Self Adaptive Ontology based Focused Crawler for Mining Services Information Discovery

Anumol. A
PG Scholar
Computer Science and Engineering
Younus College of Engineering and Technology,
Kollam, India-691010

Sheema A. S
Assistant Professor
Information Technology
Younus College of Engineering and Technology,
Kollam, India-691010

***Abstract*—It is well recognized that the Internet has become the largest marketplace in the world, and online advertising is very popular with numerous industries, including the traditional mining service industry. The amount of data and its dynamicity makes it impossible to crawl the World Wide Web (WWW) completely moreover service users may encounter three major issues heterogeneity, ubiquity, and ambiguity, when searching for mining service information over the Internet. Its a challenge in front of crawlers to crawl only the relevant pages from this information explosion. Thus a focused crawler solves this issue of relevancy by focusing on web pages for some given topic or a set of topics. Web Crawlers are one of the most crucial part of the Search Engines to collect pages from the Web. The requirement of a web crawler that downloads most relevant web pages from such a large web is still a major challenge in the field of Information Retrieval Systems. Most Web Crawlers use Keywords base approach for retrieving the information from Web. But they retrieve many irrelevant pages as well. This paper, present the framework of a novel self-adaptive semantic focused crawler SASF crawler, with the purpose of precisely and efficiently discovering, formatting, and indexing by taking into account the heterogeneous, ubiquitous and ambiguous nature of mining service information available over the Internet. The framework incorporates the technologies of semantic focused crawling and ontology learning, in order to maintain the performance of this crawler.**

***Index Terms*—Mining service industry, ontology learning, semantic focused crawler, service advertisement, service information discovery.**

1. INTRODUCTION

A focused crawler may be described as a crawler which returns relevant web pages on a traversing the web pages. Web Crawlers are one of the most crucial part used by the Search Engines to collect pages from the Web and store in database. Most Web Crawlers use Keywords base approach for retrieving the information from Web. But they retrieve many irrelevant pages using crawler. Focused crawlers in particular, have been introduced for satisfying the need of individuals (e.g. domain experts) or organizations to create and maintain subject-specific web portals or web document collections locally or for addressing complex information needs (for which a web search would yield no satisfactory results). Applications of focused crawlers also include guiding intelligent agents on the Web for locating specialized information. Typical requirements of such application users are the need for high quality and up-to-date results, while minimizing the amount of resources (time, space and network bandwidth) to carry-out the search task. Focused crawlers try to download as many pages relevant to the subject as they can, while keeping the amount of not relevant pages downloaded to a minimum number. Crawlers (also known as Robots or Spiders) are tools for assembling Web content locally. Focused crawlers in particular, have been introduced for satisfying the need of individuals (e.g. domain experts) or organizations to create and maintain subject-specific web portals or web document collections locally or for addressing complex information needs (for which a web search would yield no satisfactory results). Applications of focused crawlers also include guiding intelligent agents on the Web for locating specialized information. Typical requirements of such application users are the need for high quality and up-to-date results, while minimizing the amount of resources (time, space and network bandwidth) to carry-out the search task. Focused crawlers try to download as many pages relevant to the subject as they can.

Crawlers used by general purpose search engines retrieve massive numbers of web pages regardless of their topic. Focused crawlers work by combining both the content of the retrieved Web pages and the link structure of the Web for assigning higher visiting priority to pages with higher probability of being relevant to a given topic. This paper, present the framework of a novel self-adaptive ontology focused crawler , with the purpose of precisely and efficiently discovering, formatting, and indexing mining service information over the Internet, by taking into account the three major issues. This frameworkincorporates the technologies of semantic focused crawling and ontology learning, in order to maintain the performance of this crawler, regardless of the

variety in the Web environment.

