Review on Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery

Miss. Shwetha Jog
Research Scholar,
DPCOE,Pune, India,

Prof. Shubham Joshi,
Research Supervisor,
DPCOE, Pune, India,

***Abstract*— Web Crawlers are one of the most critical components used by the Search Engines to collect pages from the Web. It is an intelligent technique of browsing used by the Search Engine. The requirement of a web crawler that downloads most relevant web pages from 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. Focused crawler is used to collect those web pages that are relevant to a particular topic while filtering out the irrelevant pages.**

The Ontology based web crawler algorithm used to mining user queries by using separate service. The use of algorithm like page rank and other importance -metrics has scheduled a new approach in prioritizing the URL queue for downloading higher relevant pages.The idea of this paper is based on the design of an unsupervised framework for vocabulary -based ontology learning, and also a hybrid algorithm is used for matching semantically relevant concepts and metadata.

Keywords—Information discovery, Mining service, Ontology learning,SASF crawler.

INTRODUCTION

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, while keeping the amount of not relevant pages downloaded to a minimum number.

Crawlers are given a starting set of web pages (seed pages) as their input, extract outgoing links appearing in the seed pages and determine what links to visit next based on certain criteria. Web pages pointed to by these links are downloaded, and those satisfying certain relevance criteria are stored in a local repository. Crawlers continue visiting Web pages until a desired number of pages have been downloaded or until local resources (such as storage) are exhausted. 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.

The vast amount of information in the Web represents one of the most striking challenges for computer science in recent decades. On the one hand, the emergence of the Web has offered the opportunity to access more information than ever before. On the other hand, the wide range of topics covered and the diverse quality of these resources has pushed to the limit the development of new search technologies. A significant number of people use Web search engines to formulate queries (a set of terms) and review a list of suggested answers. Search engines are built from practical implementations of information retrieval techniques devised to handle large-scale Web collections. An increasing interest in the use of new specialized search engines has focused many efforts in the development of vertical search technologies.

Vertical search engines are devised to serve on specific topics of information, providing higher accuracy than general purpose (horizontal) search engines and reducing computational costs involved in query processing. Vertical search engines employ focused crawlers to collect pages relevant to specific topics. A focused crawler attempts to download (and there by consume bandwidth for) only Web pages which appear to be relevant to a given topic. In practice, this task is very difficult. At running time, the crawler must decide whether a Web page is relevant before downloading it by just examining the hyperlink that points to its actual content. Once a page is fetched, the crawler uses the content of the page or a more specific portion of its text to decide whether to follow an outgoing link. This process continuously repeated by following the most promissory links, fetching each promising Web page, and so on. In practice, the Web is traversed by conducting a best first crawling process.

I. PROBLEM DEFINITION

To design and implement an a semi-supervised approach by aggregating the unsupervised approach and the supervised ontology learning-based approach, with the purpose of automatically choosing the optimal threshold values for each concept, while keeping the optimal performance without considering the limitation of the training data set.

II. PERSPECTIVE SOLUTION

Hidden Markov Model for fetching relevance and non-relevance web pages is used. And priority crawling scores to Rank and order the relevant URLs for getting the highest priority and lowest priority is used. And finding the Similarity value based on priority values, and getting the optimal threshold value and ontology data.

III. RELATED WORK AND 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 [3]–[4]. 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 predefined topics. A survey conducted by Dong *et al.* [5] found that 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. Furthermore, the quality of ontologies may be affected by two issues. The first issue is that, as it is well known that ontology is the formal representation of specific domain knowledge [6] and ontologies are designed by domain experts, a discrepancy may exist between the domain experts' under-standing of the domain knowledge and the domain knowledge that exists in the real world. The second issue is that knowledge is dynamic and is constantly evolving, compared with relatively static ontologies. These two contradictory situations could lead to the problem that ontologies sometimes cannot precisely rep-resent real-world knowledge, considering the issues of differentiation and dynamism. The reflection of this problem in the field of semantic focused crawling is that the ontologies used by se-mantic focused crawlers cannot precisely represent the knowledge revealed in Web information, since Web information is mostly created or updated by human users with different knowledge understandings, and human users are efficient learners of new knowledge. The eventual consequence of this problem 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 semantic - focused crawlers, researchers have begun to pay attention to enhancing se-mantic-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 [7]. 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.

