A Survey in Semantic Web Technologies-Inspired Focused Crawlers

Hai Dong
Digital Ecosystems and Business Intelligence Institute
Curtin University of Technology
hai.dong@cbs.curtin.edu.au

Farookh Khadeer Hussain
Digital Ecosystems and Business Intelligence Institute
Curtin University of Technology
farookh.hussain@cbs.curtin.edu.au

Elizabeth Chang
Digital Ecosystems and Business Intelligence Institute
Curtin University of Technology
elizabeth.chang@cbs.curtin.edu.au

Abstract

Crawlers are software which can traverse the internet and retrieve webpages by hyperlinks. In the face of the inundant spam websites, traditional web crawlers cannot function well to solve this problem. Semantic focused crawlers utilize semantic web technologies to analyze the semantics of hyperlinks and web documents. This paper briefly reviews the recent studies on one category of semantic focused crawlers – ontology-based focused crawlers, which are a series of crawlers that utilize ontologies to link the fetched web documents with the ontological concepts (topics). The purpose of this is to organize and categorize web documents, or filtering irrelevant webpages with regards to the topics. A brief comparison are made among these crawlers, from six perspectives - domain, working environment, special functions, technologies utilized, evaluation metrics and evaluation results. The conclusion with respect to this comparison is made in the final section.

1. Introduction

In most popular search engines, it is well-known that the ranks of websites for general queries are directly relevant to their economic benefits. Thus, a common phenomenon has emerged that many spam websites play tricks on search engine crawlers by artificially increasing links, in order to increase their weight in the PageRanks algorithm [10]. Semantic web technologies concentrate on analyzing the semantics of web document content, which could be helpful to solve this issue.

In this paper, we briefly review the existing ontology-based focused crawlers. A comparison is made among these crawlers from the perspective of domain, working environment, special functions, technologies utilized, evaluation metrics and evaluation results, in order to survey its current research status. The conclusion with respect to this comparison is made in the final section.

2. Ontology-based focused crawlers

Generally speaking, ontology-based focused crawlers are a series of crawlers which utilize ontologies to link the fetched web documents with the ontological concepts (topics), with the purpose of organizing and categorizing web documents, or filtering irrelevant webpages with regards to the topics [9].

There have been several studies on ontology-based focused crawlers, which are briefly described below:

Ehrig and Maedche proposed an ontology-focused crawler [3] [4]. Two cycles are involved in the crawling framework. In the first cycle, users can define a crawling target by instantiating a domain-specific ontology, and limit the crawling scope by providing the URLs of crawled websites. Based on the ontology and crawling scope, the focused crawler starts to work on retrieving data from those websites, and computing the relevance between the ontological concepts and the crawled data by means of TF-IDF algorithm. Its implementation – CATYRPEL is built upon KAON – a framework for ontology-based application development.

Ardö introduced a focused crawler working for the ALVIS – an open-source prototype of peer-to-peer semantic search engine [1]. The focused crawler is used to retrieve, cluster and store relevant webpages by linking them to topics. Each topic is defined by an ontology of terms. Relevance values between the terms and document texts are computed from both global (the whole database) and local (the topic ontology) perspective, by means of the topical PageRanks algorithm [6].

978-1-4244-2917-2/08/$25.00 ©2008 IEEE

Chen and Soo designed an ontology-based information gathering agent, aiming at searching and integrating knowledge based on users' queries. An ontology is defined as the agent's domain knowledge. Users can instantiate the ontology by adding partial values to an interested concept, in order to form a query for retrieving the values in the instance's blank fields. Four basic operations are involved in the gathering process – planning, search, information extraction and integration. In the fourth operation, when an agent extract useful information (values) from semi-structured and structured web documents, it needs to determine the integration possibility of these information by predefined domain-specific heuristics and integrating rules. If all blank fields of the instance are retrieved, the instance will be returned. An agent centre controls the whole gathering process based on the returned results [2].

