Semantic Focused Crawling for Retrieving E-Commerce Information

Wei Huang[12]
[1]School of Information Management, Wuhan University, Wuhan , P.R. China
[2]School of Management, Hubei University of Technology, Wuhan , P.R. China
Email: tonny_hw@163.com

Liyi Zhang[1], Jidong Zhang[2] and Mingzhu Zhu[1]
[1]School of Information Management, Wuhan University, Wuhan , P.R. China
[2]School of Management, Hubei University of Technology, Wuhan , P.R. China
Email: lyzhang@whu.edu.cn, {datamine , zhumzhu}@163.com

***Abstract*—Focused crawling is proposed to selectively seek out pages that are relevant to a predefined set of topics without downloading all pages of the Web. With the rapid growth of the E-commerce, how to discovery the specific information such as about buyer, seller and products etc. adapting for the online business user becomes a focused issue to the information search engine. We present a novel semantic approach for building an intelligent focused crawler which deals with evaluating the page's content relevance to the E-commerce topic by the domain ontology and the hyperlinks connection to the commercial web pages by link analysis. In the process of crawling, the domain ontology can evolve automatically by machine learning based on the statistics and rules. Experiments have been performed, and the results show that our approach is more effective than the other traditional crawling algorithms, and prevents the topic-drift with higher harvest rate.**

***Index Terms*— Focused crawling, Information retrieval, E-commerce, Semantic, Machine learning**

I. Introduction

The advent of the Web has created a new medium for publishing and disseminating various sorts of information. With its exponential growth, the Web has now become the biggest human knowledge base, including commerce information. In the early stage of the Internet, it has been used for global enterprises to advertise their business areas and products to potential customers in the world. However, as the business value of the Internet has been highly evaluated, it is now being mainly used as a business interaction channel of e-commerce and supply chain management for making profits. E-commerce and e-procurement are the ways of undertaking business transactions between buyers and suppliers through the Internet beyond the limits of time and location [1],[2]. So the commerce information on the web is very important to the all participators in e-commerce, retrieving e-commerce information accurately and effectively becomes the main issue of the web search engine.

Within the last couple of years, search engine technology had to scale up dramatically in order to keep up with the growing amount of information available on the web. Web search engines are the entry point to the web for millions of people. In order for a search engine to be successful towards offering web users with the piece of information they are looking for, it is vital that the engine's index is as full and qualitative as possible. Considering the billions of pages that are available on the web and the evolution of the web population, it becomes evident that building a search engine to fit all purposes might be a tedious task. To alleviate both search engine users and search engine developers from the burden of going over large volumes of data, researchers have proposed the implementation of vertical search engines; each containing specialized indices and thus serving particular user needs. A search engine with a specialized index has consequently a more structured content and offers higher precision than a generalized search engine, as it has already been intelligently extracted from the web [3],[4]. To the e-commerce users, that is commerce vertical search engine with commercial index.

So that first step, it is very important to develop web pages discovery mechanisms based on intelligent techniques – focused crawling [5]. The goal of a focused crawler is to selectively seek out pages that are relevant to a pre-defined set of topics. The topics are specified not using keywords, but using domain knowledge documents, such as the ontology [6]. Rather than collecting and indexing all accessible Web documents to be able to answer all possible ad-hoc queries, a focused crawler analyzes its crawl boundary to find the 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 [7]. That is only to business users to search for business information.

One of the main issues of focused crawling is how to effectively traverse off topic areas and maintain a high harvest rate during the crawling process. To address this issue, an effective strategy is to apply background knowledge of crawling topics to focused crawling. Since an ontology is defined as a well organized knowledge scheme that represents high-level background knowledge with concepts and relations, ontology-based focused crawling approaches have come into research [8]. Such approaches use not only keywords, but also the E-Commerce ontology to scale the relevance between web pages and crawling topics. In the ontology-based focused crawling approaches, because concept weights and structures are heuristically predefined before being applied to calculate the relevance scores of web pages, it is difficult to acquire the optimal concept weights and adaptive structures to maintain a stable harvest rate during the crawling process. To address this issue, a learnable ontology focused crawling approach by machine learning based on the statistics and rules is proposed in this paper. In our approach, the ontology self-evolving mechanism can dynamically adapt to the particular environment of the relevant topic.

The rest of this paper is structured as follows: Section 2 reviews the background theory and work that is related to focused crawling; in Section 3 we outline the proposed architecture and its components with our algorithms; Comparisons with existing focused crawling methods on some test tasks are shown in Section 4. Finally, Section 5 discusses the conclusion and future work.

