local by it, also, the illegal downloaded information is built index to extract information and mining text, in order to monitor illegal management behavior. The monitoring system can effectively crack down on the E-commerce crimes.

However, E-commerce brings many challenges to web crawler because of a large of information. If lots of information is collected, it will require a great deal of computation, storage and many network bandwidth resources, at the same time, the E-commerce information is dynamic and related topics information is not complete, the utilization rate of the download page is also very low. Pointed, most applications don't need to fetch all pages, but the illegal information management behavior in E-commerce is only required for crawling, so it is very necessary that E-commerce monitor system based on focused crawler is researched.

FOCUSED CRAWLING AND RELATED WORK

Focused web crawler: A web crawler can automatically extract webpage through URL and traverse the web along the hyperlinks in the fetched page. A general web crawler starts to crawl from an original web or URL seed in some initial webpage, then parses the html file and extracts the sub links in the downloaded page, thus obtains a list of URL in the initial web page. The crawler extracts the new URL from the current page and places in the queue, until meet the system stop condition in the process of fetching web page.

But a focused crawler can filter irrelevant links via., a certain webpage analysis algorithm and keeps the related links and puts to the grabbed URL queue, then continues to fetch webpage from URL the queue according to a certain search strategy and repeats the process until meet the system stop condition. The grabbed webpage will be

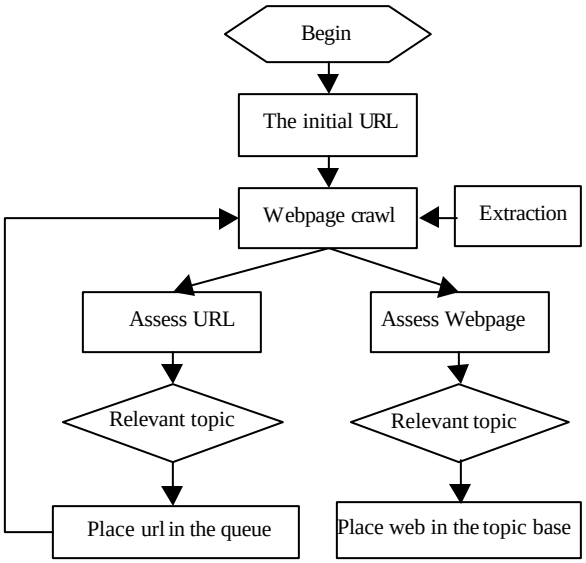

Fig. 1: Focused crawler

stored in the system to analyze, filter and index and the analysis results are also likely to feedback in order to direct to subsequent capture process. As shown in Fig. 1.

Related work: A focused crawler analyzes its crawl boundary to find some links that are likely to be most relevant for the crawl and avoids irrelevant regions of the Web. This leads to significant savings in hardware and network resources and helps keep the crawl more up-to-date. So, researchers have studied the focused crawling in the past and they proposed a number of algorithms. The first generation of crawlers on which most of the web search engines are based rely heavily on traditional graph algorithms, such as breadth-first or depth-first traversal to index the web. A core set of URLs are used as a seed set and the algorithm recursively follows hyper links down to other documents. Document content is paid little heed, since the ultimate goal of the crawl is to cover the whole web. Chakrabarti *et al.* (1999) propose the fish Search algorithm which treats every URL as fish whose survivability depends on visited page relevance and server speed (Su *et al.*, 2005) the Shark-Search algorithm improves Fish-Search as it uses vector space model in order to calculate the similarity between visited page and query (Ye *et al.*, 2004). Yang *et al.* (2010) improve the Shark-Search theme search algorithm in the search width,and link similar judgment and crawling link selecting the strategies and takes "first search, after the judgment" of the search process. (Rungsawang and Angkawattanawit, 2005) improve best-first focused crawling strategy combined with different heuristics. A generic architecture of focused crawling is proposed by Chakrabarti *et al.* (1999). This architecture includes a classifier that evaluates the relevance of a web page with respect to the focused topics and a distiller that identifies web nodes with high access points to many relevant pages within links. (Higham, 2005) Page rank is a proportion of back-link value and forward-link value of a web page. By experiments, they have found that page rank is the best parameter to be used in URL ordering process, as it means how well-known a URL is to another URL. Ma and Yuan (2011) A credibility evaluation method for web information based on improved page rank was proposed. This method not only considered the interactive structure between webs, but also took into account semantic relation between the web information and the time decay function was introduced in computing.

