Ontology Based Web Crawler

Swati Ringe[#], Nevin Francis, Palanawala Altaf H.S.A.

[#]*Swati Ringe, Fr.Conceicao Rodrigues College Of Engineering, Fr.Agnel Ashram, BandStand, Bandra-w, Mumbai-400050*

[1]swatiringe@yahoo.com
[3]altaf110@rediffmail.com
[2]jnevin777@gmail.com

Abstract— **The Web, the largest unstructured database of the world has greatly improved access to the documents. As the number of Internet users and the number of accessible web pages grow, it is becoming increasingly difficult for users to find documents that are relevant to their particular needs. Users must either browse through a hierarchy of concepts to find the information they need or submit a query to a Search Engine and wade through hundreds of results most of them irrelevant.**

Web Crawlers are one of the most crucial components used by the Search Engines to collect pages from the Web. It is an intelligent means of browsing used by the Search Engine. 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.

We propose a novel method of addressing this issue without compromising with the relevancy of the retrieved documents through the crawler. The proposed technique makes use of Semantics which helps to downloaded only relevant pages. Semantics can be provided by ontologies. An Ontology Based Web Crawler uses ontological Engineering concepts for improving its crawling performance. The Crawler, guided by an ontology describing the domain of interest, crawls the Web focusing on pages relevant to a given topic ontology. As a result ontologies found during the crawl will be relevant to the domain and produce a set of candidate mappings with the topic ontology. The crawler deals with prioritizing the URL Queue for crawling more relevant pages based on domain dependent ontologies. Also explores the possibility of using the semantic nature of the URL which has been obtained from the ontology tree to filter out the URL's Queue.

The main advantage of Ontology Based Web Crawler over other Focused crawler is that it does not need any Relevance Feedback or Training procedure in order to act intelligently. Moreover the number of extracted documents will reduce as well as the crawling time thus leading to greater search efficiency.

Keywords— **Crawler, Ontology, Search Engine, Semantic Web.**

I. INTRODUCTION

A Web Crawler is a relatively simple, automated program, or script that methodically scans or "crawls" through internet pages to create an index of the data it is looking for.

Usually for indexing, crawler-based engines consider much more factors than those they can find on the web pages. Thus before putting any web page into an index, a crawler will look how many other pages in the index are linking to the current web page, the text used in the links the user points to, what the page rank of the linking pages is, whether the page is present in some directories under related categories, etc. These "*off-the-page*" factors play a weighty part when the page is evaluated by a crawler-based engine. While theoretically, the web page developer can artificially increase the page relevance for certain keywords by adjusting the corresponding areas of the HTML code, user still have much less control over other pages in the internet that are linking to the user.

Thus off-the-page relevance prevails in the crawler's eye leading to the following **problems** faced by the general web crawler:

1. Web is increasing in size day by day. It is observed that 600 GB of text changes every month. Since web crawler fetches each and every page, it requires large storage area and it consumes much time also.
2. Hardware requirement for the crawler is too high (CPU, disks etc).
3. Crawlers cover only 30-40% of web.
4. The Search Engine which uses this general web Crawler returns links in which most of the times the first few links may not be relevant to the topic.

Crawlers have been around for a long time and have proven their usefulness and success on the Web. None the less, these general-purpose crawlers are not sufficient to tackle the stated problem. They crawl the Web in a blind and exhaustive manner. Since our goal is to find very specific data on the Web, this exhaustive approach will not find the requested information considering the current size of the Web.

Therefore, in this paper we propose a focused crawling process based on domain specification so that the crawler is guided to the relevant information and no

time is wasted on irrelevant resources. These kinds of ontology based crawlers are also referred to as preferential or heuristic-based crawlers. The heuristic we use in our nominated solution is ontology matching. Since the goal is to find information resources on the Web, we expect that most of these resources are semantically annotated and linked using an ontology hierarchy. The algorithm that we use to develop an Ontology Based Web Crawler solves the major problem of finding the relevancy of pages before the process of crawling, to an optimal level. It presents an intelligent focused crawling algorithm in which ontology is embedded to evaluate the page's relevance to the topic with a relevancy limit. Raman et al. [1] and Chang et al. [6] have also presented intelligent crawler algorithms.

