Survey on Self Adaptive Semantic Focused Crawling Using Ontology Learning

[1]S.Bhargavi,
[1]Final year, M.Tech,
[1]Department of Computer Science and Engineering,
[1]SRM University, Kattankalathur, Chennai, India

***Abstract* - The Internet today has become a vast storehouse for a scintillating amount of knowledge. It is an excellent source of information catering to the needs of people of varied interests. But this process of information retrieval does have its shortcomings too viz. heterogeneity, ubiquity and ambiguity. Thus a self-adaptive semantic focused crawler - SASF crawler that addresses these issues and optimizes the process of information discovery and indexing of the searched information is proposed. This framework encompasses the concepts of semantic focused crawling and ontology learning that helps to maintain the performance of the crawler in spite of the variety in the Web environment. The innovation here is the unsupervised vocabulary-based ontology learning and a hybrid matching algorithm that matches semantically relevant concepts and metadata. Finally the performance of the crawler is evaluated based on various parameters.**

***Index Terms* - Ontology learning, Semantic focused crawling, Hybrid matching algorithm, SASF crawler, Information retrieval**

I.INTRODUCTION

Information technology has completely revolutionized the way information is being searched and utilized in recent days. Anything and everything that is needed can be just obtained just within a few clicks courtesy of the intricately linked structure of the Web. This process of searching the Internet with the usage of search engines is facilitated by a special software called crawler. Crawlers [1] also called Robots or Spiders are the actual entities that are responsible for retrieving the relevant documents whenever a keyword search is initiated by the user. Focused crawlers are a special kind of crawlers that lead the search in a more domain specific direction. They help to maintain subject-specific web portals or web document collections locally. These repositories are meant to store complex, highly relevant, up-to-date information whilst keeping resource usage to a minimum.

Crawlers usually start their search with a set of initial pages called the seed pages. Based on the output links from these seeds the documents to be retrieved further are determined. Documents searched next are those that satisfy certain relevance criteria. This process continues until the desired number of pages are downloaded or when the local repository is exhausted. In spite of all these optimizing techniques the information discovery still has the following issues:

A. Heterogeneity

There are several versions and forms of the same information on the Web. To decide which is the version best suited to the user's needs is a challenge.

B. Ubiquity

Information is omnipresent and available everywhere on the Web. To decide which the best source of information is, is the next issue to be addressed.

C. Ambiguity

Most of these crawlers base their retrieval activity on keywords which causes some irrelevant documents also to be returned to the user. To decide which is the most relevant information, is also one of the concerns. To tackle the aforementioned issues, a self-adaptive semantic focused crawler is proposed which enables to return more relevant documents to the user, increase user satisfaction and thus its own efficiency.

The paper is organized into following sections: Section II deals with the related work in this field, Section III presents the SASF crawler and finally Section IV evaluates the performance of the crawler based on specific parameters.

II.RELATED WORK

In this section the previous work done in regard to semantic focused crawling and ontology learning are reviewed.

A semantic focused crawler is a software agent that is able to traverse the Web, and retrieve and to download related Web information non specific topics by means of semantic technologies [2][3]. 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 underlying the pre-defined topics. A survey conducted by Dong et al. [4] found that most of the crawlers in this domain make use of ontologies to represent the knowledge underlying topics and Web documents. The limitation of the ontology-based semantic focused crawlers is that the crawling performance crucially depends on the quality of ontologies.

Furthermore, the quality of ontologies may be affected by the following two issues. The first issue is that, as it is well known that ontology is the formal representation of specific domain knowledge [5] and ontology are designed by experts from domain, a discrepancy exists between the domain experts' understanding of the domain knowledge and the domain knowledge that exists in the real world. This issue is that knowledge is dynamic and is constantly evolved comparing with relatively static ontologies. Two contradictory situations could possibly leads to the problem that ontologies sometimes cannot precisely represent real-world knowledge. The eventual consequence of these problems is reflected in the gradually descending curves in the performance of semantic focused crawlers.

In order to solve the defects in ontologies and maintain or enhance the performance of the semantically focused crawlers, researchers have started to pay attention to enhancing semantic-focused crawling technologies by integrating them with ontology learning technologies. The main aim of ontology learning is semi-automatically extract the facts or and its patterns from a corpus of data and then to turn these into machine-readable ontologies [6]. Various techniques have been designed for ontology based learning, such as statistics-based techniques, linguistics (or natural language processing)-based techniques, logic-based techniques. These different types of techniques can also be divided into three techniques as, supervised, semi-supervised and unsupervised techniques from the perspective of learning control.