II. Related work

Researchers have studied the focused crawling problem in the past and proposed a number of algorithms. In [9] the authors propose the Fish Search algorithm, which treats every URL as fish whose survivability depends on visited page relevance and server speed. In [10] the Shark-Search algorithm improves Fish-Search as it uses vector space model in order to calculate the similarity between visited page and query. [11] best-first focused crawling strategy combined with different heuristics. A generic architecture of focused crawling is proposed by S. Chakrabarti [12]. 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. Brin and Page [13] Pagerank is a proportion of back-link value and forward-link value of a web page. By experiments, they have found that pagerank is the best parameter to be used in URL ordering process, as it means how well-known a URL is to another URL. At the moment, pagerank is the main technique used by the Google search engine [14] proposed a focused crawling algorithm that constructs a model named context graph for the contexts in which topically relevant pages occur on the web. However, most of the above approaches do not consider applying background knowledge such as ontology to focused crawling. Since the ontology is a description of the concepts and relationships that can well represent the background knowledge, a good way to improve the harvest rate of focused crawling is to use ontology for crawling. Several focused crawling approaches based on ontology have been proposed [15],[16]. The most prevalent focused crawling approach based on ontology is called ontology-focused crawling approach [17],[18],[19]. depicted the ontology-focused crawling framework. In this approach, the crawler exploits the web’s hyperlink structure to retrieve new pages by traversing links from previously retrieved ones. As pages are fetched, their outward links can be added to a list of unvisited pages, which is referring to as the crawl frontier. [20] developed a rule-based crawler which uses simple rules derived from inter class linkage patterns to decide its next move. Moreover their rule-based crawler support tunneling and try to learn which off topic pages can eventually lead to high quality on-topic pages. [21],[22] proposes a machine learning oriented approach for focused crawling: the Q-learning algorithm is used to guide the crawler to pass through the off-topic document to highly relevant documents. [23],[24]proposed an ontology learning framework includes a number of complementary disciplines that feed on different types of unstructured, semi-structured and fully structured data in order to support a semi-automatic, cooperative ontology engineering process. In the topic domain, [25],[26],[27] applies a fuzzy ontology framework to information retrieval system in E-Commerce. Compared to these approaches discussed above, our approach not only provides a way to apply the learnable ontology for focused crawling based on statistics and rules, but also uses link analysis which forecasts topic relevance according to the link information. This idea has not been researched in detail until now.

III. Architecture

The web crawling for focused search engine is done by a focused crawler. In the process of crawl the web networks, the priority sequence of the pages is determined upon the crawl priority score of content relevance and link prediction. As the priority sequence, the focused crawler can fetch the topic web pages from the frontier of the sequence. It is ensure that we can get higher rate of the topic harvest and prevent topic-drift. Its architecture can be described by figure 1.

