

Self-Adaptive Ontology Technique based on Crawler History

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

Abstract— Now a day's internet became very necessary in day to day life. Web crawlers play a role of critical component which is used by Search Engines to collect pages from WWW. Web crawler downloads most relevant pages from WWW. This task is important challenge. Search Engine uses this intelligent system. There are different techniques available for retrieving most important and relevant information from web. Keyword search is most used technique. By use of this technique, crawlers retrieve irrelevant pages also along with relevant pages. But Focused crawler is used to collect relevant pages of a certain topic.

Self-Adaptive Ontology Based on Crawler History is retrieves the pages by searching logically related keywords instead of using keyword search method. And parallel ontologies of a given keyword get updated automatically by learning new ontologies from previous crawled history.

Keywords—*Ontology learning, SASF crawler.*

INTRODUCTION

Crawlers are main components used by search engines. They are intelligent programs which collects information locally which is available around WWW. But as crawlers crawls all available information from the internet, some information will be irrelevant. So it leads to high usage of costly resources. So focused crawler have been introduced to satisfy individuals like student or domain experts or organization or researcher to create and maintain matter specific information. Focused crawler downloads as much as relevant pages and in parallel low irrelevant pages. Crawlers should be most intelligent to crawl most up to date information.

Crawlers take a starting set of web page urls known as seed set as input, extract outgoing links which are available from seed url pages and depends on some criteria determines whether to go for that new link or not. We pages pointed by these links are also downloaded and process is continued until number of downloaded documents reaches threshold or the local resources exhausted. General purpose crawlers retrieve a large number of web pages regardless of their topic. Focused crawler retrieve the web pages which are seems to be relevant by examining the information available around the link and structure of the link.

The main challenge is available information around the web is vast. A significant number of people make use of search engines to search for interested topic and retrieve and review a list of answers. Search engines built by using

retrieval techniques which are capable of handling large scale web collections.

I. OBJECTIVES OF THE RESEARCH

- To propose a self-learning ontology technique.
- To propose a self-learning most common words (Stop words) technique.
- To provide a better related pages and ontologies by calculating particular threshold values dynamically for each concept.

II. RELATED WORK AND LITERATURE SURVEY

A semantic focused crawler is an intelligent technique which retrieves and downloads relevant information on specific topics or keyword or search query by using logically relating to keywords and semantic technologies [1], [2]. Semantic technology helps in improving the knowledge or logically related keywords [3]–[4]. Main goal of semantic focused crawlers is to retrieve and download relevant documents or information efficiently and effectively by automatically understanding the logically related information by predefined information. A survey conducted by Dong et al. [5] found that most of the crawlers use ontologies to represent the knowledge. But the limitation of the ontology-based semantic focused crawlers is that the crawling performance depends on the quality of ontologies. Quality of the ontologies plays a very important role. It gets affected by 2 issues. They are,

- Ontology is the formal representation of specific domain knowledge [6] and ontologies are updated by domain experts, a discrepancy may exist between the domain experts' understanding of the domain knowledge and the domain knowledge that exists in the real world.
- Knowledge is dynamic and is constantly evolving, compared with relatively static ontologies.

Because of these 2 situations, ontologies fail in representing real world knowledge. And it directly creates a problem in crawlers where semantic crawlers which use these ontologies cannot effectively represent knowledge. Because web information is created and updated by manually and there are lots of chances getting different. This leads decreases in performance and efficiency of semantic focused crawler. In order to solve this problem, ontologies should get updated automatically instead of updating manually.

III. PROPOSED ARCHITECTURE/PROTOTYPE

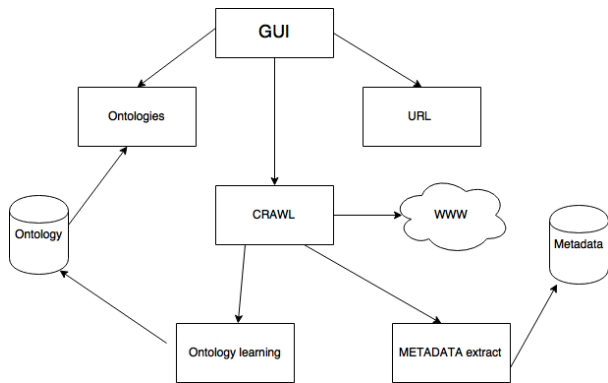


Fig1. Proposed system architecture.