Zheng et al. [7] proposed a supervised ontology-learning based focused crawler that aims to maintain the harvest rate of the crawler in the crawling process. The important theme of this crawler is to construct an artificial neural network (ANN) model to determine the relatedness between a Web document and ontology. Given a domain-specific ontology and a topic represented by a concept in this ontology is a set of concepts that are relevant and are selected to represent the back ground knowledge of the topic by counting the distance between the topic concept and the other concepts in the ontology. This crawler then calculates the frequency of those relevant concepts occurring in the visited Web documents. Next, the authors used the back propagation algorithm to train a three-layer feed forward ANN model. The output of this ANN is the relevant between the topic and a Web document. The training process is considered to be following a supervised paradigm, in which the ANN model is trained by labeled Web documents. The training process will not stop till the root mean square error (RMSE) is less than0.01.The limitations of this approach are:

1) It can be used in enhancing the harvest rate of crawlers but does not have the function of classification;
2) It cannot be used in evolving ontologies by enriching its vocabulary of those ontologies; and
3) The supervised learning may not work within an uncontrolled Web environment with unpredicted new terms.

Suet al. [8] proposed unsupervised ontology learning based focused crawler in order to compute the relevance scores between topics and Web documents. Given specific domain ontology and a topic represented by a concept in this particular ontology, the score that is relevant between a Web document and the topic is the weighted sum of the occurrence frequencies of all the concepts of the ontology in the Web document. Next, this crawler makes use of reinforcement learning, a probabilistic framework for learning optimal decision making from rewards or punishments [9], in order to train. The learning step follows an unsupervised paradigm, in which the crawler is used to download a number of Web documents and learn statistics based on these Web documents. The learning step can also be repeated many times.

Compared with Zhengetal.[7], approach, this approach is capable of classifying Web documents by means of the concepts in an ontology, in order to learn the weights of relations between concepts. The limitations of Su et al.'s approach are:

1) It cannot be used in enriching the vocabulary of those created ontologies
2) Though the unsupervised learning paradigm can work in an uncontrolled Web environment, it does not work well when there are numerous new terms emerge or when ontologies have a limited range of vocabulary.

By means of a comparative analysis of those two ontology based focused crawlers, a common limitation is found, which is that none of the two crawlers is able to really evolve ontologies by enriching their vocabulary contents. It is revealed that both of the approaches attempt to use learning models to deduce the quantitative relationship between the occurrence frequencies of the concepts in ontology and their topic, which may not be applicable in the real world Web development environment. When large number of unpredictable new terms outside the scope of the vocabulary of ontology emerges in Web documents, these approaches cannot determine the relation between the new terms and their topic, and it cannot make use of those new terms for which the determination of relatedness, which could result in the decline of their performance.

III. SYSTEM COMPONENTS AND WORKFLOW

A. System Components

In this section, the system architecture and the workflow of the proposed SASF crawler is described. It needs to be noted that this crawler is built upon the semantic focused crawler designed by Dong et al [2], [3]. The differences between this work and the previous work are that the previous research work created a purely semantic focused crawler, that does not have an ontology-based learning function to automatically evolve the utilized ontology.

In this research, mining service ontology and a mining service metadata schema are designed to solve the problem of self-adaptive service information discovery. An overview of the system architecture and the workflow is shown in Fig. 1. SASF crawler consists of two knowledge bases – a Mining Service Ontology Base and a Mining Service Meta database.

An ontology learning- based focused crawler [10] is proposed, in order to discover precisely, its format and its index relevant Web documents in the uncontrolled Web environment. The proposed SASF crawler is built upon the architecture of a semantic focused crawler. The Mining Service Ontology Base is used to store the service ontology, which is used to represent the specific domain knowledge. The Mining Service Metadata Base is used to store the automatically generated and indexed service metadata. Service metadata is the abstraction of an actual information published in a Web document.