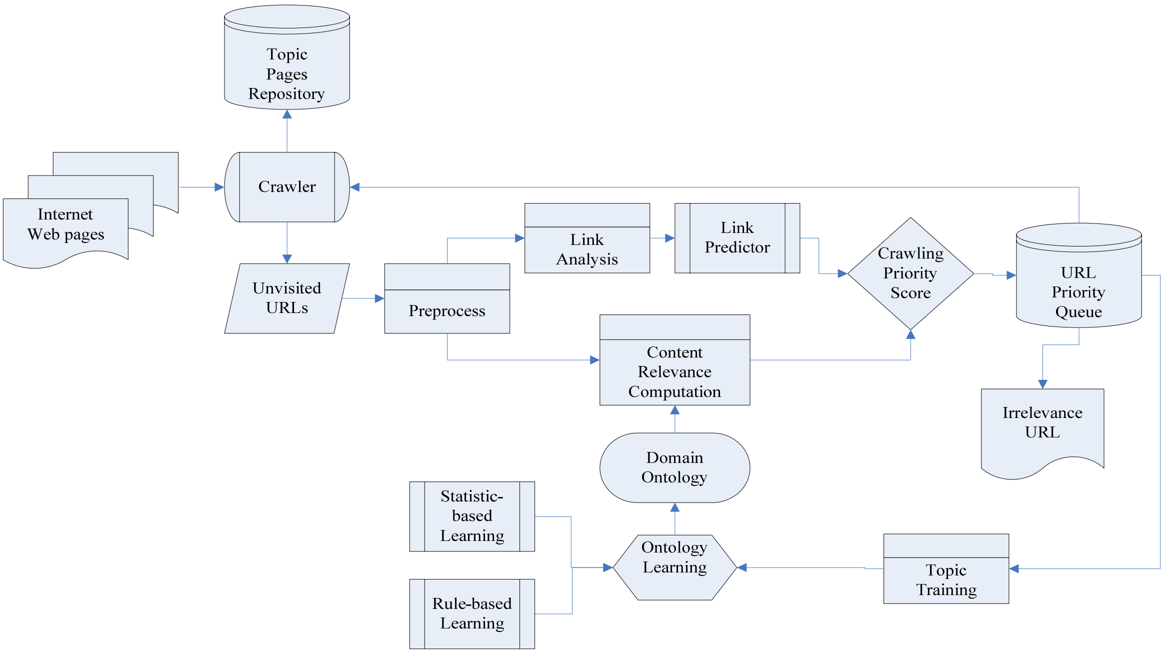

Figure 1. Architecture of focused crawler based on learnable ontology

Firstly, crawler begins to fetch the unvisited web pages to the preprocess module, which parses the link and content information. The preprocessing module includes functions of parsing web pages from html to text, making POS (Part-Of-Speech) tagging for web pages, stop words removal, and word stemming. HTML Parser is employed to process the html by removing all the html tags. After preprocessing of the visited web pages, the entities (words occurring in the set of relevant concepts) and the links' structures are extracted and counted. Then the link analysis module gets the all link information and formulates the link priority score by the predictor. The content entities will be get into the content relevance computation module and calculate the content relevance score by the entities mapping to the E-Commerce ontology. Ontology based algorithm is critical to the relevance computation. In order to make the ontology adaptive to the context of E-Commerce, an ontology learning process which is based on the machine learning algorithms such as statistic-based and rule-based must be done. The crawl priority module which will summarize the score of crawl priority based on the link and content priority score arranges to the sequence of crawling according to the crawl priority score. The larger crawl score pages will be visited more quickly. Finally, the more focused web pages will be downloaded into the topic repository which prepares for the focused information retrieval with higher recall ratio. We now describe the five core issues in detail.

A E-Commerce features

E-commerce is an extensive domain relatively, and how to ensure that there is no topic-drift when focus retrieving E-commerce information as much as possible, which requires a more optimized E-Commerce feature extraction and expression mechanism.

It has been widely accepted that ontology is the core element for the Semantic Web. Ontology is defined as a well organized knowledge scheme that represents high-level background knowledge with concepts and relations. It can well contain and express the features of the E-Commerce domain. This will have to be extended for the relevance measure of our focused crawler. An ontology is an explicit specification of some topic. For our purposes, it is a formal and declarative representation which includes the vocabulary (or names) for referring to the terms in that subject area and the logical statements that describe what the terms are, how they are related to each other, and how they can or can not be related to each other. Ontology therefore provides a vocabulary for representing and communicating knowledge about some topic and a set of relationships that hold among the terms in that vocabulary.

The on-line transactions is the main portion of the E-Commerce, so buyers, sellers and the product information is the core classes of E-Commerce ontology, and the on-line payment, off-line delivery ways and contracts for the sale are also essential. Above all, E-Commerce Ontology's structure can be described in detailed by figure 2:

