A Novel Architecture of Ontology-based Semantic Web Crawler

Ram Kumar Rana
IIMT Institute of Engg. & Technology, Meerut, India

Nidhi Tyagi
Shobhit University, Meerut, India

ABSTRACT
Finding meaningful information among the billions of information resources on the Web is a difficult task due to growing popularity of the Internet. The future of World Wide Web (WWW) is the Semantic Web, where ontologies are used to assign (agreed) meaning to the content of the Web. On the Semantic Web, data will inevitably be linked to many different ontologies, and information processing across ontologies is not possible without knowing the semantic mappings between them. As the resources on the Semantic Web are annotated using these ontologies, new search techniques are required to find specific information. For this, architecture has been proposed for ontology based semantic web crawler. This architecture can exploit the semantic metadata to efficiently discover and extract information from the Semantic Web. In this paper Semantic matching between content of downloaded web page and ontology is used to guide the crawler towards relevant information.

General Terms:
Information Retrieval, Search Engine, Web Crawling and Semantic Matching.

Keywords: Ontology, Semantic Web, Crawler.

1. INTRODUCTION
The World Wide Web is an architectural framework for accessing linked documents spread out over millions of machines all over the Internet. The popularity of WWW is largely dependent on the search engines. Search engines are the gateways to the huge information repository at the internet. Search engine consist of four discrete components: Crawling, Indexing, Ranking and query-processing.

The earliest Web search engines relied for retrieving information from Web pages on the bases of matching with the words in the search query. As the Web continues to grow, and as the diversity of users increases search engines must utilize semantic clues to satisfy user's information needs. Semantic search, which takes into account the interests of the user as well as the specific context in which the search is issued, is the next step in providing users with the most relevant information possible.

Currently the general purpose search engines strive as entry points for the web pages perform the coverage of information that is as broad as possible. They use Web crawlers to maintain their index databases. These crawlers are blind and exhaustive in their approach, with comprehensiveness as their major goal. A URL (Uniform Resource Locator) that is a URI (Uniform Resource Identifier) specifies where an identified resource is available. In order to search most relevant information, crawlers can be more selective about the URL they fetch and refer as to be crawled this mechanism for retrieving such URL appears in [1] [2] [3] [4].

The Semantic Web is known for being a web of Semantic Web Documents (SWDs) those are freely available on the Semantic Web and are described in Resource Description Framework (RDF) or any other syntax of semantic web [5]. Today's search engines deal with SWDs poorly, since they have been developed to process text documents. Most make no attempt to parse web documents into appropriate tokens and none take advantage of the structural and semantic information encoded in a SWD. This paper proposes ontology based semantic web crawler architecture for effective crawling, parsing, analyzing and classification of semantic data.

1.1 The Semantic Web
First time the term "Semantic Web" came up was in 1998 when Tim Berners-Lee published the Roadmap to the Semantic Web on the homepage of the World Wide Web Consortium (W3C). Tim Berners-Lee defines the Semantic Web as follows: "The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"[6].

The Semantic Web is envisioned as a next generation of the Web in which information is machine readable, and automated agents can retrieve, extract, and combine information from the Web [7]. It is an evolving extension of the World Wide Web in which the semantics of information and services on the Web is defined, making it possible for the Web to understand and satisfy the requests of people and machines to use its content. One of the basic pillars of the Semantic Web concept is the idea of having explicit semantic information on the Web pages that can be used by intelligent agents in order to solve complex problems of Information Retrieval and Question Answering. In consequence, the final objective of the Semantic Web is to be able to semantically analyze and catalog the Web contents. This requires a set of structures to model the knowledge, and a linkage between the knowledge and contents. In this manner the Semantic Web relies on two basic components, ontologies and semantic annotations. It relies on ontologies in order to interpret the textual content of a resource regardless of its format. Even though there have been many conceptual approximations in the field of Semantic Web in which it is assumed that resources have been semantically annotated. So, in order to take profit from the Web resources which are currently available, the extraction of features from plain text, as proposed in this work, goes through the semantic analysis of its content and in association with the concepts of ontologies.