2. LITERATURE SURVEY

A semantic focused crawler is a software agent that is able to traverse the Web, and retrieve as well as download related Web information on specific topics by means of semantic technologies [1][2]. Since semantic technologies provide shared knowledge for enhancing the interoperability between heterogeneous components, semantic technologies have been broadly applied in the field of industrial automation. The goal of semantic focused crawlers is to precisely and efficiently retrieve and download relevant Web information by automatically understanding the semantics underlying the Web information and the semantics[5] underlying the predefined topics. Most of the crawlers in this domain make use of ontologies to represent the knowledge underlying topics and web documents. However, the limitation of the ontology-based semantic focused crawlers is that the crawling performance crucially depends on the quality of ontologies. This is because ontologies are often designed by domain experts. A discrepancy may exist between the domain experts in understanding of domain knowledge. Moreover Knowledge is dynamic. Information in web are updating day by day and hence this technique will not work in real web Environment. Problems may arise when new terms beyond the concepts in ontology arises.

In order to solve the defects in ontologies and maintain or enhance the performance of semantic-focused crawlers, researchers have begun to pay attention to enhancing semantic-focused crawling technologies by integrating them with ontology learning technologies. The goal of ontology learning is to semi-automatically extract facts or patterns from a corpus of data and turn these into machine-readable ontologies[3].

Zheng et al. [4] proposed a supervised ontology-learning based focused crawler that aims to maintain the harvest rate of the crawler in the crawling process. The main idea of this crawler is to construct an artificial neural network (ANN)model to determine the relatedness between a Web document and an ontology. The main idea of this crawler is to construct an artificial neural network (ANN)model to determine the relatedness between a Web document and an ontology. Given a domain-specific ontology and a topic represented by a concept in the ontology, a set of relevant concepts are selected to represent the background knowledge of the topic by counting the distance between the topic concept and the other concepts in the ontology. The crawler then calculates the term frequency of the relevant concepts occurring in the visited Web documents. Backpropogation algorithm is used to train three-layer feedforward ANN model.The output of the ANN is the relevance score between the topic and a Web document. The training process follows a supervised paradigm, whereby the ANN is trained by labeled Web documents. The training will not stop until the root mean square error (RMSE) is less than 0.01. The limitations of this approach are: 1) it can only be used to enhance the harvest rate of crawling but does not have the function of classification; 2) it cannot be used to evolve ontologies by enriching the vocabulary of ontologies; and 3) the supervised learning may not work within an uncontrolled Web environment with unpredicted new terms.

In order to solve the defects in ontologies and maintain or enhance the performance of semantic - focused crawlers, researchers have begun to pay attention to enhancing semantic- focused crawling technologies by integrating them with ontology learning technologies. The goal of ontology learning is to semi-automatically extract facts or patterns from a corpus of data and turn these into machine-readable ontologies. Various techniques have been designed for ontology learning, such as statistics-based techniques, logic based techniques, etc. These techniques can also be classified into supervised techniques, semi-supervised techniques, and unsupervised techniques from the perspective of learning control. Obviously, ontology-learning-based techniques can be used to solve the issue of semantic-focused crawling, by learning new knowledge from crawled documents and integrating the new knowledge with ontologies in order to constantly refine the ontologies.

3. PROPOSED SYSTEM

The purpose of this self adaptive ontology focused crawler is to precisely and efficiently discovering, formatting, and indexing mining service information over the Internet. The flowchart of the System architecture is shown in Fig:1

Figure 1.

The Mining Service Ontology Base is used to store a mining service ontology, which is the representation of specific mining service domain knowledge. Concepts in the mining service ontology are organized in a hierarchical structure, and these concepts are associated by a generalization/specialization relationship. The hierarchical structure of ontology is as shown in fig:2 Each concept in the mining

Figure 2.

service ontology represents a mining service sub-domain and have a concept description. The conceptDescription property is a datatype property used to store the textual descriptions of a mining service concept, which consists of several phrases in order to concisely summarize the discriminate features of the corresponding mining service subdomain. The contents of each conceptDescription property are manually specified by domain experts and this will be used to calculate the similarity value between a mining service concept and a mining service metadata.

