ISSN No. 0976-5697

Volume 3, No. 5, Sept-Oct 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

An Efficient Semantic Web Search Engine

Shaik Mahaboob Basha*
Computer Science & Engineering
Nimra College of Science and Engineering
Ibrahimpatnam, (AP), India
2mahaboob@gmail.com

Sayed Yasin
Associate Professor, Computer Science & Engineering
Nimra College of Science and Engineering
Ibrahimpatnam, (AP), India
Sdyasin761@gmail.com

Md. Amanatulla
Computer Science & Engineering
Nimra College of Science and Engineering
Ibrahimpatnam, (AP), India
amanatulla@gmail.com

Abstract: Search Engines are the key to finding specific information on the vast expanse of the World Wide Web (WWW). Most of the search engines search for key words to answers the queries from users. There are many of search engines available today, but retrieving meaningful information is difficult. However to overcome this problem in traditional search engines to retrieve meaningful information intelligently, semantic web technologies are phase a vital role. The proposed work is aimed at design of an efficient search engine named as An Efficient Semantic Web Search Engine (AESWSE).

Keywords: Semantic Web, Search Engines, Information retrieval, Knowledge Discovery, Intelligent Search.

I. INTRODUCTION

Search Engines are the key to find specific information on the vast expanse of the World Wide Web. Without sophisticated Search Engines, it would be virtually impossible to locate any thing on the web without knowing specific URL. When people use the term Search engine in relation to the web, they are usually referring to the actual search forms that searches through databases of HTML documents. Today's traditional Information retrieval systems are not "intelligent" enough to retrieve the data on the Semantic Web. Thus there is a need for specialized search engines which can search the Semantic Web. The Semantic Web is an extension of World Wide Web. The major philosophical difference between the Semantic Web and the World Wide Web is that the Semantic Web is supposed to provide machine accessible meaning for its constructs whereas in the World Wide Web this meaning is provided by external mechanisms. This meaning is largely based on the meaning of the names which, in the Semantic Web [1], are URL references.

The problems described in previous sections can be resolved by maintaining metadata repository for the pages that contain domain knowledge from trusted sources. Search Engines instead of searching keywords on the web page will now search metadata for the required information. We, in this work, developed search engine that is based on this concept. Our search engine first searches the pages and then gets the result by searching for the metadata. The metadata recording could either be made manual or automated. The manual system requires input of information from the administrator of web site. This solution is improper since it can compromise the reliability and efficiency. An automated system can be developed by employing Agents that can gather information from the trusted web sites.

II. BACKGROUND

Google [2], Yahoo [3] and Bing [4] have been out there which handles the queries after processing the keywords, which makes them keyword based search engine. They only search information given on the web page. Recently, some research group's start delivering results from their semantics based search engines; however most of them are in their initial stages. Current web is the biggest global database that lacks the existence of a semantic structure and hence it makes difficult for the machine to understand the query provided by the user .When the information was distributed in web, we have two kinds of research problems in search engine i.e. how can a search engine map a query to documents where information is available but does not retrieve in intelligent and meaning full information? The query results produced by search engines are distributed across different documents that may be connected with hyperlink. How search engine can recognize efficiently such a distributed results? In general, Web Search Engines are based on the Robots i.e. Crawler-based search engines are those that use automated software agents that visit a Web site, read the information on the actual site, read the site's meta tags and also follow the links that the site connects to performing indexing on all linked Web sites as well.

The crawler returns all that information back to a central depository, where the data is indexed. The crawler will periodically return to the sites to check for any information that has changed. The frequency with which this happens is

determined by the administrators of the search engine. This works efficiently for searching the documents in the Universal level but in the case of searching activity within the Organization, Crawlers fail to maintain the complete information in the databases. Moreover, the Search engines are implemented mostly in C or C++ or PYTHON for the efficiency reasons.