B. *Workflow*

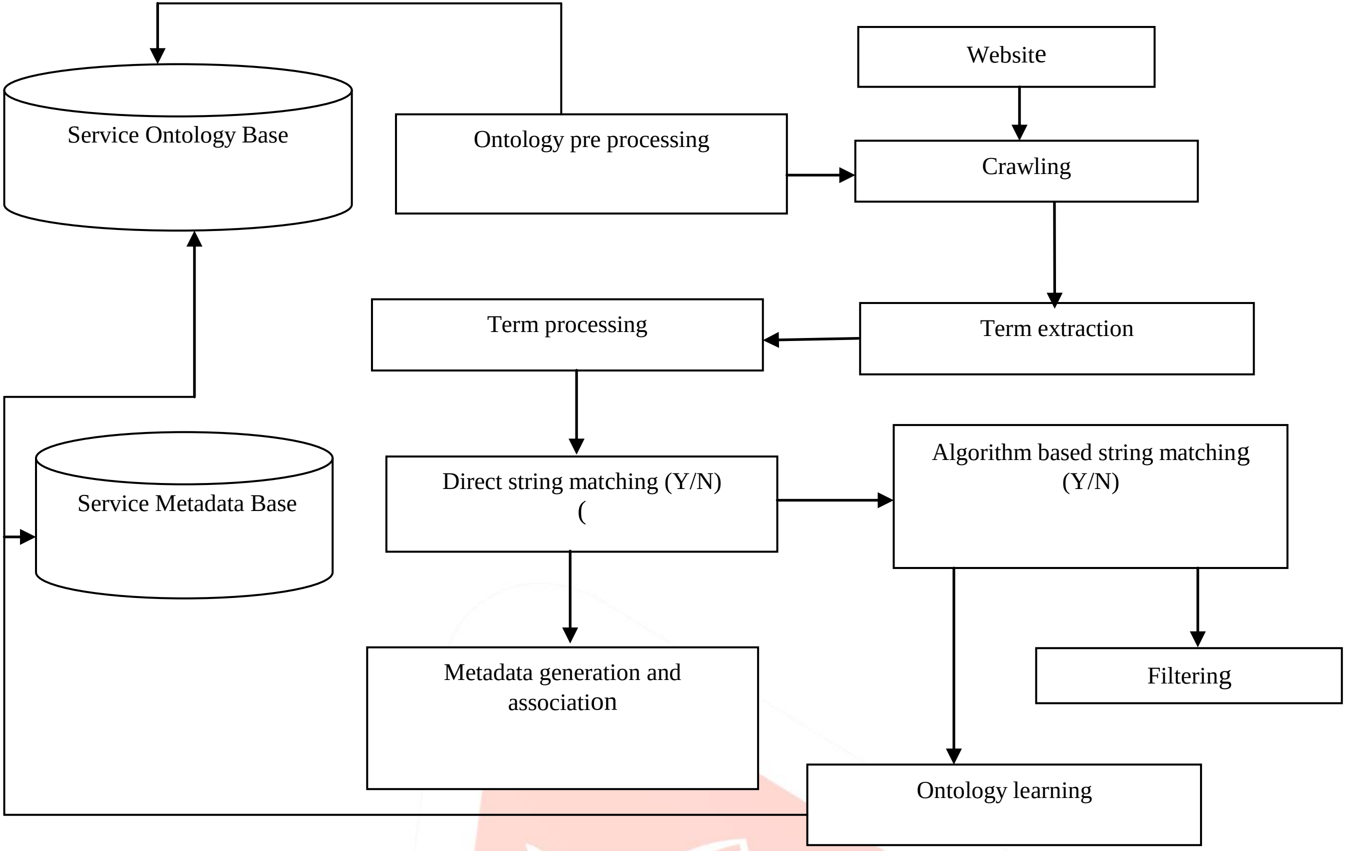

Fig: 1Architecture of semantic crawler using ontology

The system consists of five modules:

1. Ontology pre-processing
2. Crawling
3. Term extraction
4. Term processing
5. Metadata association and ontology learning

1. Ontology pre-processing:

The contents of each concept in the ontology are processed before matching the metadata and the concepts. This processing is realized by implementing tokenization, part-of-speech (POS), tagging of words, useless word filtering, stemming, and inter related meaning searching for the particular concepts.

2. Crawling:

In this phase k Web pages are downloaded from the Internet at one time.