1.2 Ontology
The term Ontology was borrowed from philosophy and the term Ontology was initially used by AI practitioners and is now one of the fundaments of the Semantic Web. It is not possible to imagine the Semantic Web without ontology because fundamental concepts of semantic Web are ontologies or in other words it can said that Semantic Web is the biggest research project involving ontologies.

One can find many different definitions for the concept of ontology applied to information systems, each emphasizing a specific aspect its author judged as being more important.

For instance, Gruber (1993) [8] defines an ontology as a formal specification of a conceptualization or, in other words, a declarative representation of knowledge relevant to a particular domain. Uschold and Gruninger (1996) define ontology as a shared understanding of some domain of interest. Sowa (2000) defines ontology as a product of a study of things that exist or may exist in some domain. Ontology provides the "well-defined meaning" to the information contained in the Web having benefit that different parties over the internet now have "shared" definitions about certain key concepts.

With so many possibilities for defining what ontology is, one way of avoiding ambiguity is to focus on the objectives being sought when using it. For the purposes of the present research effort, the most important aspect of ontologies is their role as a structured form of knowledge representation. Ontologies are used for the purpose of interoperability among systems based on different schemas and comprehensively describing knowledge about a domain in a structured and sharable way; ideally in a format that can be read and processed by a computer. Semantic matching [9] based on Ontology can be used to design a crawler in order to give better result. So, meaningful information can be retrieved with such ontology based crawler because ontology is used for matching with semantics description of webpage.

1.3 Crawler

A crawler is a program that downloads and stores Web pages, often for a Web search engine. Roughly, a crawler starts off by placing an initial set of URLs, in a queue, where all URLs to be retrieved are kept and prioritized. From this queue, the crawler gets a URL (in some order), downloads the page, extracts any URLs in the downloaded page, and puts the new URLs in the queue. This process is repeated until the crawler decides to stop. Collected pages are later used for other applications, such as a Web search engine. Figure 1. represents the general architecture of the web crawler involving a scheduler and a multi-threaded downloader.

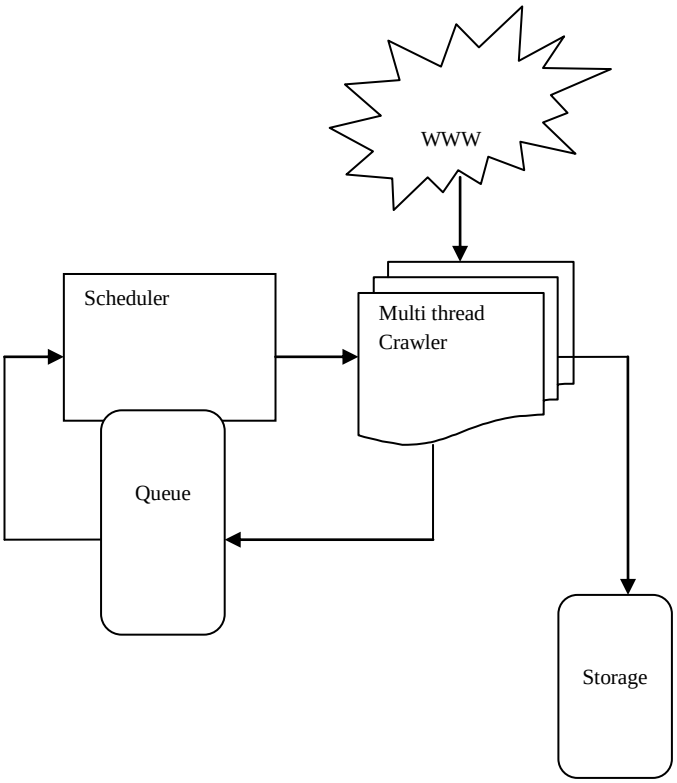

Figure.1 General Architecture of Web crawler, involving a scheduler and a multi-threaded downloader.