A focused crawler and web crawler are very important in the search engine and it is applied to other application areas. Pan and Xu (2010) to fully exploit the information contained in a blog, the idea of Chinese Blog search engine (CBSS) is proposed and the architecture of CBSS is designed. The traditional web crawler is improved by using rules definition, regular expression. (Zhao and Nie, 2011) a content analysis based Chinese BBS topic detection algorithm, including obtaining BBS information by web crawlers, processing BBS information with URL and Xpath based webpage templates, realizing BBS information participles by ICTLAS, clustering BBS topics by Carrot, analyzing hot topics based on the power spectrum and predicting topics based on time sequences. A webpage detection system based on honey client is designed, in which spider is combined with honey pot. In the system, spider is used to collect source of urls, then client engine automatically created Internet Explorer processes and device-driven detector is used to detect mal wares coming through Interne t Explorer. In the end, the malicious web page's url is added to the black list and the mal ware database is enlarged (Sun *et al.*, 2011). Dong *et al.* (2012) proposes an ontology-learning-based focused crawling approach, enabling Web-crawler-based online service advertising information discovery and classification in the web environment, by taking into account the characteristics of service advertising information.

E-COMMERCE MONITORING MODEL BASED ON THE TOPIC CRAWLER

E-commerce monitoring technology based on topic crawler platform monitors webpage which uses the strategy and special computer program and the crawling

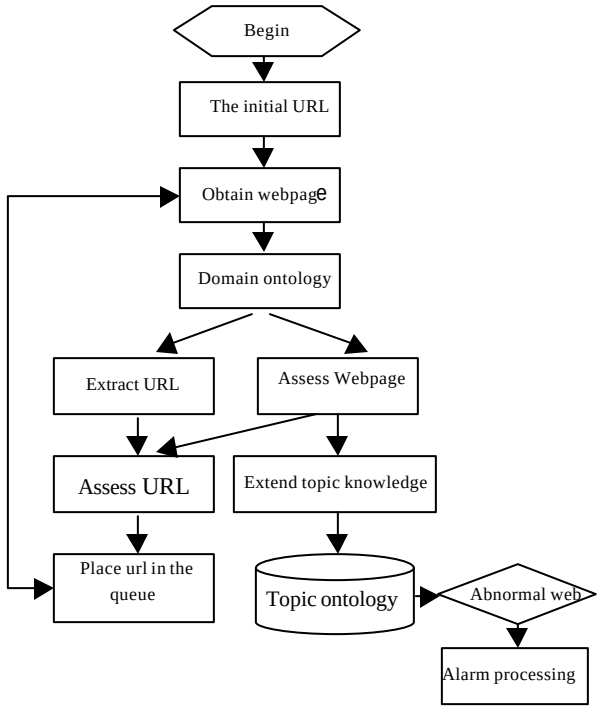

Fig. 2: System architecture

information is constructed using ontology, then assess relevant webpage and the extracted url and continues to expand subject knowledge, finally form a topic ontology base and to retrieve monitoring information for the user and monitoring information that meet the requirements is displayed to the user. User can customize his own task,including search websites and search theme.The system monitors according to monitoring period that the user defined and the results are returned to the user immediately. The system can realize the real-time monitoring according to the user's requirements and monitoring cycle, the system architecture is shown in Fig. 2.

KEY TECHNOLOGY OF E-COMMERCE MONITOR SYSTEM

Representation model of the system: How do users describe monitoring subject before focused crawler starts to work. A method of description subject is called thematic representation model and the topic using theme monitoring representation model is called subject knowledge. A topic granularity is the larger in focused crawler. Ontology is a conceptual model and it can make concept semantic via describing the concepts and the relations between the concepts. Therefore, domain ontology is more suitable to describe representation model of theme monitoring information. Traditional focused crawler usually uses representation model based on the topic in text mining that is a feature vector model of the relevant topic which discovers new relevant webpage through calculating text similarity and a crawler begins to fetch on-line relevant webpage, when a topic document is very discrete and less relevant url, traditional focused crawler may lead to premature convergence and miss a lot of useful information. In addition, a user is often difficult to accurately describe a topic, Once a given sample document is not reasonable and the characteristics are continuously extend and a result is the subject produces serious theme offset. In fact, it is large that the user wants to fetch the topic granularity and the vocabulary on a web page is a class concept, also, they may be regard as the keywords of the concepts.