3. Term extraction:

This module extracts the required information from the downloaded Web pages, according to the service metadata schema and the service provider metadata schema, in order to prepare the property values to generate a new group of metadata.

4. Term processing:

The contents of the metadata are processed in order to prepare for subsequent concept of matching metadata. The implementation of this process is similar to the implementation of the preprocessing processing technique its major difference is that the term processing does not need the function of synonym searching for two major reasons:

1) The synonyms of the terms in the concept-Description properties of concepts have already been retrieved in preprocessing and 2) The computational cost of the synonym searching for the metadata is relatively high and this may influence the scalability of the SASF crawler, as term processing is a real-time process.

5. Metadata association and ontology learning:

First of all, the direct string matching process examines whether or not the contents of the metadata are included in that of a concept. If the answer is 'yes', then the concept and the meta data are regarded as semantically relevant data.

By means of generating metadata and its association process, the metadata can also be generated and it is stored in the mining service metadata base as well as it is being associated with the concept. If the answer is 'no', an algorithm-based string matching process will be invoked to check the semantic relatedness between the metadata and their concept, by means of a concept- based metadata semantic similarity algorithm. If the concept and the metadata are semantically relevant, the contents of the metadata can be regarded as anew value for the concept. The metadata is thus allowed to go through the metadata generation and association process; otherwise the metadata is regarded as semantically irrelevant to the concepts used. The above process is repeated until all the concepts in the mining service ontology have been compared with those metadata. If none of these concepts is semantically relevant to those metadata, then those metadata is regarded as semantically non-relevant to the mining service domain and will be dropped.

IV. SYSTEM EVALUATION

The parameters for comparison between this crawler and the existing ontology-learning-based focused crawlers are adopted from the field of IR and need to be re-defined in order to be applied in the scenario of ontology-based focused crawling.

Harvest rate is used to measure the harvesting ability of a crawler. Precision is used to measure the preciseness of a crawler. Recall issued to measure the effectiveness of a crawler. Harmonic mean is a measure of the aggregated performance of a crawler. Fallout is used to measure the inaccuracy of a crawler. Crawling time issued to measure the efficiency of the crawler used. The average crawling time of the SASF crawler for Web page is defined as the time interval from processing the Web page from the Crawling process to the Metadata Generation and Association process or to the Filtering process. From Fig.2 it is clearly visible that the self-adaptive model is far superior in performance than the probabilistic model in all of the parameters.