The two main data structures used in above architecture are the Web page storage and the URL Queue [10]. WebCrawler assists users in their Web navigation by automating the task of link traversal. Web crawler creates a collection of WebPages those are indexed and searched by search engine for fulfilling searchers' queries. A user issues a query to a pre-computed index and retrieves a list of documents that match the query. Thus web crawler is the most important part in of a search engine and plays a vital role in information retrieval.

2. RELATED WORK

The enormous success of Google and similar search engines for the Web has demonstrated the value of crawling and indexing its documents. In case of Semantic Web crawling and indexing weave to say that it was poorly handled by current search engines: for instance, query answering for keywords does not allow matching results based on semantics content. This creates a scope for semantic web crawler. Some efforts by different researchers in the same direction are given below.

Focused crawler [11] discover semantic web data, by using some sort of heuristic to rate pages according to their relevance to a given topic. This crawler should stay focused around the given topic, so that irrelevant pages should not pursued by the crawler.

Further, to find best result LS Crawler [12] performs the searching on the semantic basis. It enhances the process of determining the relevancy of the documents before downloading.

Most of the work to enhance page relevancy is done by improving page rank [13] by indexing module of search engine that keeps information about the ranking of pages. So the crawler can be more selective if high rank pages are crawled first like PageRank algorithm used in Google [14], a page has a high rank if the sum of the ranks of its back-links is high. The benefits of Google PageRank are the greatest for

under specified queries, for example: 'Stanford University' query using PageRank lists the university home page first.

Considers an ontology-based algorithm for page relevance computation where relevance of the page with regard to user selected entities of interest is computed by using several measures on ontology graph (e.g. direct match, taxonomic and more complex relationships). The harvest rate is improved compared to the baseline focused crawler (that decides on page relevance by a simple binary keyword match) [15].

The Swoogle [16] is a search engine for Semantic Web ontologies, documents, terms and data published on the Web. Swoogle employs a system of crawlers to discover RDF documents and HTML documents with embedded RDF content.

Slug [17] is a web crawler designed for harvesting semantic web content. Implemented in Java using the Jena API, Slug provides a configurable, modular framework that allows a great degree of flexibility in configuring the retrieval, processing and storage of harvested content. The framework provides an RDF vocabulary for describing crawler configurations and collects metadata concerning crawling activity. Crawler metadata allows for reporting and analysis of crawling progress, as well as more efficient retrieval through the storage of HTTP caching data.

The critical look at the available literature reveals that:

- Current well developed and understood web crawling and indexing techniques are not directly applicable, for semantic web crawling since they focus almost exclusively on text indexing.
- A semantic web crawler differs from a traditional web crawler as: the format of the source material it is traversing.
- Transformation of World Wide Web into semantic web has been almost impossible due many practical reasons like independent ontologies.
- Crawler treats user search requests without full context and do not focus on the topic that's why results returned are ambiguous and not satisfy the interest of the user.
- In other words, to be able to answer queries which exploit the semantics of Semantic Web sources, different crawling and indexing techniques compared to conventional search engines are necessary.
- However little is known about the structure or growth of the Web of SWDs. There is a strong need to prepare more agreed meaning web document notation like RDF, N3, RDFS etc.

Our approach is to provide a crawler for the collection of semantic base information so that the crawling process is effective. This can be achieved with the help of a sematic matching process between semantic descriptions of web pages ontology.

3. PROPOSED WORK

Generally, a user finds no obvious connection between the page available from the web and the pages of interest. For example consider a situation where a user seeks a document on keyword 'mouse'. On submitting the query, the search engine starts looking in DOC index for similar matching documents. Suppose two agents are involved in this search by traditional search engine using syntactic crawling, they will show list of documents to user as shown in Figure 2.