However, existing model based on domain ontology is difficult to direct to the focused crawler. Firstly, most of the existing ontology database hasn't uniform format and very complicated. Secondly, many areas don't construct a complete ontology base, at the same time, the user could not pre-construct complete domain ontology data base before crawling. Therefore, we define a simplified ontology model in this study which build the theme model of information representation according to fetching relevant monitoring information for a user. It fit to build ontology according to the demand and without the full range of web information. It is convenient for a user to describe a theme and it fit to extend subject knowledge.

Definition1. Suppose that O = (R, U, C) is a topic knowledge, R is the keywords Set of the themes directly related, U is the keywords Set of entirely unrelated, C is the keywords Set of indirect correlation.

If a webpage belong to a field in E-commerce which often contains some key words in the field. For example, the related field document about "Etam cotton-padded clothes" and it must contain the keywords "Etam, cotton, clothes" and R is referred to the keywords set representing domain feature. In order to limit for the crawl range, it must contain at least one keyword in R if the document is topic relevant.

U is a set of filtering irrelevant with the theme of the document and it must not contain any keyword in E if a document is topic related. For example, the topic knowledge about "Etam cotton-padded clothes", U should include "duck down and Pants", so it can filter the content.

C is a set of extension theme related keywords and it doesn't directly contain the keywords in A, but it associates with set A, it is used to extend subject knowledge according to the correlation.

From the above analysis, for the topic representation model, users only need to set a number of keywords in the initial and describe easily the topic.

Here, is the definition of ontology about commodity in E-commerce:

```
<owl: Class rdf: ID = "clothes" >
< owl: onProperty rdf: resource = "# Brand " >
< owl: comment > Etam < /owl: comment >
< owl: onProperty rdf: resource = "# Fabric " >
<owl:comment> Polyester fiber < /owl: comment>
< owl: onProperty rdf: resource = "# Color " >
< owl: comment > Deep orange < /owl: comment>
< owl: onProperty rdf: resource = "# Thickness " >
<owl: comment > Conventional < /owl: comment>
<owl: onProperty rdf: resource = "# Evaluation " >
< owl: comment > Very good < /owl: comment>
...
```

Correlation evaluation based on information quantity: A web document contains the theme interrelated information from the Internet which including image, animation, text, audio, hyperlinks and other general information in a webpage, but most important information is still text. First from the webpage, filtering the html label, advertisement, navigation and extracts the text scanning the obtained web document and filters the unrelated information and the word "noise". The expression domain feature words are usually some noun, therefore, only extract nouns as feature words when we analyses a webpage text or URL anchor text.

In order to reduce irrelevant content, the keywords set must contain at least one subject knowledge belong to R set and no belong to U when these content is related theme webpage according to definition 1. The correlation value is 0 if don't meet above conditions of the webpage, When meet the conditions of the webpage, we introduce information quality. The attribute keywords of the represented topic Webpage is viewed as a sequence in the model and it indicates the difference between the theme web page Using information quality, thus these web pages sequences were compared and classified.

Information quality of E-commerce page sequence: Information entropy is used to measure the uncertainty of a system (Chen *et al*., 2013). We improve it based on the others' research in this study and define information quality by information theory, in order to distinguish Shannon information entropy, we call the uncertainty measure for information quality.

Concept, information quality related formulas and their properties are introduced. The attribute keywords of the represented topic Webpage are viewed as a sequence in the model and it indicates the difference between the theme web page Using information quality, thus the similarity of web page sequences are assessed.

Definition 2: A web page W = {T, C, L}, it can describe overall information page, T is webpage topic, C is webpage text, C = $\{c_1, c_2,\ldots, c_n\}$, c_i is the i effective keyword of webpage text information, L is anchor text of link (link text), Here the context of link also contain in L.

