Ontology Based Approach for Services Information Discovery using Hybrid Self Adaptive Semantic Focused Crawler

Swapnil V. Patil
Department of Computer Engineering
JSPM'S JSCOE,
Pune, India
swapnilpatil96@gmail.com

Sharmila M. Shinde
Department of Computer Engineering
JSPM'S JSCOE,
Pune, India
sharmi_anant@yahoo.co.uk

Abstract— Focused crawling is aimed at specifically searching out pages that are relevant to a predefined set of topics. Since ontology is an all around framed information representation, ontology based focused crawling methodologies have come into exploration. Crawling is one of the essential systems for building information stockpiles. The reason for semantic focused crawler is naturally finding, commenting and ordering the administration data with the Semantic Web advances. Here, a framework of a hybrid self-adaptive semantic focused crawler – HSASF crawler, with the inspiration driving viably discovering, and sorting out administration organization information over the Internet, by considering the three essential issues has been displayed. A semi-supervised system has been planned with the inspiration driving subsequently selecting the ideal limit values for each idea, while considering the optimal performance without considering the constraint of the preparation of data set.

Keywords- *Hybrid Self-adaptive semantic focused crawler (HSASF), semantic web, ontology learning, dynamic threshold.*

*****

I. INTRODUCTION

Now a days, the exploration of focused crawler approaches the field of semantic web, alongside the presence of expanding semantic web reports and the fast improvement of ontology markup dialects. Internet is a repository containing a huge measure of archives and hyperlinked documents. The data on the web is shared in view of the client interest. The search through these chronicles considers the user queries furthermore, retrieves documents that are connected. A web crawler is a product specialist that can consequently peruse and download site pages from the web. A focused crawler must incredible the likelihood that an unvisited page will be applicable before really downloading the page.

Focused crawling is gone for specifically searching out pages that are significant to a predefined arrangement of themes. Since ontology is an all around framed information representation, ontology based focused crawling methodologies have come into exploration. Crawling is one of the essential systems for building information stockpiles. Semantic focused crawler makes utilization of domain ontologies to speak to topical maps and connection Web pages with important ontological ideas for the choice and categorization purposes. Moreover, ontologies can be consequently overhauled in the crawling procedure.

Past research work made a simply semantic focused crawler, not having an ontology learning capacity to naturally advance the used ontology. This research has a goal to cure this deficiency. Previously related work utilized the service ontology and the service metadata forms, particularly intended for the transportation administration space and the medicinal services administration domain.

There is need to concentrate on finding successfully and precise data over the web. Likewise focus universal threshold value dynamically for idea metadata relatedness. It is necessary to design framework that enables the crawler to work in an uncontrolled web.

It is well recognized that information technology has a profound effect on the way business is led, and the Internet has become the wide marketplace in the world [1]. Innovative business experts have understood the business utilizations of the Internet both for their clients and key accomplices, transforming the Internet into a huge shopping center with an immense list. Purchasers have the capacity to peruse an immense scope of items and administration promotions over the Internet, and purchase this merchandise specifically through online exchange frameworks. Service advertisements frame an impressive piece of the promoting which happens over the Internet and have the accompanying components: Heterogeneity, Ubiquity, and Ambiguity [1].

In this paper, we present the framework of a hybrid self-versatile semantic focused crawler – HSASF crawler, with the motivation behind accurately and proficiently finding, and organizing data over the Internet, by considering the three noteworthy issues. This structure fuses the innovations of semantic focused crawler and ontology adjusting, taking into consideration the important aim to keep up the execution of this crawler, paying little heed to the mixture in the Web environment. The developments of this research lie in the outline of an unsupervised framework for vocabulary-based ontology learning, and a hybrid algorithm for coordinating semantically applicable concepts and metadata. A semi-supervised methodology has been designed has been planned with the motivation behind consequently selecting the optimal threshold values for every concept, while taking into account the optimal performance without considering the constraint of the preparation data set. That is to determine universal threshold value dynamically for concept metadata relatedness.