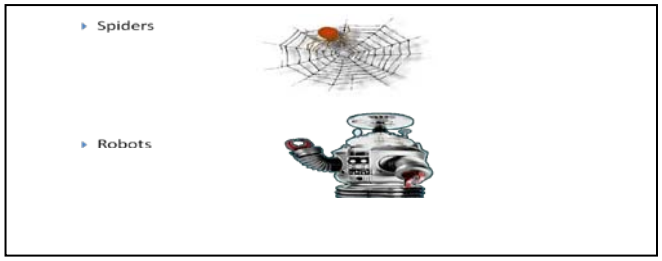

Figure 1: Spiders and Robots

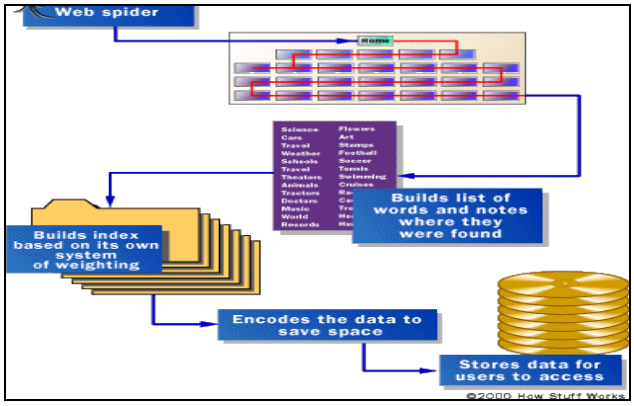

Figure 2: How do Search Engine Works

Here in this case the Human Powered Search Engine which is more appropriate for intranet search is to be implemented using JSP[8] (Java Server Pages) to produce the faster and unambiguous results. The following figure depicts the semantic web frame work it also referred as the semantic web layercake by W3C.

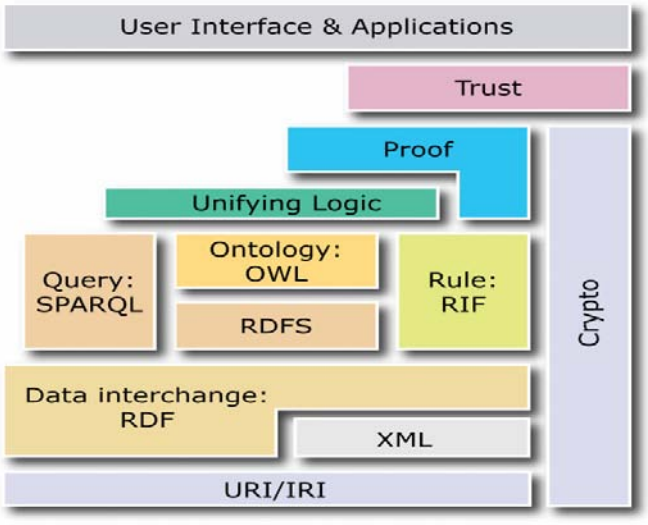

Figure.3: Semantic Web Frame work

III. CURRENT WEB & LIMITATIONS

Present World Wide Web is the longest global database that lacks the existence of a semantic structure and hence it becomes difficult for the machine to understand the information provided by the user in the form of search strings. As for results, the search engines return the ambiguous or partially ambiguous result data set; Semantic web is being to be developed to overcome the following problems for current web.

a. The web content lacks a proper structure regarding the representation of information.
b. Ambiguity of information resulting from poor interconnection of information.
c. Automatic information transfer is lacking.
d. Usability to deal with enormous number of users and content ensuring trust at all levels.
e. Incapability of machines to understand the provided information due to lack of a universal format.