Definition 3: A web page P = $\{f_1, f_2,\ldots, f_k\}$, P is a information sequence, $f_i \epsilon K$, k is the length of a sequence, K = $\{w_1, w_2,\ldots, w_q\}$, q is q keywords. For example, a web page T = {C, D}, C = {A, C, D, C, B, A, B, B, D, C}, L = {A, B, D}, the sequence of the page P = {C, D, A, C, D, C, B, A, B, B, D, C, A, B, D}, it can be thought as a function F: F(1) = C, F(2) = D,..., F(15) = D, A, B, C, DϵK = {A, B, C, D}.

Suppose that P = $\{f_1, f_2,\ldots, f_k\}$ is an isometric webpage sequence using the attribute value, it can be defined order set N = {1, 2,..., n} and an equivalence relation Ond(P):

$$Ind(P) = \{i, j \epsilon N\times N | \forall k \epsilon \{1, 2,\ldots, K\}, f_k(i) = f_k(j)\},$$

Obviously:

$$Ind\ (P) = \bigcap_{k=1}^{k} Ind(\{f_k\}), f_k = f_1,$$

If and only if $Ind(\{f_k\}) = Ind(f_1)$.

Unequal Webpage sequence P = $\{ f_1, f_2,\ldots, f_k\}$, the maximum length sequences for n

Each sequence is filled using the null symbol,and make $_p$ isometric web page sequence using the attribute value, it can be defined order set N = {1, 2,..., n} and a compatibility relation Sim(P):

$$Sim(P) = \{i, j \epsilon N\times N | \forall k, fk(i) = fk(j)\}, \text{ or } fk(i) = * \text{ pr } fk(j) = *\}$$

Obviously:

$$Sim\ (P) = \bigcap_{k=1}^{k} Sim(\{f_k\}),$$

$S_p(i)$ is compatible class, Therefore, all consistent set $S = \{S_p(i) | i \epsilon N\}$ form A cover N without a partition N.

Definition 4: Suppose that A = $\{f_1, f_2,\ldots, f_m\}$ is isometric web page sequence database, whose length is n, p$\subseteq$A, U/Ind(P) = $\{X_1, X_2,\ldots, X_t\}$, we define information quality of the web page sequence set Pas follows:

$$E(P) = \sum_{i=1}^{t} \frac{|X_i|}{n}\frac{|X_i^c|}{n} = \sum_{i=1}^{t} \frac{|X_i|}{n}\left(1 - \frac{|X_i|}{n}\right) \quad (1)$$

X_i^c is X_i's complement and X_i^c = N- X_i, $|X_i|$ is set X_i's base, Supposing that Q$\subseteq$A and U/Ind(Q) = $\{Y_1, Y_2,\ldots, Y_k\}$.

We define condition information quality of the web page sequence set P, Q as follows:

$$E(Q/P) = \sum_{i=1}^{1}\sum_{j=1}^{k}\frac{|X_i \cap Y_j|}{n}\frac{|X_i - Y_j|}{n} \quad (2)$$

We define interaction information quality of the web page sequence set P, Q as follows:

$$E(Q,P) = \sum_{i=1}^{1}\sum_{j=1}^{k}\frac{|X_i \cap Y_j|}{n}\left(1-\frac{|X_i - Y_j|}{n}\right) \quad (3)$$

We define joint information quality of the web page sequence set P, Q as follows:

$$E(P \cup Q) = \sum_{i=1}^{1}\sum_{j=1}^{k}\frac{|X_i \cap Y_j|}{n}\left(1-\frac{|X_i \cap Y_j|}{n}\right) \quad (4)$$

Definition 5: Suppose that $A = \{f_1, f_2,\ldots, f_m\}$ is a Unequal Webpage sequence database, its the maximum length is n, for $P \subseteq A$, $S(P) = \{S_P(1), S_P(2),\ldots, S_P(n)\}$, We define information quality of the web page sequence set P as follows:

$$E(P) = \sum_{i=1}^{n}\frac{1}{n}\left(-\frac{|S_p(i)|}{n}\right) \quad (5)$$

Also, for $Q \subseteq A$, $S(Q) = \{S_Q(1), S_Q(2),\ldots, S_Q(n)\}$, we define condition information quality of the web page sequence set P, Q as follows:

$$E(Q/P) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{n}\left(-\frac{|S_p(i) - S_Q(j)|}{n}\right) \quad (6)$$