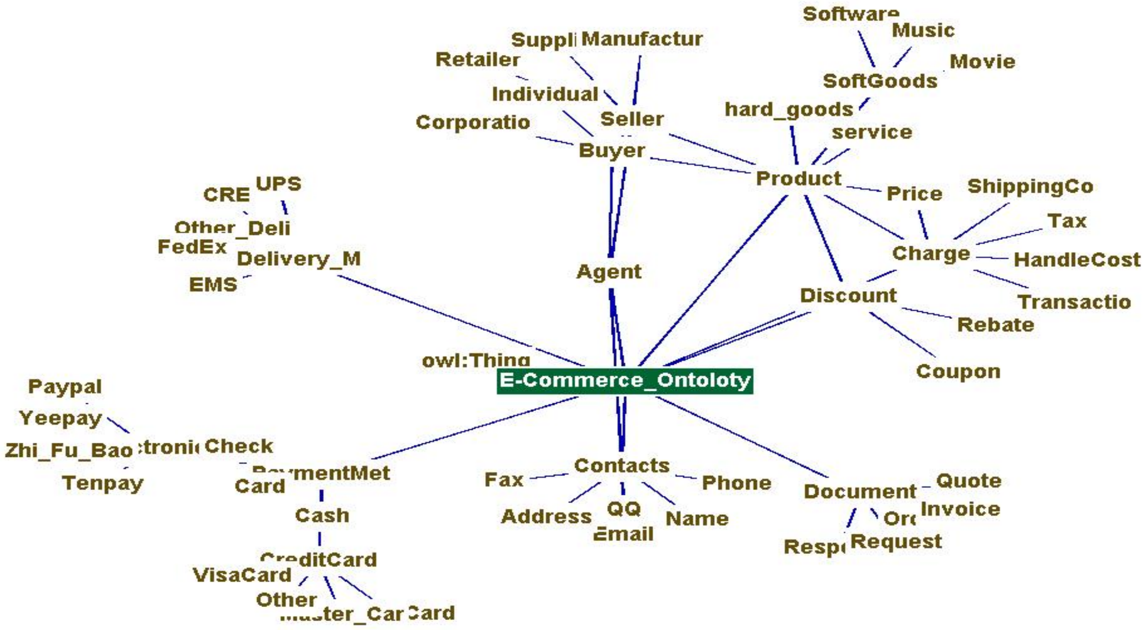

Figure 2. E-Commerce Ontology Structure

Definition 1 (Semantic Level) Ontology O: {C, R, Hc, Hr}, which includes: classes C; relation functions R: R∈C×C; a class hierarchy Hc : Hc∈C×C, Hc (Cm, Cn) means that Cm is a sub-class of Cn; and a relation hierarchy Hr : Hr∈R×R, Hr(Rm, Rn) means that Rm is a sub-relation of Rn.

Definition 2 (Lexical Level) Lexicon L: {Lc, Lr, Fc, Fr}, with lexical entries Lc, Lr for classes C, relations R and functions Fc: F(C) = {L | C×L∈F} and Fr: F(R) = {L | R×L∈F}.

B Content Relevance Computation Based on Ontology

This is an important component of the content relevance computation module, in general the relevance measure is a function which tries to map the content lexical entities against the existing ontology to gain an overall content relevance score: Spc. The measures and the associated relevance computation strategies are described in detail as following.

To give a detailed overview on the relevance computation in our framework, we introduce a formal model of our notion of ontology and metadata, where a specific focus is set on the interaction of ontology and associated metadata with natural language. As the architecture in Figure 1, hypertext has already been adjusted using the preprocessing module. This step is a lookup of entities of lexicon of the ontology. Every time a word matches an entry in the lexicon and therefore an entity of the ontology structure, the statistical information of responding entity is counted.

Definition 3 (Frequency of Class) Frequency of class Ck in document D:

$f_{C_k}^{D} = N_{\omega_1^k}^{D} + N_{\omega_2^k}^{D} ...... + N_{\omega_i^k}^{D}$, with a set of words according to class Ck : $\omega_1^k, \omega_2^k ......\omega_i^k$; and the times that word ω_i^k appears in document D: $N(\omega_i^k)^D$.

Definition 4 (Weight of Class) The weight of class Ck is the difference between the entities in the ontology, it is defined as Wck with n being the discount factor and d(Ck,T) being defined as the distance between an entity of the ontology and the topic entity T :

$$W_{C_k} = 100\% * n^{d(C_k,T)} \tag{1}$$

Definition 5 (Relevancy Function) Priority content relevance score of document D:

$$S_{pc} = \sum_{C_k \in D} (f_{C_k}^{D} * W_{C_k}) \tag{2}$$

In formula (2): weight of class Ck: Wck and the frequency of class Ck in the document D: $f_{C_k}^{D}$.

In general, the relevance measure is a function that tries to map the content of a Web document against the E-Commerce ontology to gain a content relevance score. The value of score is larger, the content of web pages will be more relativity with the crawling topic.

C Ontology learning

Although using ontology as the background knowledge has actually improved the performance of focused Web crawler, the result mostly depends on the quality of the ontology which is predefined. If the ontology characterizes the specific topic poorly, low performance is inevitable. So we must find an approach which can evolve the original ontology to match the real topic environment. There are two methods for ontology learning, one is based on crawling process statistics, the other is based on the knowledge rules lib.

1) *Statistics-Based Learning*

The statistics-based learning module relies on the feedback from Web environment to topic specific ontology during the crawling process. From the feedback statistic topic-hit data we can learn the weight of classes in ontology and propagate evolution according to the ontology graph. As the definition in the relevance computation, weight of class denotes how the class is relevant to the topic. It is evident that those classes and sub-classes which always play the different roles during the crawling process. Thus our research tries to apply reinforcement learning, a framework for learning statistic weight of ontology to solve this problem.

Therefore, a task is defined by training data from preprocess and priority sequence module. In our research, training data consists of content based learning information K and relevancy entities frequencies. After preprocessing of the visited web pages, the entities (words occurring in the set of relevant concepts) are extracted and counted. We get the term frequencies of relevant concepts in the visited web pages. After calculating the term frequencies of relevant concepts, we can calculate the weight of each class.