II. Related Work

A lot of work has been done by various researchers on semantic focused crawler. Here, we give an overview about the domain of semantic focused crawling and ontology-learning-based focused crawling, and review related work done on ontology learning-based focused crawling.

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 [11], [12].

The main purpose of semantic focused crawlers is to precisely and efficiently extract and download relevant Web information by automatically understanding the semantics underlying the Web information and the semantics underlying the previous fields. Taking into consideration the previously present service information becomes a vital issue in Digital Ecosystems. In order to solve this problem, a conceptual framework for a semantic focused crawler, with the purpose of automatically discovering, annotating and classifying the service information with the Semantic Web technologies was presented [11].

Ontology-based focused crawlers refer to a group of focused crawlers that link web documents with related ontology concepts, with the purpose of filtering and categorizing web documents [18]. Ontology is a well-formed knowledge representation, ontology-based focused crawling approach. It uses predefined concept weights for the calculation the relevant scores of web pages. But, during the crawling process it is not easy to get the optimal concept weights in order to maintain a stable harvest rate. To address this issue, we proposed a learnable focused crawling framework based on ontology. An ANN (artificial neural network) was constructed using a domain-specific ontology and applied to the classification of web pages [19].

C Su et al. presented an intelligent focused crawler algorithm in which we embed ontology to evaluate the page's relevance to the topic [20]. Such a capability would be of keen interest in focused crawling and resource discovery, as it can fine-tune the relevance of unvisited URLs in the crawl frontier. Given specific domain ontology and a topic represented by a concept in this ontology, the relevance score 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 [20].

The limitations of Su *et al.*'s approach are: 1) it cannot be used to enrich the vocabulary of ontologies; 2) although the unsupervised learning paradigm can work in an uncontrolled Web environment, which may not work well when numerous new terms emerge or when ontologies have a limited range of vocabulary [20].

System uses ontology & query data for information extraction. The query executed by crawler and the data is extracted from web or internet. The service users may come across three major issues – heterogeneity, ubiquity, and ambiguity [1], when searching for mining service information over the Internet. So, a framework of a novel self-adaptive semantic focused crawler with the purpose of precisely and efficiently discovering, formatting, and indexing service information over the Internet, by taking into account the three major issues [1].

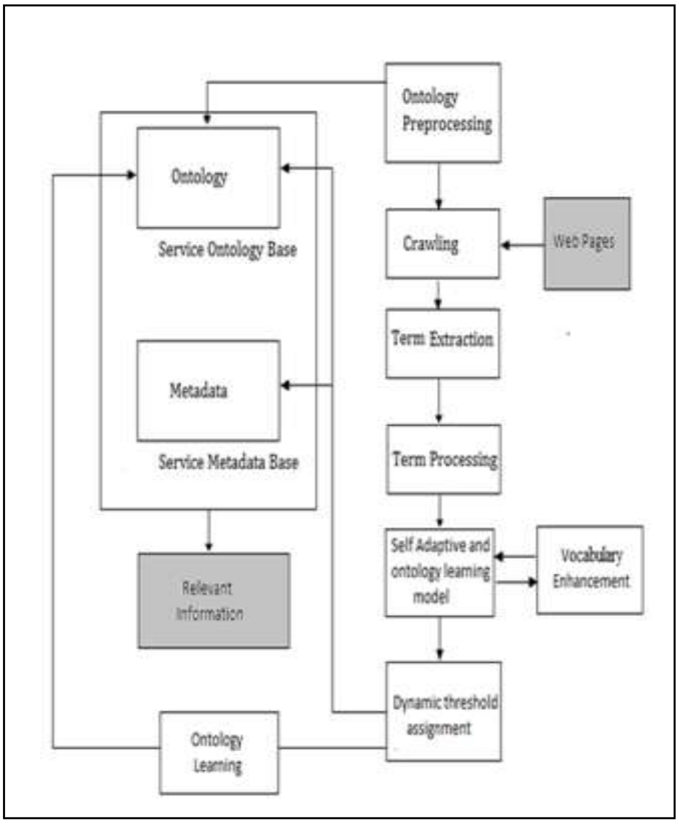

Figure 1. General System Architecture

III. Objectives

Our work is related to designing framework for self adaptive semantic focused crawler based on ontology learning. Some of the major objectives are:

A. Effectively and accurate information discovery over the internet.

B. Determine universal threshold value dynamically for concept metadata relatedness.

C. To enrich the vocabulary of the service ontology by surveying those unmatched but relevant service descriptions, in order to further improve the performance of the crawler

D. To enable the crawler to work in an uncontrolled web.

IV. System Implementation

An overview of the system architecture and the workflow is depicted in Fig. 1. The architecture includes the following steps:

- First step is pre-processing, which is to process the contents of the concept description property of each concept in the ontology before matching the metadata and the concepts.

- Crawling and term extraction are the next steps.
- In the term processing, which is to process the content of the service Description property of the metadata.
- The remaining workflow can be incorporated as a self-adaptive metadata association and ontology learning process.
- Next step is **v**ocabulary Enhancement provide a huge amount of additive data with deals to enrich the vocabulary of the mining service ontology.
- Determine universal Threshold value.

Workflow:

A. Preprocessing

Initially in preprocessing, before matching the metadata and the concepts it processes the contents of the concept Description property of every concept in the ontology. This processing is attained by the usage of WordNet Library in order to implement tokenization, part-of speech (POS) tagging, nonsense word filtering, stemming, and similar word searching for the conceptDescription property values of the concepts.

B. Term Extraction

The second step is crawling whereas the third step is term extraction. The vital moto of these two processes is to download Web pages from the Internet at same time, and to get the relevant data from the web pages which has been downloaded, according to the mining administration metadata planand the mining service provider metadata schema, for the purpose of preparing the property values to generate a new group of metadata.

C. Term Processing

Further term processing is carried out, which processes the content of the serviceDescription property of the metadata to construct for the purpose of subsequent concept-metadata matching. The implementation of this process and the implementation of the preprocessing process are likely to be similar.

D. Self-adaptive metadata Association and ontology learning process

The rest of the workflow can be combined as a self-adaptive metadata association and ontology learning technique. Here, first of all it is detected whether or not the contents of the serviceDescription property of metadata are involved in the conceptDescription with the use of direct string matching process and learned ConceptDescription properties of a concept.

E. Vocabulary enhancement

Here, a vast amount of additive data contained in the Vocabulary Enhancement leads in enriching the vocabulary of the mining service ontology by making a survey of unmatched but related service descriptions which can be useful to substantially improve the performance of the crawler.

F. Dynamic Threshold Assignment

This is a threshold value set or derived for relevant accuracy of report by crawler. Here, a universal threshold value has been found for the concept-metadata semantic similarity algorithm in order to set up a boundary for determining concept-metadata relatedness.

V. Ontology Steps

Algorithm is as follows:

A. Input the URL as query.
B. Use web pages of the defined type (html, php, jsp etc.) then generate queue and it is added to queue.
C. Process the page content with parser.
D. As per the response from the server if it is ok then read the file of ontology and matches the content of web page with the terms of ontology.
E. Relevance Score will be getting from corresponding web page and add the web page to index and caches file to a folder. Use cache and index for efficient searching.

Relevance Score can be calculated by following algorithm. Let P is Webpage and RELEVANCE P = 0 (Relevance Score). LIMIT will be set by us for checking relevancy of a Webpage and is a numerical value. Different results are obtained by fetching the same Website with different limits.

```
1. Read first term (T) from the ontology and give it the weight (W) according to the weight Table which contains LEVEL, ONTOLOGY TERMS and WEIGHTS.
2. Calculate term (T) and its synonyms occurrences in the Webpage P. Let FREQUENCY is the number of Occurrence.
3. Multiply the number of occurrence calculated at step 2 with the weight W. Let call this SCORE. Then SCORE = FREQUENCY * W.
4. Now RELEVANCE P = RELEVANCE P + SCORE.
5. Select the next term and weight from the weight table and go to step 2, Visit all the terms in the weight table.
6. Check If RELEVANCE P ¡ LIMIT then the Webpage is discarded
   Else
   The page is Use to Download. End
```

Provided for some include, discover all website pages from the web with sites. Unmistakably, this issue is harder than the past issue since in the event that we can take care of this issue, we can tackle the past issue like precision, heterogeneity.