We define interaction information quality of the web page sequence set P, Q as follows:

$$E(Q, P) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{n}\left(-\frac{|S_p(i) \cup S_Q(j)|}{n^2}\right) \quad (7)$$

We define joint information quality of the web page sequence set P, Q as follows:

$$E(P \cup Q) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{n}\left(-\frac{|S_p(i) \cap S_Q(j)|}{n^2}\right) \quad (8)$$

Construct the correlation degree model:

Definition 6: Suppose that $A = \{f_1, f_2,\ldots, f_m\}$ is a webpage sequence database, for P, $Q \subseteq A$, we define similarity coefficient of the web page sequence set P, Q as follows:

$$\rho(P,Q) = \frac{E(P, Q)}{\sqrt{E(P)}\sqrt{E(Q)}} \quad (9)$$

If E(P) = 0 or E(Q) = 0, then r(P, Q) = 1.

EXPERIMENTS AND THE ANALYSIS

According to the above ideas, the E-commerce website is tested. In this study, we select a goods webpage as "Taobao" and it can be divided into three different test data. These webpage is parsed and tagged by data analyzer according to the topic and we fetch four words of describing the same theme webpage after data pretreatment.

In order to facilitate the statistics and these keywords can represent the characteristics of the theme webpage which is replaced by A, B, C, D, respectively. The theme webpage sequences are constructed according to the method and P0 is the original information of the theme webpage. Select the 100 relevant webpage and the high correlation theme webpage were selected 7 items using this algorithm and they are sorted by random and nature.

Form the Fig. 3, the experiment results show that the value of correlation threshold of τ is lower, the number of the webpage is more when crawling stop, but the correlation page is higher. Conversely the value of correlation threshold of τ is higher, the number of the fetched webpage is less, but the relevance page is more. From Fig. 1 know, the system can recommend higher correlation webpage when τ is greater than 0.7 and the webpage of weak correlation or no correlation will be filtered out. At the same time, for the rationality of detection algorithm, the experimental personnel review webpage content of the theme page P0 and they make a judgment according to the high correlation, correlation, low related and unrelated four grades by taking P0 as the standard, lastly the first 7 pages for fetching are evaluated and their results is high correlation.