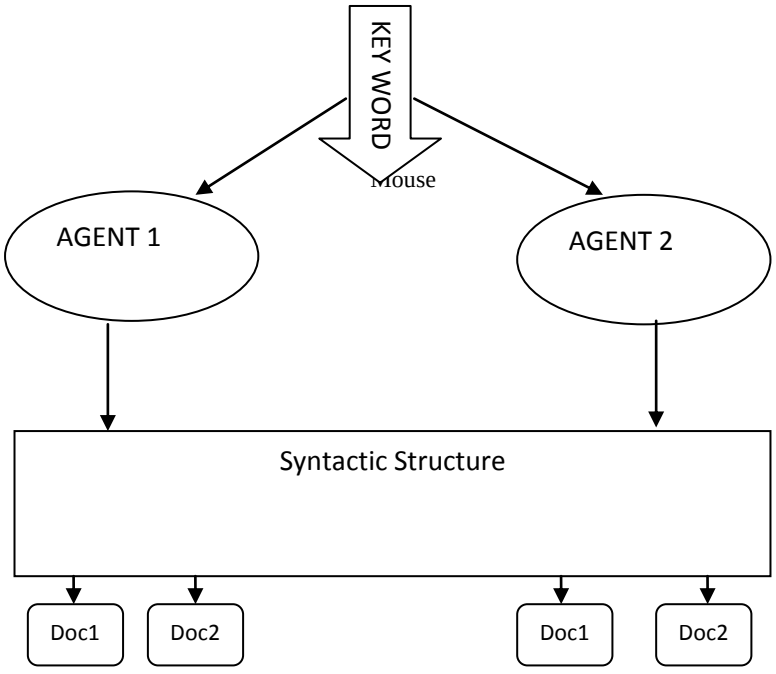

Figure.2. Syntactic search results for keyword 'mouse'

Result shown in form of DOC1 and DOC2 both had information about mouse, but these results were not sufficient to show which link the user must pursue to get exactly the same information which user desires. There may be different meaning of keyword 'mouse' like *RAT* and *PC mouse* but DOC1 and DOC2 do not specify this. When information seeker receive such results user may be in ambiguous state. The reason behind this ambiguity is that the information resources were not interpreted semantically by search agents and semantic attribute were not associated with DOC1 and DOC2.

Above example clearly specify that Semantic Web data representation format like RDF are poorly handled by current crawler. That's why other module of search engine which take

input from web crawler like indexing and query answering based on keywords does not allow exploiting the semantics inherent to structured content. Consequently, currently well developed web crawling and indexing techniques are not directly applicable, as they focus almost exclusively on text indexing. Many efforts were made to solve this ambiguity like focus crawler, domain specific crawlers and context specific crawler as discussed in last section but result were not optimized.

It has been observed that ontology based semantic web crawler is solution of such problem. A semantic search performed for the same keyword 'mouse' been represented in figure 3.