VI. Experimental Results

A. *Crawling Time*

Crawling time is one of the factor which is used to measure the efficiency of a crawler. Table I shows the crawling time of various data sets on traditional system and proposed system. To measure system crawling time various queries are used and their respective crawling time recorded. The time is measured in seconds.

Here, we compare Traditional system versus proposed system. Figure 2 shows a graph for crawling time. A graph is plotted with X-axis against Y-axis. X-axis represents the query whereas Y-axis represents time in milliseconds.

TABLE I. Crawling Time

Data set	Traditional System	Proposed system
Apple store	3845	3045
bankaccount	2000	1900
Boxing day	2300	1930
Computer Mouse	2675	2285
desert	2814	2401
Java	1022	900

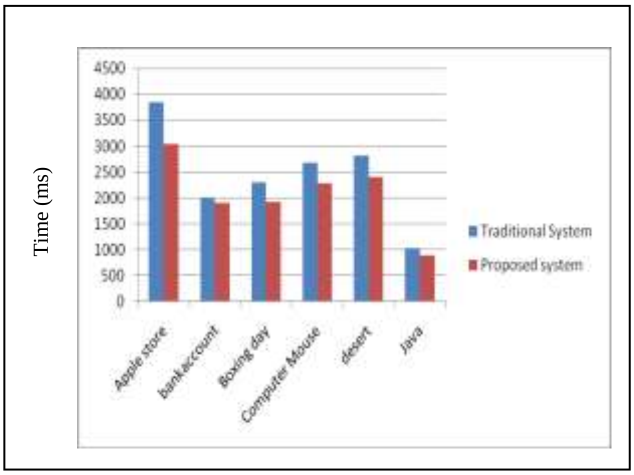

Figure 2. Crawling Time

B. *Exact Time Accuracy*

Table II shows the execution time required for various queries. Execution time is measured in milliseconds. The execution time as well as the exact time accuracy for traditional system and the propose system is shown.

Figure 3 shows a graph for exact time accuracy. A graph is plotted with X-axis against Y-axis. X-axis represents the query whereas Y-axis represents time in milliseconds.

TABLE II Execution time

Data set	Traditional System		Proposed system	
	Execution Time(ms)	*Accuracy*	*Execution Time(ms)*	*Accuracy*
Apple store	5600	69	4703	80
bankaccount	5437	84	8765	96
Boxing day	1543	87	1543	91
Computer Mouse	2387	80	2387	84
desert	4975	69	3294	87
Java	6432	74	7000	80
nurse	5002	79	4740	82
nursejob	8234	84	7543	84
storm	4863	85	4873	87
toothpaste	6363	79	6196	83