The serviceDescription property is a datatype property which contains the texts extracted from the mining service advertisements by the focused crawler. This property will be used for the subsequent concept-metadata similarity computation.

3.1. System Workflow

As can be seen in Fig:1 the first step is preprocessing, which is to process the contents of the conceptDescription property of each concept in the ontology before matching the metadata and the concepts. This processing is realized by using JavaWordNet Library8 (JWNL) to implement tokenization, part-of-speech (POS) tagging, nonsense word filtering, stemming, and synonym searching for the conceptDescription property values of the concepts.

The second and third steps are crawling and term extraction. The aim of these two processes is to download Web pages from the Internet at one time and to extract the required information from the downloaded Web pages. The next step is term processing, which is to process the content of the serviceDescription property of the metadata in order to prepare for subsequent concept-metadata matching.

In the algorithm-based string matching process consist of three types of string matching processes-1) Semantic based matching, 2) Statistics based matching and the top of these two matching 3) A hybrid matching is also performed.

In semantic based matching the semantic relatedness between the concept and the metadata is determined by comparing their algorithm-based property similarity values with a threshold value.If the maximum similarity value between the serviceDescription property value of a metadata and the conceptDescription property values of a concept is higher than the threshold value, themetadata and the concept are regarded as semantically relevant; otherwise not. The key idea of the SeSM algorithm is to measure the text similarity between a concept description and a service description, by means of WordNet. As the concept description and the service description can be regarded as two groups of terms after the preprocessing and term processing phase, first of all, we need to examine the semantic similarity between any two terms from these two groups. Since terms (or concepts) in WordNet are organized in a hierarchical structure, in which concepts have the relationships of hypernym/hyponym, it is possible to assess the similarity between two concepts by comparing their relative position in WordNet.

Next step is statistics based string matching is performed. In statistics based matching the maximum probability that Concept description and service description coexist in a web page.

On top of the Semantic based string matching and Statistics based string matching, a hybrid matching is required to seek the maximum similarity values from the two Semantic based matching and Statistics based matching.

4. SYSTEM EVALUATION

This section evaluate the proposed crawler crawler by comparing its performance with that of the existing ontology- learning-based focused crawlers of Zheng et al. and Su et al. 1)Testing Data Source:one common defect of the existing ontology-learning-based focused crawlers is that these crawlers are not able to work in an uncontrolled Web environment with unpredicted new terms, due to the limitations of the adopted ontology learning approaches. Hence, our proposed crawler aims to remedy this defect, by combining a real-time SeSM algorithm, and an unsupervised StSM algorithm. In order to evaluate our model and the existing models in the uncontrolled Web environment, we choose two mainstream business advertising websites Australian Kompass14 (abbreviated as Kompass below) and Australian Yellowpages, as the test data source. There are around 800 downloadable mining-related service or product advertisements registered in Kompass, and around 3200

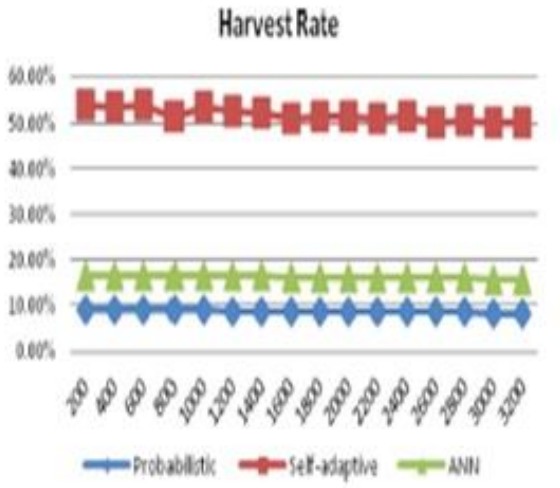

Figure 3.

similar advertisements registered in Yellowpages. All of them are published in English.