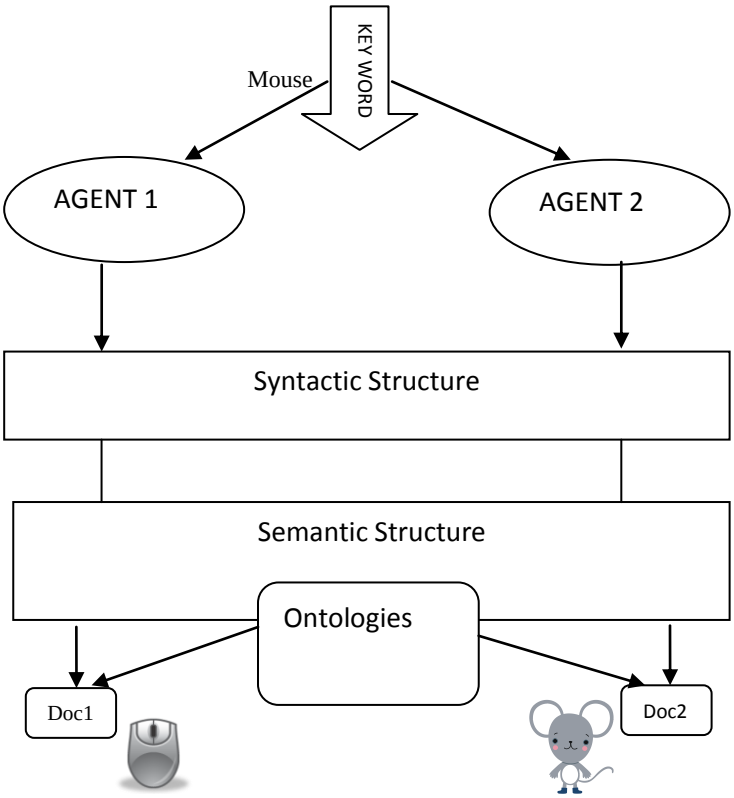

Figure.3 Semantic search result for keyword 'mouse'

The ambiguity faced by user in Figure.2 results is solved here in Figure.3, as agents search for keyword '*mouse'* is based on semantic structure which has semantic description of the terms. Thus result shown by Figure.3DOC1 and DOC2 have associated information that will help user to decide which link the user must pursue to find exactly the same information as needed.

Figure 4 shows the proposed architecture of ontology based semantic web crawler and the functionality of the main components is represented in the table 1.

Table1 1: **Components of proposed architecture with their functionality.**

Component	**Functionality**
Ontology DB	This module contains a set of pre-defined ontologies in any form may be hierarchical classes giving well defined meaning for any entity. Ontologies given here are used in matching.
Ontological Classifier	First a web page (downloaded by downloader) is parsed for URLs extraction. During this parsing we also extract semantic contents of the page. These semantic contents provide description of downloaded web page. This module process information extracted above with the help of its two sub-modules given in next two rows.
Semantic Matching	An important module which matches semantic content (page description) and ontologies for page relevancy computation. Matching process is based on certain parameter like common definition matching. Ontologies are used here to give shared meaning. Semantic matching submit all URL to next module i.e. Classification or filter module.
Classification OR Filter	It passes URL of those pages only which are meaningful as suggested by previous module. Basically it provides filtered URLs to URL frontiers.

URL Frontier	A to do list of crawler which contains set of URLs. This is a queue data-structure storing URL according to their priority. A page downloader fetches URL from this queue.
Multi-threated Downloader	A program that downloads web pages from World Wide Web using http request. It generates more than one thread for given URLs.
Web-Page Repository	A collection of downloaded web documents. This repository is used by various module of searching engine like indexing. Downloader also checks before downloading a new page weather it has been downloaded earlier or not.