IV. PROPOSED WORK

To design a semi-supervised approach by aggregating the unsupervised approach and the supervised ontology learning-based approach, with the purpose of automatically choosing the optimal threshold values for each concept, while keeping the optimal performance without considering the limitation of the training data set. Hidden Markov Model is proposed to determine the relevance and non-relevance of a retrieved web page.

And priority crawling score is used to Rank and order the relevant URLs for getting the highest priority and lowest priority and finding the Similarity value based on priority values. Finally getting the optimal threshold value and ontology data. In order to evaluate the retrieval effectiveness, precision Pr corresponds to the fraction of top r ranked web pages that are relevant to the query over the total number of retrieved web pages. The time spent for crawling and analyzing the semantic web pages is another important element to measure.

V. PROPOSED ARCHITECTURE/PROTOTYPE

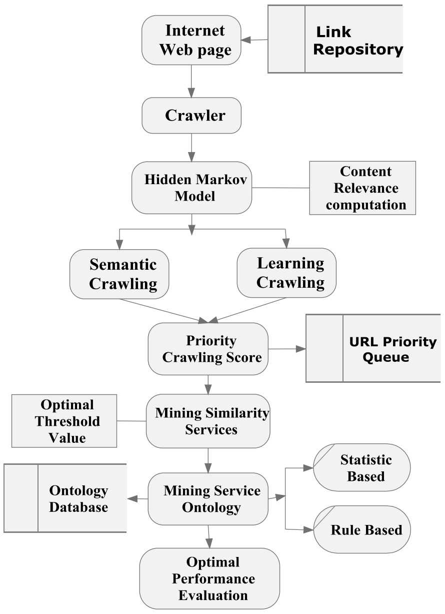

Fig1. Proposed system architecture.

VI. SCOPE OF WORK

Need to determine the relevance and no relevance of a retrieved web page. For this purpose proposed the Hidden Markov Model for fetching relevance and non-relevance web pages. And used priority crawling score to Rank and order the relevant URLs for getting the highest priority and lowest priority. And finding the Similarity value based on priority values. Finally getting the optimal threshold value and ontology data.

VII. DISCUSSION

Focused crawler is used to collect those web pages that are relevant to a particular topic while filtering out the irrelevant. Proposed system is intend to design a semi-supervised approach by aggregating the unsupervised approach and the supervised ontology learning-based approach, with the purpose of automatically choosing the optimal threshold values for each concept, while keeping the optimal performance without considering the limitation of the training data set. In the proposed system, we need to determine the relevance and non-relevance of a retrieved web page. For this purpose we propose the Hidden Markov Model for fetching relevance and non-relevance we pages by implementing. And we use priority crawling score to Rank and order the relevant URLs for getting the highest priority and lowest priority to get the optimal threshold value.

VIII. REFERENCES

[1] H. Dong and F. K. Hussain, “Focused crawling for automatic service discovery, annotation, and classification in industrial digital ecosys-tems,” *IEEE Trans. Ind. Electron.*, vol. 58, no. 6, pp. 2106–2116, 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. 687–704, 2011.

[3] 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.

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

[5] 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.

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

[7] 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.

[8] *Hai Dong, Member, IEEE, and FarookhKhadeer Hussain,“* Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery”, IEEE Trasncations On Industrial Informatics, VOL. 10, no. 2, may 2014

IX. AUTHOR BIBLIOGRAPHY

	Miss. Shwetha Jog She has completed Bachelor of Engineering in Computer Science and Engineering from VTU University, Karnataka. Currently pursuing Master of Engineering from SavitribaiPhule Pune University.
	Prof. Shubham Joshi He has completed Bachelor of Engineering in Computer Engineering and Master of Engineering in Information Security. Currently He is pursuing Ph.D. He has published 22 research papers, 01 Book. He is Microsoft certified professionalandChairperso nof Publicity & Web Hosting, IEEE MPSubsection, Reviewer of the International conferences and Journals.Woking as a PG Co-ordinator at DPCOE,Pune.