Tane et al. proposed a new ontology management system – Courseware Watchdog. One important component of the system is an ontology-based focused crawler. By means of the crawler, a user can specify his/her preference, by assigning weights to the concepts of an ontology. By means of the interrelations between concepts within the ontology, the weights of other concepts can be calculated. Once a webpage is fetched, its text and URL descriptions are matched with the weighted ontological concepts. Thus, the weights of the webpage and its URLs are measured, ranked and clustered according to the concepts. In addition, the webpage relations can be viewed by linking the webpages to the ontology concepts that appear in the webpages [8].

THESUS aims to organize online documents by linking their URLs to hierarchical ontology concepts, which are seen as thematic subsets. A web crawler is used in the document acquisition component of the system. The mechanism of this crawler is as follows: first, the crawler extracts the URLs and their descriptive texts from the initial set of documents; then the descriptive text of one URL are matched with one of the ontological concepts, and the URL is linked to concept. A threshold of maximum times of recursions or maximum number of documents is set as an ending requirement [5].

As many ontology-based crawlers' stored ontologies cannot completely define crawling targets, Su et al. designed an ontology-learning focused crawler. The ontology learning mechanism originates from the theory of reinforcement learning – a decision-making framework based on reward or punishment points. First of all, a weight namely the distance between a concept and a topic is predefined. Then the crawler starts to retrieve documents based on the relevance values between each document and the ontological concept, which is computed by considering both the concept's weight and term frequencies in each document. An interest ratio to this topic is obtained by predicting the probability of a crawling event hitting the topic after the crawling process. Finally the weight is recomputed by considering an evolved weight based on the original weight and the interest ratio [7].

From the above introductions, we can observe that most of these crawlers utilize various ontology-document link analysis technologies to control crawling scope, cluster and retrieve web documents according to users' specific interest.

3. Comparison of the ontology-based focused crawlers

In this section, we will make a brief survey on these ontology-based focused crawlers from six perspectives – domain, working environment, special functions, technologies utilized, evaluation metrics, and evaluation results. The comparison result is shown in Table 1.

From the comparison table, it is observed that none of the crawlers are domain-specific. In other words, these crawlers can be used in any domains for any crawling topics. This multi-domain adaptability could be beneficial for the future development of these crawlers. For working environments, some of the crawlers are encapsulated in larger systems, while others are designed as separate tools. For special functions, most crawlers' ontological concepts' weights on query topics can be customized in order to highlight users' specific preference. One crawler can also provide the function that the crawled knowledge can be integrated according to domain-specific heuristic and rules, which could be useful to enhance the precision and reduce the recall. Another crawler can flexibly evolve the weights between concepts and topics through an ontology-leaning model. This could be helpful to solve the problem that predefined ontologies sometimes cannot completely inosculate the crawling topics. For utilized technologies, these crawlers use various technologies to satisfy different function requirements, except the commonly used ontology technology. In addition, the TF-IDF and PageRanks algorithm are adopted for the retrieved web documents ranking. While most crawlers do not provide evaluation methods, we still find that harvest rate is the primary metric to measure the crawlers' performance. Finally, the evaluation result shows that some ontology-based focused crawlers show some progress, compared with some traditional web crawlers.

4. Conclusion

This paper reviews part of the current researches on semantic focused crawlers – ontology-based focused crawler researches. In addition, we make a

brief comparison among these crawlers, from the perspective of domain, working environment, special functions, technologies utilized, evaluation metrics and evaluation results. From the comparison, it is found that various innovative technologies have been adopted to enhance the customizability and functionality of these crawlers. However, these researches are still in early stages, as there is no significant progress shown in their testing result.

Apart from the ontology-based crawlers, there are other categories of semantic focused crawlers including metadata abstraction crawlers and so on. The more detailed survey on these crawlers can be found in paper [9].

5. Acknowledgements

This research was supported by Digital Ecosystems and Business Institute (DEBII) in Curtin University of Technology (Australian CRICOS provider code: 00301J).