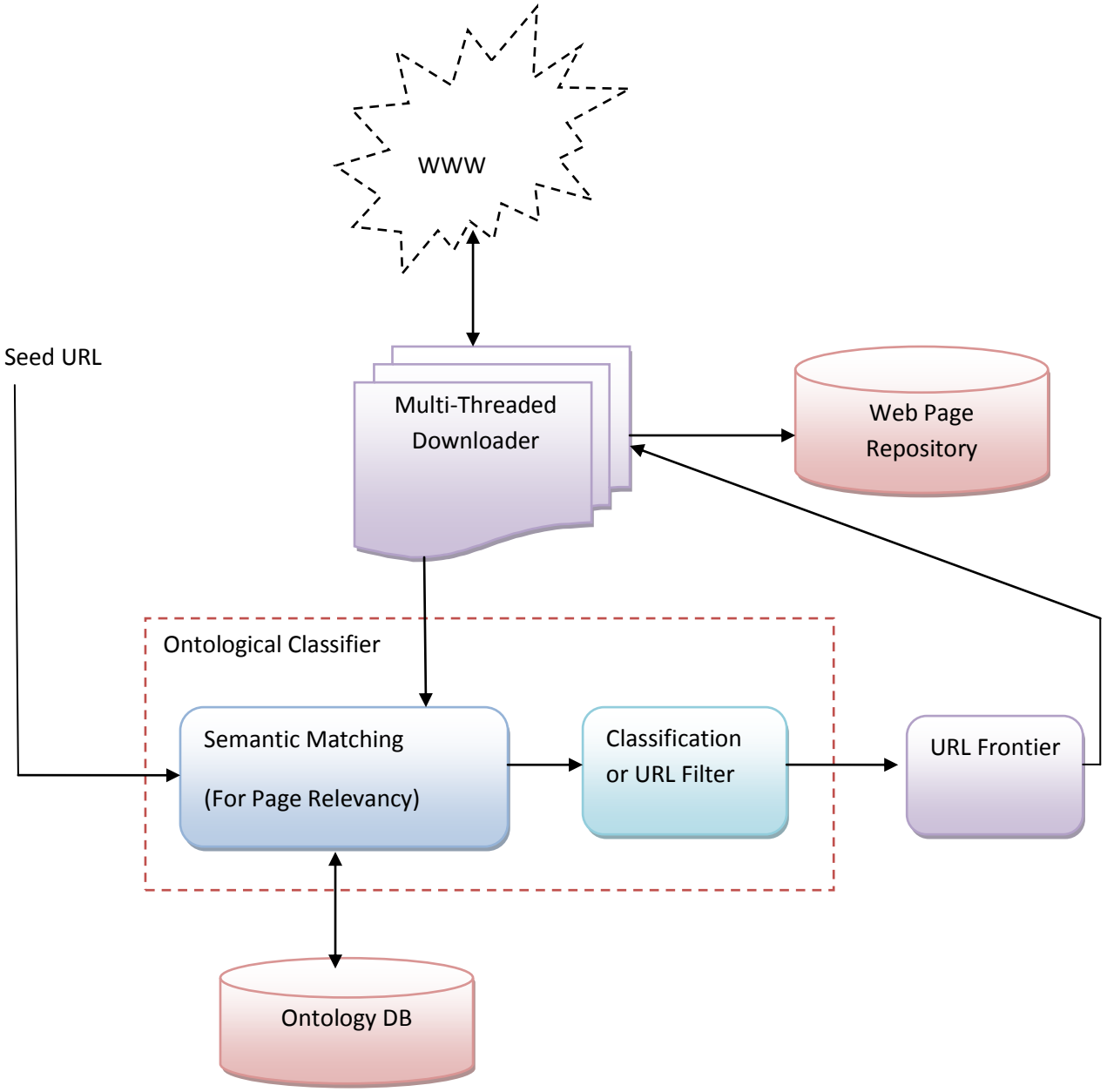

Figure.4 Proposed architecture of ontology based semantic web crawler

First time, crawling work starts when seed URLs are submitted to semantic matching process, where Seed URLs are semantically matched with ontologies. This matching will provide a semantic description to URLs. The process may add few semantic parameters to URL so that they can fetch most relevant SWDs. The parameters added here can be Web Ontology Language (OWL) [19], RDF(S) parameters. Semantic matching produces those URLs which contain addresses to relevant semantic web pages, which increases the possibility to find many other relevant URLs by the crawler. URLs given by semantic matching process are entered into classification process which classifies them into groups based

on ontology again. The similar type of URLs which agreed on same meaning is classified and entered into the URL Frontier which is the to-do list of a crawler that contains the URLs of unvisited pages. After completion of above crawling loops which was started with seed URLs, we got some new URLs in URL frontier. Now next time crawling loops working is different as discussed below.