II. ONTOLOGY

Ontology is a formal, explicit specification of shared conceptualization. Ontology provides a common vocabulary of an area and defines, with different level of formality, the meaning of terms and relationships between them. Ontologies were developed in Artificial Intelligence to facilitate Knowledge sharing and reuse. Since the early 1990"s, ontologies have become a popular research topic. They have been studied by several Artificial Intelligence research Communities, including Knowledge engineering, natural-language processing and Knowledge Representation. It helps in describes a Semantic Web- based Knowledge management architecture and a suite of innovative tools for semantic information Processing [2].

III. THE PROPOSED ALGORITHM OVERVIEW

In Ontology Based Web Crawler the web pages are first checked for validity (i.e. of the type html, php, jsp etc). If it is valid then it is parsed and the parsed content is matched with the ontology. If the page is relevant it is indexed otherwise it is not considered.

Hence the algorithm is as follows:

1. Get the seed URL.
2. If the web page is valid that is it is of the defined type (html, php, jsp etc.) then it is added to queue.
3. Parse the content.
4. Get the response from the server if it is ok then read the protégé file of ontology. and match the content of web page with the terms of ontology.
5. Count the Relevance Score of web page and add the web page to index and caches file to a folder. With the help of cache and index searching can be done.

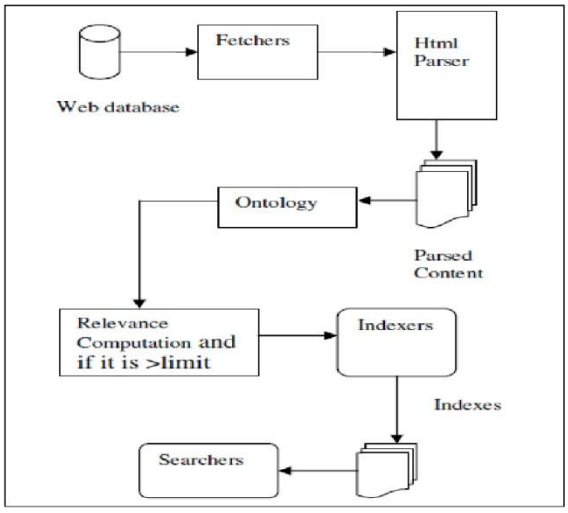

Figure 1: Architecture of Ontology Based WebCrawler.

For calculating the Relevance Score following algorithm is used.

A. Algorithm for Calculating the Relevance Score

We take reference from the work done by Ganesh et al. [8] and Debajyoti et al [7]. Let P is Webpage and RELEVANCE_P = 0 (Relevance Score). LIMIT is a numerical value that will be set by us for checking relevancy of a Webpage. Different results are obtained by crawling the same Website against 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 how many times the term (T) and its synonyms occur in the Webpage P. Let the number of occurrence is calculated in FREQUENCY.
3. Multiply the number of occurrence calculated at step 2 with the weight W. Let call this SCORE. Then SCORE = FREQUENCY * W.
4. Add this term weight to RELEVANCE_P. So new RELEVANCE_P = RELEVANCE_P + SCORE.
5. Select the next term and weight from the weight table and go to step 2, until all the terms in the weight table are visited.
6. If RELEVANCE_P < LIMIT then
   The Webpage is discarded
   Else
   The page is downloaded.
7. End.

IV. RELATED WORK

We present a case study of how the suggested crawler computes the relevancy of the Web page given in reference [9] which has the file named in reference [5] for the search Keyword "Hidden Web" using the Ontology (shown below) for Hidden Web stored in our Knowledge Base. We use an open source platform Protégé. Hidden Web is the root class in our ontology. It has subclasses Hidden, Web, Crawler, Architecture. The Crawler class is further subdivided into Index, URLs and Search. Contents has subclasses Domain, Weights and Labels.

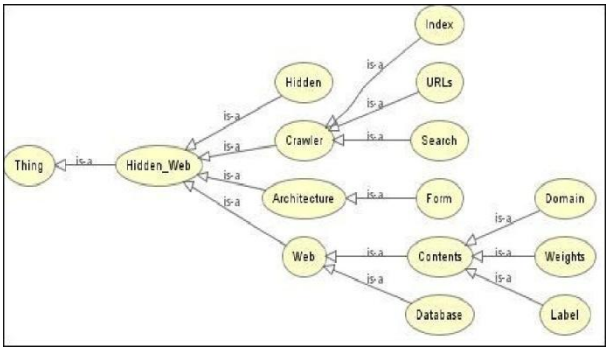

Figure 2: Graphical Representation of the Ontology for Hidden Web

The entire computation of the algorithm for the relevancy computation is shown in the tabular form below.