From Definition 4, we can get the feedback statistic topic-hit data and calculate the topic ratio σ :

$$\sigma(t,c_o) = \frac{Nt \cdot n_o}{Nc \cdot n_{to}} \tag{3}$$

In formula (3):

Nc: number of URLs crawled;

Nt: number of URLs crawled which hit the topic t;

no: number of web pages crawled in which class o occurs;

nto: number of web pages crawled in which class o occurs, and which hit the topic t.

From formula (1), the evolution weight class is defined as follow:

$$W_{co}^{k} = \sigma(t,c_o) \cdot Wc_k \tag{4}$$

From the statistics-based learning we can proceed to evolve the ontology to help crawler work more and more efficiently.

2) *Rule-Based Learning*

The rule-based ontology learning contains the rules and methods to extract lexical and ontological elements from domain knowledge lib and concept relation database [12]. It consists of class learning rules, properties learning rules and hierarchy learning rules. It is fixed and is written once by the system programmer and will not be changed during the system's performance. During process of crawling, the domain ontology's structure could be optimized for the context of crawling by the rule-based learning module. The overall process of our ontology learning is presented in Figure 3.

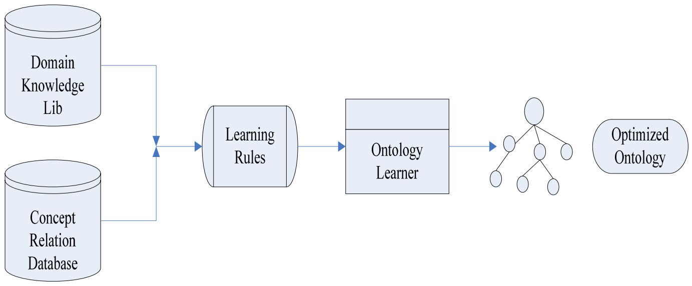

Figure 3. Ontology Rule-Based Learning Process

Rule 1 (Class learning) For relations R1, R2,⋯, Ri in relation database, supposed that P1= pkey(R1), P2 = pkey(R2),⋯, Pi = pkey(Ri) if R1(P1) = R2(P2) =⋯= Ri(Pi), then the information spread across R1, R2,⋯Ri should be integrated into an ontological class Ci. The ontological class Ci can be created based on the relation Ri, if there does not exist a relation that can be integrated with Ri, that is to say, Ri should not be integrated into the class Ci, and one of the following conditions can be satisfied:

(i) |pkey(Ri)| = 1;

(ii) |pkey(Ri)| > 1, and there exists Ai, where Ai∈ pkey(Ri) and Ai∉ fkey(Ri).

Rule1 integrates the information in several relations into one ontological class, when these relations are used to describe one entity. And that if the relation Ri is used to describe an entity, instead of relationship between relations, then it can be mapped into one ontological class.

Rule 2 (Properties learning): For relations Ri and Rj, if Ri(Ai)⊆Rj(Aj) and Ai⊄ pkey(Ri) are satisfied, then an object property P can be created based on Ai. Suppose that the classes corresponding to Ri and Rj are Ci and Cj respectively, the domain and range of P are Ci and Cj.

For relations Ri and Rj, two ontological objects property "has-part" and "is-part-of" can be created, if the two conditions are satisfied:

(i) |pkey(Ri)|>1;

(ii) fkey(Ri)∈pkey(Ri), where fkey(Ri) referring to Rj.

Suppose that the classes corresponding to Ri and Rj are Ci and Cj respectively, the domain and range of "is-part-of" are Ci and Cj and the domain and range of "has-part" are Cj and Ci. Properties "has-part" and "is-part-of" are two inverse properties.

Rule 3 (Hierarchy learning): For relation Ri and Rj, supposed that Pi = pkey(Ri), Pj = pkey(Rj), if Rule 1 does not applied and Ri(Pi)⊆Rj(Pj) is satisfied, then the class/property corresponding to Ri is a subclass/subproperty of the class/property corresponding to Rj.

Based the rules we can make the E-Commerce ontology more adaptive for the topic in the process of crawling, it is useful to the whole focused crawling.

D Link analysis

The hyperlink labels of a web page include the topic features. An important label is the text around links such as anchor text. They are the descriptions to pointed web pages. If the topic word turns up in these texts, the page that the link points to is likely the topic page. So we can collect topic relations from the hyperlinks, justly it is the function of the link analysis module.