6. References

[1] A. Ardö, "Focused crawling in the ALVIS semantic search engine," in *2nd European Semantic Web Conference (ESWC 2005)*, Heraklion, 2005.

[2] Y.-J. Chen and V.-W. Soo, "Ontology-based information gathering agents," in *Web Intelligence: research and development*, N. Z. e. al., Ed. Maebashi: Springer-Verlag, 2001, pp. 423-427.

[3] M. Ehrig and A. Maedche, "Ontology-focused crawling of web documents," in *ACM Symposium on Applied Computing (SAC 2003)*, Melbourne, 2003.

[4] M. Ehrig, A. Maedche, S. Handschuh, L. Stojanovic, and R. Volz, "Ontology-focused crawling of web documents and RDF-based metadata," in *Intenational Semantic Web Conference 2002 (ISWC 2002)*, Sardinia, 2002.

[5] M. Halkidi, B. Nguyen, I. Varlamis, and M. Vazirgiannis, "THESUS: organizing web document collections based on link semantics," *The VLDB Journal,* vol. 12, pp. 320–332, 2003.

[6] G. S. Pedersen, A. Ardö, M. Cromme, M. Taylor, and W. Buntine, "ALVIS - superpeer semantic search engine," in *Research and Advanced Technology for Digital Libraries*, J. G. e. al., Ed. Alicante: Springer-Verlag, 2006, pp. 461-462.

[7] C. Su, Y. Gao, J. Yang, and B. Luo, "An efficient adaptive focused crawler based on ontology learning," in *the 5th International Conference on Hybrid Intelligent Systems (HIS'05)*, Rio de Janeiro, 2005.

[8] J. Tane, C. Schmitz, and G. Stumme, "Semantic resource management for the web: an elearning application," in *WWW2004*, New York, 2004.

[9] H. Dong, F. K. Hussain, and E. Chang, "State of the art in semantic focused crawlers," in *2009 IEEE International Conference on Industrial Technology (ICIT 2009)*, Gippsland, in press.

[10] A. Batzios, C. Dimou, A. L. Symeonidis, and P. A. Mitkas, "BioCrawler: An intelligent crawler for the semantic web," *Expert Systems with Applications,* vol. In Press, Corrected Proof, p. 908.

TABLE I
COMPARISON OF THE ONTOLOGY-BASED FOCUSED CRAWLERS

Name	Ontology-focused Crawler	ALVIS Cralwer	Information Gathering Agent	Courseware Watchdog Crawler	THESUS Crawler	Ontology-learning Focused Crawler
Domain	General	General	General	General	General	General
Working Environment	General	ALVIS search engine	General	Courseware Watchdog	THESUS	General
Special Functions	User can adjust the relevance computation strategy if s/he has special needs.	Using both of the gloabal and local ranking algorithm	An agent centre controls whole crawling process; three sorts of search procedures are provided for targeted information with different reliabilities; knowledge integration is realized based on domain-specific heuristic and rules.	Weighting similarity values between URLs and ontological concepts, and between parent pages and children pages.	Assigning weight to ontological concepts based on users' preference; weighting ranking and clustering webpages based on the weighted concepts.	Weights and propagation between concepts and topics can be altered through the crawling procedure by the ontology learning model and algorithms.
Technologies Utilized	TF-IDF for relevance computatation; KAON for prototype implementation.	PageRanks for web documents ranking.	Agent technology	Ontology and association metric for weighting similarity values between URLs and ontological concepts, and between parent pages and children pages.	Ontology for weighting, ranking and clustering webpages.	Reinforcement learning moudle for ontology evolvement.
Evaluation Metrics	Harvest rate	Not provided.	Not provided.	Not provided.	Not provided.	Harvest rate, crawling time
Evaluation Results	Less than 35% at the beginning and reduces to less than 15% along with the rise of crawled webpages.	Not provided.	Not provided.	Not provided.	Not provided.	Harvest rate in ontology-learnable crawlers are greater than normal ontology crawlers, but their time costs are also longer.