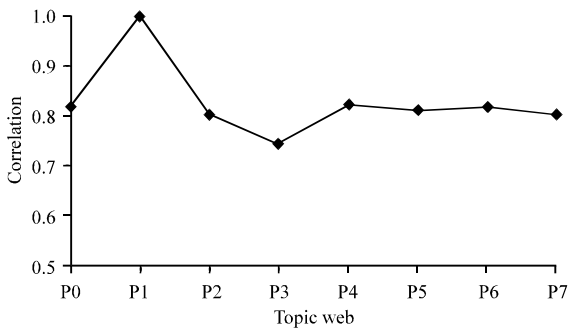

Fig. 3: Correlation of the topic page

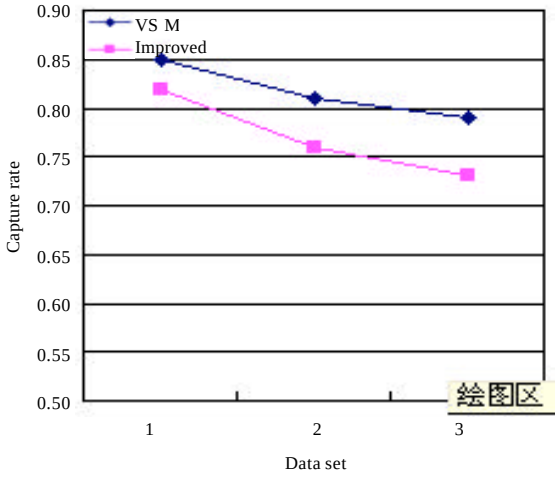

Fig. 4: Capture rate of webpage

In order to further test the correctness and effectiveness of the system, Vector Space Model as a contrast, for fair comparison, the experimental data set and the way of sampling are the same. The crawler threads of parallel crawling is forty, the depth of grasping is five and the system monitoring cycle is 24 h. The three different test data show that the system using the traditional VSM algorithm grasp lots of webpage, but the correlation webpage is not much. Compared with the system using method in this study, the fetched webpage quantity of the latter reduced by about 3.5-7.6%. As shown in Fig. 4. For reducing the irrelevant webpage, it is improved greatly the speed and quality. Webpage with the theme of "corresponding to the purchase order" is highly related, irrelevant and low correlation webpage is filtered, so it is a higher search accuracy and strong real-time monitoring.

CONCLUSION

In order to monitor E-commerce effectively, focused crawler technology is applied. The technology is studied deeply and E-commerce monitor system based on focused web crawler is proposed which is designed many the new algorithm with unique characteristics based on the original technique. To the user better express and describe the topic information, representation model of theme monitoring information based on ontology is proposed which analyses the topic correlation from semantic concept level and key words are mapped to semantic concept level. To analyze theme correlation effectively, correlation evaluation of the theme monitoring information based on information quantity is provided. The simulation results show that the method can not only analyze similarity of the web page for crawling fast and effectively, but also improve the efficiency of the theme crawler for crawling.

ACKNOWLEDGMENT

The authors would like to thank for financial support by science and technology plan project of Hunan Province (2013FJ3032),Education Department scientific research projects of Hunan Province (11C1184),Xiangnan University Research Fund (45).

REFERENCES

Rungsawang, A. and N. Angkawattanawit, 2005. Learnable topic-specific web crawler. J. Network Comput. Appl., 28: 97-114.

Chakrabarti, S., M. van der Berg and B. Dom, 1999. Focused crawling: A new approach to topic-specific Web resource discovery. Comput. Networks, 31: 1623-1640.

Chen, X.G., J.L. Zhang and J.R. Cheng, 2013. Analysis of similarity of DNA sequences based on information quantity. Appl. Res. Comput., 30: 1381-1384.

De Bra, P., G.J. Houben, Y. Kornatzky and R. Post, 1994. Information retrieval in distributed hypertexts. Proceeding of the 4th RIAO International Conference, October 11-13, 1994, USA., pp: 481-491.

Higham, D.J., 2005. Google PageRank as mean playing time for pinball on the reverse web. Applied Math. Lett., 18: 1359-1362.

Dong, H., F.K. Hussain and E. Chang, 2012. Ontology-Learning-Based focused crawling for online service advertising information discovery and classification. Proceedings of the 10th International Conference on Service-Oriented Computing, November 12-15, 2012, Shanghai, China, pp 591-598.

Ye, Y.M., F.Y. Ma, Y.M. Lu, M. Chiu and J.H. Huang, 2004. iSurfer: A focused web crawler based on incremental learning from positive samples. Proceedings of the 6th Asia-Pacific Web Conference on Advanced Web Technologies and Applications, April 14-17, 2004, Hangzhou, China, pp: 122-134.

Yang, R.G., Y. Song and X.Z. Meng, 2010. Multimedia topic search algorithm based on improved shark-search. Comput. Eng. Appl., 46: 152-154.

Ma,W.Y. and F. Yuan, 2011. A credibility evaluation method for web information based on improved pagerank. J. Zhengzhou Univ. (Nat. Sci. Edn.), 43: 43-47.

Pan, B. and L.L. Xu, 2010. Study of Chinese blog search engine. Comput. Eng. Des., 31: 1718-1721.

Su, C., Y. Gao, J. Yang and B. Luo, 2005. An efficient adaptive focused crawler based on ontology learning. Proceedings of the 5th International Conference on Hybrid Intelligent Systems, November 6-9, 2005, China, pp: 6-9.

Sun, X.Y., Y. Wang, Y.F. Zhu and D.Y. Wu, 2011. Design and implementation of mal-webpage detection system based on honeyclient. Comput. Appl., 28: 242-246.

Zhao,Y.H. and Z. Nie, 2011. Designand imp ementation of chinese bbs. Comput. Appl. Software, 28: 242-246.