The web is a hypertext system based on Internet. The biggest difference between hypertext system and ordinary text system is that the former has a lot of hyperlinks. Study shows, hyperlinked structure on the web has contained not only the characteristic of page importance, but also the topic relevance information. Because of the distribution characteristic of pages on the Web, we can have such conclusions: Page A pointed by Page B has a good relevance with the topic of A. Page linked by a topic page will be a worthy one. Therefore, we calculate the score of a page by PageRank. The algorithm has an effective use of huge link relations in the Web. It can be described as: one link from A pointing to B is considered as an A' providing vote to B, then we can judge the page importance according to the number of vote. The formula is:

PR (A) = (1-D) + D(PR (T1)/C (T1) + ... + PR (Tn)/C (Tn)) (5)

In formula (5),

PR(A)：The PageRank value of page A;

PR(Ti)：ThePageRank value of Ti (Topic pages) Pointing to page A;

C(Ti)：The number of out links of page Ti;

D: Damping(D∈(0,1))

The link priority score Spl of each candidate URL ui is calculated based on the link context (PageRank(n)) and the relevance of its parent pages (the pages which link to ui), and is defined as:

$$S_{pl}(u_i) = PR(u_i) + \sum_{j=1,d_j \to u_i}^{t} \lambda(d_j) PRS(d_j)/t \tag{6}$$

In formula(6),

dj is represented as a term frequency vector: dj = <tf1j, tf2j ….. tfnj >;

PR(ui) is the link context score of ui;

PRS(dj) is positive relevance score as:

$$PRS(d_j) = \frac{\sum_{k=1}^{|d_j|} tf_{ki}\omega_k}{\sqrt{\sum_{k=1}^{|d_j|}(tf_k)^2 \cdot \sum_{k=1}^{|d_j|}(\omega_k)^2}} \tag{7}$$

ω k is the weight for term tk calculated by the TF*IDF weighting scheme;

t is the total number of the parent pages with PRS value higher than the predefined threshold and λ (dj) is computed as:

$$\lambda(d_j) = \frac{t^+ + \theta}{t + \theta} \tag{8}$$

where t+ is the number of relevant child pages of dj that have been crawled, θ is a normalization factor set as 0.5. PR(ui) is calculated by formula (7) as the PRS value of the link context of ui, where the link context of ui is built by concatenating all the anchor texts of ui from its parent pages. The term $\sum_{j=1,d_j \to u_i}^{t} \lambda(d_j) PRS(d_j)/t$ refers to the contribution from the relevant parent pages to its child page. The factor λ (dj) controls the degree of the contribution called 'online link reinforcement'. If λ (dj) is high, dj will contribute more relevance to ui. This case is called strong link reinforcement; otherwise it is weak link reinforcement.

So we can get link priority score Spl by the link analysis module, it will be useful to calculate the crawl priority score.

E Crawl priority score

The aggregate interest ratio is a (weighted) product of the interest ratios for each of the individual priority score. We can equally combine the preferences by summing the weighted logarithms of individual factors.

$$S_p = r_1 \cdot S_{pc} + r_2 \cdot S_{pl} \tag{9}$$

Spc and Spl come from the formula (2) (6), the values r1 and r2are weights that are used in order to normalize the different factors. By increasing the weight of a given factor, we can increase the importance of the corresponding priority.

Finally, the priority sequence of the pages is determined upon the crawl priority score Sp, the focused crawler could fetch the topic web pages as more as possible

IV. PERFORMANCE EVALUATION

The primary factor which was used to evaluate the performance of the crawling system was the harvest rate P(C), which is the percentage of the web pages crawled satisfying the topic. We compare our approach with the standard keyword-based crawling, breadth-first crawling, and pagerank crawling approach. The seed URL is the E-Commerce web site "www.dangdang.com", and the predefine topic is "Book" product. The result is shown in detailed by figure4:

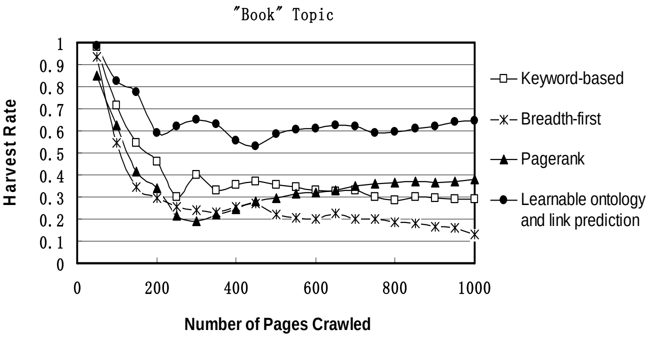

Figure 4. Comparison of several crawls in harvest rates

The crawl task described above has been processed several times to ensure the quality of the results. The four approaches maintain the higher harvest rate in the beginning as the related starting URL. But with the crawler processing and more pages being perceived, learning module gradually shows its effectiveness, our approach's harvest rate goes higher step by step and maintain the highest rate of the four approaches'. It is strongly testified the rationality of evolvement propagation.