Each time (first time onwards) crawler, based on proposed architecture start when multi-threaded downloader fetches a URL from newly constructed URL frontier. After that it sends http request to download this page from web server. If requested web document (page) exists and free from any permission issue like *ROBOT.TXT* then it is download after checking its availability with web page repository(to avoid date duplication). After downloading web page it is submitted to Ontological Classifier where it is parsed in order to fetch URLs and semantic description. This semantic description is extracted and matched with the help of ontologies. On the basis of this matching we compute page relevancy. If a page is found semantically relevant then its URL are entered into classification process which classifies and filters them. Filtered URLs are submitted into URL Frontier. These URL links extracted from web pages are stored here with their priority. This priority queue or queue with scheduler based on priority is searched for URL every time when a page downloader starts to download a page. This loop of crawling will terminate when running time of crawler expire or URL queue is empty.

4. CONCLUSION

Architecture for ontology based semantic web crawler populates the web page repository with most promising recourses has been proposed for ontology based semantic web crawler. This repository can exploit the semantic metadata to efficiently discover and extract information resources on the Web. Semantic matching between downloaded web page contents and ontologies guides the crawler for extracting relevant information which provide scope for better search engine.

5. REFERENCES

[1] C. C. Aggarwal, F. Al-Garawi, and P. S. Yu, “Intelligent crawling on the world wide web with arbitrary predicates,” in World Wide Web, 2001, pp. 96–105. .Available: iteseer.ist.psu.edu/aggarwal01intelligent.html.

[2] M. Ehrig and A. Maedche, “Ontology-focused crawling of web documents,” in Proc. of the Symposium on Applied Computing, March,Florida, USA, 2003.

[3] V. H. Tuulos, "Design and Implementation of a Content-Based Search Engine", http://www.cs.helsinki.fi/u/tuulos/tuulos-thesis.pdf, (retrieved may 2008),2007.

[4] Debajyoti, Arup Biswas, Sukanta “A New Approach to Design Domain Specific Ontology Based Web Crawler”, 10th International Conference on Information Technology – 2007 IEEE.

[5] W3C. (2011). Resource Description Framework (RDF). http://www.w3.org/RDF/.

[6] Nigel Shadbolt , Wendey hall, Tim Berners-Lee “the semantic web revisited” .IEEE intelligent system(2006).

[7] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. Scientific American, 284(5):34{43, 2001.

[8] Thomas R. Gruber, A translation approach to portable ontology specifications, KnowledgeAcquisition 5 (1993), no. 2, 199–220.

[9] J. Euzenat and P. Shvaiko. Ontology matching. Springer, 2007.

[10] L. P. Junghoo Cho, Hector Garcia-Molina, “Efficient crawling through URL ordering,” Stanford University, 1998.

[11] Q. Xu and W. Zuo, “First-order focused crawling,” in WWW ’07:Proceedings of the 16th international conference on World Wide Web, pp. 1159–1160, 2007.

[12] M. Yuvarani, N.Ch.S.N.Iyengar, A.Kannan, “LSCrawler: A Framework for an Enhanced Focused Web Crawler based on Link Semantics”. Paper presented at International Conference on Web Intelligence (IEEE/WIC/ACM), 2006 pp 794-800

[13] M. Bianchini, M. Gori, and F. Scarselli. Inside PageRank. ACM Transactions on Internet Technology, 2003.

[14] L. Page, S. Brin, R. Motwani, T. Winograd. "The PageRank Citation Ranking: Bringing Order to the Web", Stanford Digital Library Technologies Project.

[15] M. Ehrig and A. Maedche, “Ontology-focused crawling of web documents,” in Proc. of the Symposium on Applied Computing, March,Florida, USA, 2003.

[16] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. C. Doshi, and J. Sachs. "Swoogle: A semantic web search and metadata engine for the semantic web". In Proc. 13th ACM Conf. on Information and Knowledge Management, Nov. 2004.

[17] Leigh Dodds Slug: A Semantic Web Crawler 2006. http://www.ldodds.com/projects/slug/slug-a-semantic-web-crawler.pdf

[18] Michael K. Smith, Chris Welty, and Deborah L. McGuinness, Editors, W3C Recommendation, 2004.