2)Test Environment and Targets: Owing to the primary objective of the ontology-learning-based focused crawlers, that is, to precisely and efficiently download and index relevant Web information, we subsequently employ the ontology learning and Web page classification models used in each of the focused crawlers, namely the ANN model used in Zheng et al.s crawler, the probabilistic model used in Su et al.s crawler, and our selfadaptive Ontology based focused model, together in our Ontology based crawler framework, for the task of metadata harvesting and indexing (or classification) (i.e., the Metadata Generation and Ontology Learning process in Fig. 1), and compare their performance in this process. It is recognized that all of these models need a training process before harvesting and classification, since the ANN model follows a supervised training paradigm, and the other two models follow an unsupervised training paradigm. Therefore, we use the Kompass website as the training data source, and label the Web pages from this website for the ANN training. Following that, we test and compare the performance of these models by using the unlabelled data source from Yellowpages, with the purpose of evaluating their capability in an uncontrolled environment. In addition, by means of a series of experiments, we find that 0.6 is the optimal threshold value for the self-adaptive model to determine the relatedness between a pair of service description and concept description.

3)Harvest Rate: The graphic representation of the comparison of the probabilistic, self-adaptive and ANN models on harvest rate, along with the increasing number of visited Web pages, is shown in Fig.3.

It needs to be noted that the harvest rate concerns only the crawling ability, not the accuracy, of a crawler. A high proportion (around 40%) ofWeb pages are peer-reviewed as non-mining-service-related Web pages in the unlabeled data source,which is part of the reason that the overall harvest rates of these three models are all below 60%. It can be seen that the self-adaptive model has the optimal performance (more than 50%), compared to the ANN model (around 16%) and the probabilistic model (between 8% and 9%). This proves that the self-adaptive model has a positive impact on improving the crawling ability of the ontology focused crawler, as more service descriptions extracted from the Web pages are matched to the learned concept descriptions.

5. CONCLUSION

In this paper, we presented a self adaptive ontology-learning based focused crawler, for service information discovery in the mining service industry.This approach involved an innovative unsupervised ontology learning framework for vocabulary-based ontology learning, and a novel concept-metadata matching algorithm, which combines a semantic-similarity-based SeSM algorithm and a probability-based StSM algorithm for associating semantically relevant mining service concepts and mining service metadata. This approach enables the crawler to work in an uncontrolled environment where the numerous new terms and ontologies used by the crawler have a limited range of vocabulary. Subsequently, we conduct a series of experiments to empirically evaluate the performance of the self adaptive ontology based crawler.

6. FUTURE WORK

Apart from the work done towards this system, future work mainly comprises of the following objectives:

In future we try to find a universal threshold value to set up a boundary for determining concept-metadata relatedness. And also we try to introduce a new matching algorithm so as to improve the performnace of the ontology based crawler.

REFERENCES

[1] H. Dong and F. K. Hussain, Focused crawling for automatic service discovery, annotation, and classification in industrial digital ecosystems, IEEE Trans. Ind. Electron., vol. 58, no. 6, pp. 21062116,
Jun. 2011.

[2] H. Dong, F. K. Hussain, and E. Chang, A framework for discovering and classifying ubiquitous services in digital health ecosystems, J. Comput. Syst. Sci., vol. 77, pp. 687704, 2011.

[3] W. Wong, W. Liu, and M. Bennamoun, Ontology learning from text: A look back and into the future, ACM Comput. Surveys, vol. 44, pp. 20:136, 2012.

[4] H.-T. Zheng, B.-Y. Kang, and H.-G. Kim, An ontology-based approach to learnable focused crawling, Inf. Sciences, vol. 178, pp. 45124522,2008.

[5] J. L. M. Lastra and M. Delamer, Semantic web services in factory automation: Fundamental insights and research roadmap, IEEE Trans. Ind. Informat., vol. 2, no. 1, pp. 111, Feb. 2006.