Another performance metrics we provide is the performance time cost, figure 5 illustrates this in details.

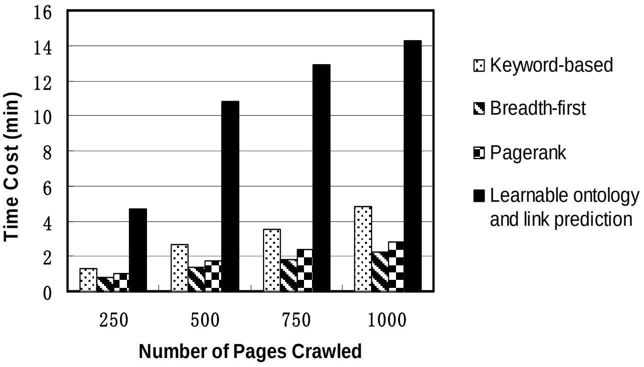

Figure 5. Comparison of several crawls in time cost

We can see that our approach needs much more time than other algorithms, if all crawlers are called for crawling equal number of pages. Because the ontology learning and link analysis module cost time to adaptive the domain context. But if marked with computation time and number of topics hitting pages, the efficiency of our approach is the best.

V. Conclusion

In order to maintain a high harvest rate during the crawling process, we propose ontology based intelligent crawling approach which includes a self-evolving mechanism that can dynamically adapt ontology to the particular environment of the relevant topic. The process of E-Commerce domain ontology evolving is based on machine learning including statistics and rules based. The framework of our approach combines content and hyperlink analysis, building the crawling sequence according to the score of content relevance and link priority. The empirical evaluation shows that our approach outperformed the keyword-based crawling, breadth-first crawling, and pagerank crawling. Based on the experiment results, our approach improves the efficiency of the focused crawler, which can be employed to build a topic-hit data collection for the E-Commerce domain information retrieval.

In the future, our efforts will be focused on improving the efficiency of the crawler and reducing the time of crawling. We look forward optimizing the mechanism of priority score computing and enhancing the precision of the topic repository of crawling.

Acknowledgement

The paper is supported by the National Natural Science Foundation of China (No.70803009), and the MOE Project of Key Research Institute of Humanities & Social Science at Universities (No.07JJD870220).

References

[1] Chakrabarti, S., van den Berg, M., Dom, B., "Focused crawling: a new approach to topic-specific Web resource discovery", Computer Networks, 31(11-16), 1999, pp.1623-1640.

[2] Rennie, J., McCallum, A.K., "Using reinforcement learning to spider the Web efficiently", 16th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, 1999, pp. 335-343.

[3] Diligenti, M., Coetzee, F., Lawrence, S., LeeGiles, C., Gori, M., "Focused crawling using context graphs", 26th International Conference on Very LargeDatabases, Cairo, Egypt, 2000, pp.527-534.

[4] Aggarwal, C.C., Al-Garawi, F., Yu, P.S., "Intelligent crawling on the world wide web with arbitrary predicates", Proceedings of the 10th international conference on World Wide Web, ACM Press, New York , 2001,pp. 96-105.

[5] Charu C. Aggarwal, "On the design of a learning crawler for topical resource discovery", ACM Transactions on Information Systems, Vol. 19, No. 3, 2001, pp.286-309.

[6] Maedche, A., Ehrig, M., Handschuh, S., Stojanovic, L., Volz, R., "Ontology-focused crawling of documents and relational metadata", Proceedings of the Eleventh International World Wide Web Conference WWW-2002, Hawaii, 2002, pp.347-354.

[7] S. Chakrabarti, K. Punera, and M. Subramanyam. "Accelerated focused crawling through online relevance feedback", International conference of WWW, ACM, Hawaii, 2002, pp.251-259.

[8] Ehrig, M., Maedche, A., "Ontology-focused crawling of web documents", SAC 2003: Proceedings of the 2003 ACM symposium on Applied computing, ACM Press, New York, 2003, pp.1174-1178.

[9] J.X. Yu, "iSurfer: A focused web crawler based on incremental learning from positive samples", APWeb 2004, LNCS 3007, 2004, pp.122-134.

[10] Su, C., Gao, Y., Yang, J., Luo, B., "An efficient adaptive focused crawler based on ontology learning", Fifth International Conference of Hybrid Intelligent Systems, 2005, pp.6-9.