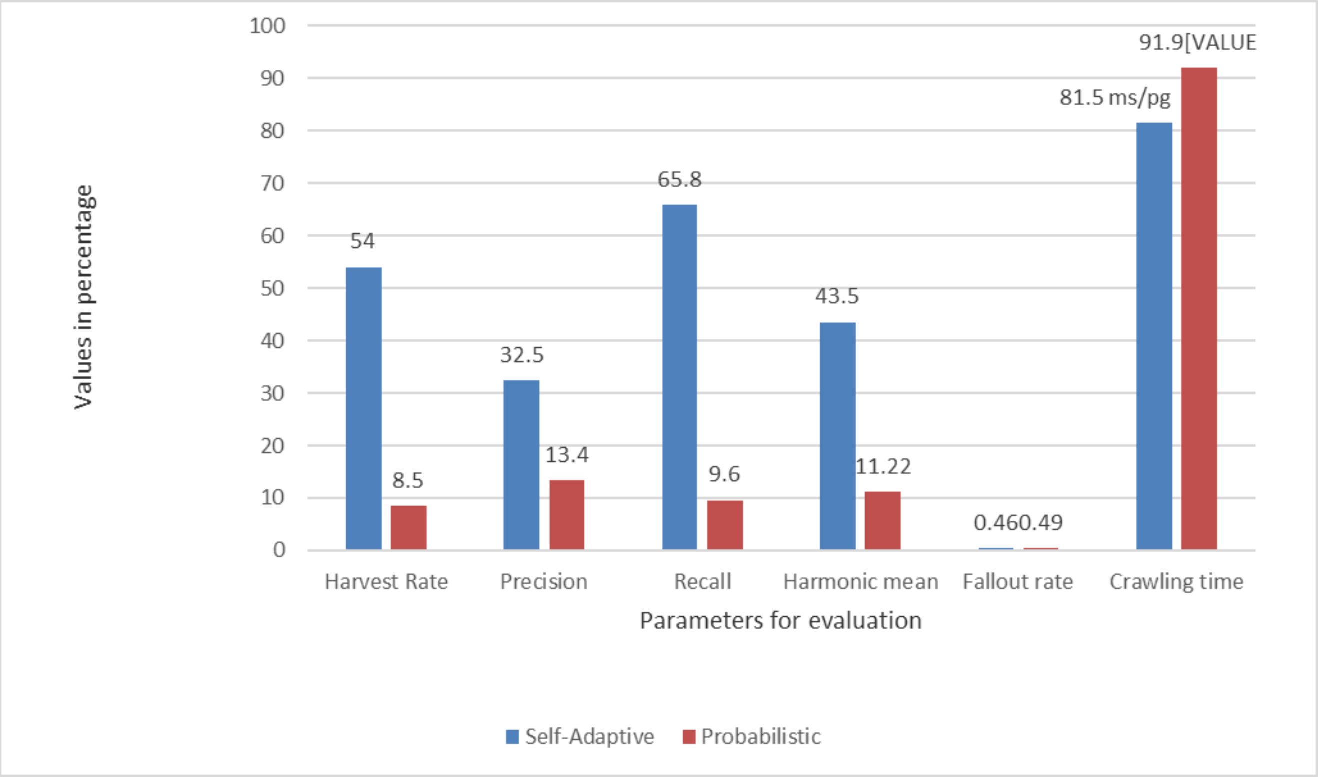

Fig.2: Comparative performance analysis between the different crawler models

The same goes for the comparison between self-adaptive and ANN models except for their crawling times. The ANN model offers a crawling time of 77ms/page for the first 200 pages whilst the self-adaptive model offers 81.5ms/page. But as the number of pages crawled increases the crawling time of the ANN model falls slowly. On the other hand self-adaptive model gains speed after 200 pages and improves on its crawling time due to the ontology learning incorporated. Thus in the long run the self-adaptive model outdoes the ANN model too. On the whole the performance curves of the self-adaptive model are relatively smooth; regardless of its variety in the Web pages that has been visited.

V.REFERENCES

[1] Shwetha Jog,Shubham Joshi,"Review on Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery",International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-Vol. 3 Issue 12, December-2014

[2] H. Dong and F. K. Hussain, "Focused crawling for automatic service discovery, annotation, and classification in industrial digital ecosystems," IEEETrans.Industrial .Electronics, vol.58, no.6, pp.2106–2116, Jun. 2011.

[3] H.Dong, F.K.Hussain ,and E.Chang,"A framework for discovering and classifying ubiquitous services in digital health ecosystems," J. Computer Syst. Sci., vol. 77, pp. 687–704, 2011.

[4] H. Dong, F. Hussain, and E. Chang, O. Gervasi, D. Taniar, B. Murgante, A.Lagana, Y.Mun, and M.Gavrilova, Eds., "State of the art in semantic focused crawlers," in Proc. ICCSA 2009, Berlin, Germany, 2009, vol. 5593, pp. 910–924.

[5] T. R. Gruber, "A translation approach to portable ontology specifications," Knowledge Acquisition, vol.5, pp.199–220, 1993.

[6] W.Wong, W.Liu, and M.Bennamoun, "Ontology learning from text: A look back and into the future," ACM Computer Surveys, vol. 44, pp. 20:1–36, 2012.

[7] H.-T. Zheng, B.-Y. Kang, and H.-G. Kim, "An ontology-based approach to learnable focused crawling," Inf. Sciences, vol. 178, pp. 4512–4522, 2008.

[8] C. Su, Y. Gao, J. Yang, and B. Luo, “An efficient adaptive focused crawler based on ontology learning,” in Proc. 5th Int. Conf. Hybrid Intell.Syst.(HIS’05),RiodeJaneiro,Brazil,2005,pp.73–78.

[9] J. Rennie and A. McCallum, “Using reinforcement learning to spider the Web efficiently,” in Proc. 16th Int. Conf. Mach. Learning (ICML ’99), Bled, Slovenia, 1999, pp. 335–343.

[10] Hai Dong and Farookh Khadeer Hussain, “Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery” IEEE Transactions on Industrial Informatics, Vol. 10, No. 2, May 2014.