TABLE 1: ONTOLOGY WEIGHT TABLE FOR HIDDEN WEB

Level	Ontology Terms	Weight (W)	Frequency (F)	Score W*F
1	HiddenWeb	0.1	13	1.3
2	Hidden	0.4	19	7.6
2	Crawler	0.4	21	8.4
2	Architecture	0.4	03	1.2
2	Web	0.4	26	10.4
3	Index	0.6	04	2.4
3	URLs	0.6	03	1.8
3	Search	0.6	11	6.6
3	Form	0.6	30	18.0
3	Contents	0.6	09	5.4
3	Database	0.6	02	1.2
4	Domains	1.0	04	4.0
4	Weights	1.0	04	4.0
4	Label	1.0	11	11.0

Here, RELEVANCE_P = $\sum$ score = 83.3

We take our LIMIT to be 10 for more accurate results. In this case RELEVANCE_P > LIMIT. (Actually RELEVANCE_P is more than 8 times the LIMIT value**).** Hence the Crawler concludes that the page is very much relevant to the search query and starts downloading the page.

V. FUTURE SCOPE

Though we believe that our projected crawler takes care of everything an efficient crawler needs, there is still a window of improvement in our crawler that can be addressed. In our relevancy calculation algorithm, we have to set the weight of the ontology term manually. A mechanism can be devised such that after reading the ontology and after visiting certain Web pages it can provide the weight of the ontology term automatically. Also, the processing time of the Web crawler can be improved. In our algorithm the ontology remains static; ontology can be evolved dynamically by adding new concepts and relations while visiting Web pages.

VI. CONCLUSION

The main aim of our paper is to retrieve relevant Web pages and discards the irrelevant ones. We have developed an ontology based crawler which retrieves Web pages according to a relevancy calculation algorithm and discards the irrelevant Web pages. In doing this we have use the concept of Ontology which provides the meaning of terms and relationship between them. We believe that our aimed crawler will not only be helpful in exploiting fewer web pages such that only relevant pages are retrieved but also will be an important component for the future „Semantic Web" which is going to become very popular in the years to come. Hence, such an improved crawler suggested by us in this paper can help in applications areas like Social Networking Portal, Online Library for Books Information etc. and can add to the benefits of them in their respective fields.

REFERENCES

[1] Raman Kumar Goyal[1], Vikas Gupta[2], Vipul Sharma[3], Pardeep Mittal[4], "*Ontology Based Web Retrieval*", [1]Lecturer (Information Technology), RIEIT, Railmajra, [2]AP (CSE), RIEIT, Railmajra,[3]Student, UIET, Panjab University, Chandigarh, [4]AP (CSE), BFCET, Bathinda.

[2] Felix Van de Maele, "*Ontology-Based Crawler for the Semantic Web*", Faculty of Science, Department of Applied Computer Science, Vrije Universiteit Brussel, May 2006.

[3] Marc Ehrig, Alexander Maedche, "*Ontology Focused Crawling of Web Documents*", University of Karslruhe, Germany.

[4] Jan Paralic, Ivan Kostial, "*Ontology Based Information Retrieval*", Department of Cybernetics and AI, Technical University of Kosice, Letna 9, 040 11 Kosice, Slovakia.

[5] Sriram Raghavan,Hector GarciaMolina, "*Crawling the Hidden Web*", Computer Science Department, Stanford University, USA.

[6] Chang Su, Yang Gao, Jianmei Yang, Bin Luo "An Efficient Adaptive Focussed Crawler Based on Ontology Learning", Proceedings of the Fifth International Conference on Hybrid Intelligence Systems- 2005 IEEE.

[7] Debajyoti, Arup Biswas, Sukanta "A New Approach to Design Domain Specific Ontology Based Web Crawler", 10th InternationalConference on Information Technology – 2007 IEEE.

[8] Ganesh S, Jayaraj M, Aghila G "Ontology Based Web Crawler" Information Technology;Coding & Computing, 2004 volume 2, 2004 page (s) -337-341-IEEE. *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.23.2091&rep=rep1&type=pdf.*