[11] A. Rungsawang, N. Angkawattanawit, "Learnable topic-specific web crawler", Journal of Network and Computer Applications, 2005, pp.97-114.

[12] Man Li, "Learning ontology from relational database", Proceedings of 2005 International Conference on Machine Learning and Cybernetics, 2005, pp.3410-3415.

[13] Higham, D.J., “Google PageRank as mean playing time for pinball on the reverse web”, Applied Mathematics Letters, 18(12), 2005, pp.1359-1362.

[14] Valentina Tamma, “Ontologies for supporting negotiation in e-commerce”, Engineering Applications of Artificial Intelligence, 2005, pp.223-236.

[15] Hsu, C.-C., Wu, F., “Topic-specific crawling on the web with the measurements of the relevancy context graph”, Inf. Syst., 31(4), 2006, pp.232-246.

[16] Yiyao Lu, “Clustering e-commerce search engines based on their search interface pages using WISE-Cluster”, Data & Knowledge Engineering, 2006, pp.231-246.

[17] Andreia Malucelli, Daniel Palzer, “Ontology-based Services to help solving the heterogeneity problem in e-commerce negotiations”, Electronic Commerce Research and Applications, 2006, pp.29-43.

[18] Almpanidis, G., Kotropoulos, C., Pitas, I. “Combining text and link analysis for focused crawling-An application for vertical search engines”, Information Systems, 2007, pp.886-908.

[19] Lin, K., Zheng, J., “Ontology based learnable focused crawler”, Journal of Computational Information System, 2007, pp.1173-1180.

[20] Wei Fang, “Ontology-based focused crawling of deep web sources”, KSEM 2007, LNAI 4798, 2007, pp.514-519.

[21] K.C. Chang et al., “Efficiently crawling strategy for focused searching engine”, APWeb/WAIM 2007 Ws, LNCS 4537, 2007, pp.25-36.

[22] P. Brusilovsky, A. Kobsa, “The Adaptive Web”, LNCS 4321, Spring Press, 2007, pp.231-262.

[23] Lina Zhou, “Ontology learning: state of the art and open issues”, Inf Technol Manage, 2007, pp.241-252.

[24] Xuejun Nie, “A Domain Adaptive Ontology Learning Framework”, IEEE International Conference on Networking, Sensing and Control, 2008, pp.1726-1729.

[25] Hai-Tao Zheng, “Learnable focused crawling based on ontology”, AIRS 2008, LNCS 4993, 2008, pp.264-275.

[26] Jun Zhai, “Application of Fuzzy Ontology to Information Retrieval for Electronic Commerce”, Electronic Commerce and Security, 2008, pp.221-225.

[27] E. Kapetanios, V. Sugumaran, “An ontology-based focused crawler”, NLDB 2008, LNCS 5039, 2008, pp.376-379.

Wei Huang was born on November, 1979. He received the B.S. and Master degrees from Hubei University of Technology, Wuhan, China, in 2002 and 2005, respectively. He is currently a PH.D candidate of electronic commerce, Wuhan University, Wuhan, China. His research interests include information system, e-commerce and information retrieval.

He has worked at Hubei University of Technology since 2005. Now, he is a teacher of Information & E-commerce in School of Management.

Mr. Huang is a member of Association of Hubei Electronic Commerce.

Liyi Zhang was born on August, 1965. He graduated from Wuhan University of Hydraulic & Electric Engineering to obtain Bachelor Degree in 1988, from Xi’an jiaotong University to obtain the Master Degree in 1991 with specialty of Pattern Recognition & AI, and from Wuhan University to obtain PH.D in 1999 with specialty of System Engineering. His research interests include information system, e-commerce and information retrieval.

During 1991-2000, he taughe at Wuhan University of Hydraulic & Electric Engineering . Now, he is a professor and DEAN OF DEPARTMENT of Information & E-commerce in School of Information Management, Wuhan University. He has published five books, over 40 Journal papers. In addition, he has organized several conferences in the emerging areas of Electronic Commerce.

Mr. Zhang is a member of E-commerce Major Guiding Committee of China, the Secretary-general of Association of Hubei Electronic Commerce, and a member of AIS(Association of Information System).

Jidong Zhang He graduated from Wuhan University to obtain Bachelor Degree in 1998, and got the Master Degree in 2004, obtained PH.D in 2007. His research interests include information system, and electronic library.

He is an assistant professor of Hubei University of Technology.

Mingzhu Zhu received the B.S. degree from Wuhan University, Wuhan, China, in 2008.

He is currently a postgraduate candidate of electronic commerce, Wuhan University, Wuhan, China. His research interests include management information system, electronic commerce and information retrieval.