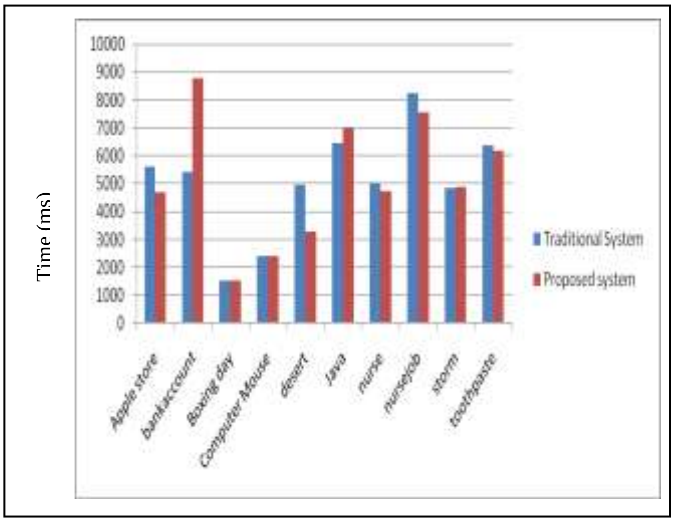

Figure 3. Exact Time Accuracy

VII. Conclusion

Our work focused on adequately and precise data revelation over the web. Additionally focus general edge esteem alterably for idea metadata relatedness furthermore, improve the vocabulary of the ontology base by looking over those unmatched yet, applicable administration portrayals, to further enhance the execution of the crawler and configuration system to empower the crawler to work in an uncontrolled web.

Acknowledgment

I thank my project guide Prof. S.M.Shinde and M.E. coordinator Prof. M.D.Ingle for the guidelines in completion of this paper. I also wish to record my thanks to our Head of Department Prof. H.A.Hingoliwala for consistent encouragement and ideas.

References

[1] Hai Dong, Member, IEEE, and Farookh Khadeer Hussain, Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery , Vol. 10, MAY 2014

[2] R. C. Judd, "The case for redefining services," *J. Marketing*, vol. 28, pp. 58–59, 1964.

[3] T. P. Hill, “On goods and services,” *Rev. Income Wealth*, vol. 23, pp. 315–38, 1977.

[4] C. H. Lovelock, “Classifying services to gain strategic marketing insights,”*J. Marketing*, vol. 47, pp. 9–20, 1983.

[5] H. Dong, F. K. Hussain, and E. Chang, “A service search engine for the industrial digital ecosystems,” *IEEE Trans. Ind. Electron.*, vol. 58, no. 6, pp. 2183–2196, Jun. 2011.

[6] Mining Services in the US: Market Research Report IBISWorld2011.

[7] B. Fabian, T. Ermakova, and C. Muller, “SHARDIS – A privacy-enhanced discovery service for RFID-based product information,” *IEEE Trans. Ind. Informat.*, to be published.

[8] H. L. Goh, K. K. Tan, S. Huang, and C. W. d. Silva, “Development of Bluewave: A wireless protocol for industrial automation,” *IEEE Trans. Ind. Informat.*, vol. 2, no. 4, pp. 221–230, Nov. 2006.

[9] M. Ruta, F. Scioscia, E. D. Sciascio, and G. Loseto, “Semantic-based enhancement of ISO/IEC 14543–3 EIB/KNX standard for building automation,” *IEEE Trans. Ind. Informat.*, vol. 7, no. 4, pp. 731–739, Nov. 2011.

[10] I. M. Delamer and J. L. M. Lastra, “Service-oriented architecture for distributed publish/subscribe middleware in electronics production,” *IEEE Trans. Ind. Informat.*, vol. 2, no. 4, pp. 281–294, Nov. 2006.

[11] 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. 2106–2116, Jun. 2011.

[12] 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. 687–704, 2011.

[13] 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. 1–11, Feb. 2006.

[14] S. Runde and A. Fay, “Software support for building automation requirements engineering—An application of semanticweb technologies in automation,” *IEEE Trans. Ind. Informat.*, vol. 7, no. 4, pp. 723–730, Nov. 2011.

[15] M. Ruta, F. Scioscia, E. Di Sciascio, and G. Loseto, “Semantic-based enhancement of ISO/IEC 14543–3 EIB/KNX standard for building automation,” *IEEE Trans. Ind. Informat.*, vol. 7, no. 4, pp. 731–739, Nov. 2011.

[16] 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.

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

[18] W. Wong, W. Liu, and M. Bennamoun, “Ontology learning from text: A look back and into the future,” *ACM Comput. Surveys*, vol. 44, pp. 20:1–36, 2012.

[19] 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.

[20] 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)*, Rio de Janeiro, Brazil, 2005, pp. 73–78.

[21] 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.

.