Proposed system aim is to extend existing work and implement a self-learning ontology to keep the ontology up to date with existing knowledge in world. This system is able to increase the number of crawled related documents and also calculating optimal threshold value dynamically for learning logical ontologies. By use of this system user will get a high number of related documents crawled. Because of using self-Adapted ontologies from previous crawled history instead of using keyword search technique. Keyword search technique crawls for a specific given keyword whereas this technique able to crawl semantically related ontologies which are self-adapted from previous crawl history.

IV. MATHEMATICAL MODEL

Let S is represented as System, $S = \{K,U,D,W,CA,O\}$,Such that:

K= Input keyword to be crawled
 Ok= Ontologies for the Keyword K
 U= URL list to indicating source to crawl
 For each D1 to DM find W11 W1M
 Generate C1 = Count(W11)
 Output C1
 for i from 2 to M do
 Generate Ci = Count(W1i)
 Output Ci
 end for

Output: D1 to DM indicating crawled related documents O1 to OM learned ontologies for given keyword K

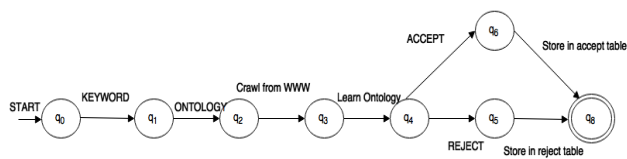


Fig2. Mathematical model

V. RESULTS

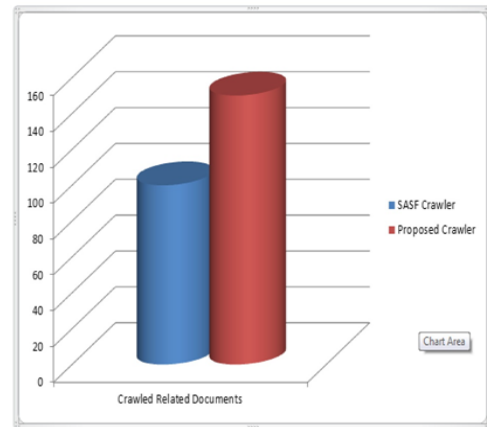


Fig3. Number of crawled documents

VI. CONCLUSION


The main objective of this research was to throw some light on web crawling algorithm. Web crawler is most essential component of the search engine. And search engines are critical part of our day to day life. Web crawler crawls all the available documents from www. But if someone interested in only some particular domain or some area it is waste of time and costly resources like memory, CPU time and network. And it uses a keyword search based approach.

Enhanced Self Adaptive Ontology based Focused Crawler, learns new keywords or ontologies by extracting the unmatched but related keywords. Hence the number of semantically related documents crawled by the crawler is high as compared to other crawlers.

VII. REFERENCES

- [1] Hai Dong, Member, IEEE, and FarookhKhadeer Hussain,Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery. VOL. 10, no. 2, may 2014
- [2] 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, Jun. 2011
- [3] 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, 2011.
- [4] Razali, A.M. and S. Ali, Semantic web services in factory automation: Fundamental insights and research roadmap. IEEE Trans.Ind. Informat., vol. 2, no. 1, Feb. 2006.
- [5] Saad Ali, S.N., Semantic-based enhancement of ISO/IEC 14543?3 EIB/KNX standard for building au-tomation. IEEE Trans. Ind. Informat., vol. 7, no. 4, Nov. 2011.
- [6] H. Dong, F. Hussain, and E. Chang, O. Gervasi, D. Taniar, B. Mur-gante, 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.
- [7] Srinivas, K., B.K. Rani, and A. Govrdhan, A translation approach to portable ontology specifica-tions. Knowledge Acquisition, vol. 5, 1993.
- [8] Das, R., I. Turkoglu, and A. Sengur, Ontology learning from text: A look back and into the future. ACM Comput. Surveys, vol. 44, 2012.

VIII. AUTHOR BIBLIOGRAPHY

	<p>Miss. Shwetha Jog She has completed Bachelor of Engineering in Computer Science and Engineering from VTU University, Karnataka. Currently pursuing Master of Engineering from Savitribai Phule Pune University.</p>
--	--