Hakia[5] is a general purpose semantic search engine that search structured text like Wikipedia.Hakia calls itself a “meaning-based (semantic) search engine” [6] . They’re trying to provide search results based on meaning match, rather than by the popularity of search terms. The presented news, Blogs, Credible, and galleries are processed by hakia's proprietary core semantic technology called QDEXing [6]. It can process any kind of digital artifact by its Semantic Rank technology using third party API feeds [7]. Currently, a couple of intelligent search engines are designed and implemented for different Working environments and the mechanisms that realize this search engine are distinct. The proposed system maintains the information related to the web pages or any other documents. In addition to this the documents are subjected to the process of forward indexing in which the words present in each and every document are identified, given an unique id and maintained in the database with the combination of the document id which is given at the initial stage of the preprocessing. This process returns the searching activity to produce faster results. For each indexed document, the page rank will be calculated on basing the number of inbound and outbound links of each page initialing the search process follows the exact matching, and then it is for the tokens. In an unsuccessful case, approximate matching will be performed. The following applications are used to track the information of the documents

A. Document Preprocessing
B. Query Matching
C. Report generation with Page Ranking

A. Document Preprocessing Module:

This module mainly contains the preprocessing activity of the web pages or any other documents. This module again contains the following sub modules.

a. Deploying the information of the Documents.
b. Forward indexing of the Documents.
c. Finding the page ranks of each document.

B. Deploying the Information of the Documents:

Initially for any searching activity, database has to be maintained, and for that entire information about the web pages

and any other documents that are to be subjected to the search criteria have to be placed in database. The details include the abstract for the page, type of the page, page id, this makes the process of Semantic search meaningful. Etc.and number of inbound and outbound links for that page.

C. Forward Indexing the Documents:

This is the process of indexing the words present in each and every page and giving the word id for each word uniquely. This makes the process of Semantic search meaningful. In addition to this, the synonyms for the words, number of hits of each word in each page, maximum size of the word in that page, font type of the word in that page are also stored. These factors are maintained in the database so as to facilitate the page ranking.

D. Finding The Page Ranks Of Each Document:

After storing the above information it is always needed to find the page rank for each page. This will be based on the data structure which maintains the inbound links and outbound links of every page, etc.

E. Query Matching:

In semantic web search the query matching is another important task it will specifies about the exact matching of the Query, Token matching with the database and approximate match of the Query. This module involves the following activities.

a. Exact matching of the Query
b. Token matching with the database
c. Approximate match of the Query

F. Exact Matching Of The Query:

In semantic web search Query matching is another essential task. This is nothing but the straight forward method of looking database for the entered query.

G. Token Matching With the Database:

In general, the user query rarely matches the database which contains the stemmed abstract. So tokens are to be searched even though exact matching is possible. But at the same time, redundant matches are to be neglected.

H. Approxiamte match of the query:

This is the worst case of the searching process. In this case we have to search different combinations of the small parts of the query as well as small parts of the tokens present in the query are to be searched. This is only performed when search process gives unsuccessful results in the above two cases.

I. Report Generation with Page Ranking:

This module mainly consists of the following activities.

a. Query dependent page ranking.
b. Query independent page ranking

Here the matched documents are presented with different criteria. Based on the user choice, the search results will be displayed subject to the specified ranking algorithm.

J. Query Dependent Page Ranking:

In general users search some keywords rapidly on a particular document. These are known as Social Annotations.

These are maintained in the database at the time of preprocessing. Whenever the user enters the query, the matching activity is performed and finally the results are displayed. But here in this case the results are displayed after comparing the query with the social annotations for that page. This will be done dynamically and social similarity is calculated and hence the results are sorted.

K. Query Independent Page Ranking:

Searching will be performed without considering the entered query and are only dependent on the calculated page rank by page ranking algorithm and varying parameter like, number of keyword hits in that page, font size of the token, font type of the token in that page which are already stored at the time of preprocessing itself.

IV. TYPES OF SEMANTIC SEARCH ENGINES

Semantic is the process of communicating enough meaning to result in an action. A sequence of symbols can be used to communicate meaning, and this communication can then affect behavior. Semantics has been driving the next generation of the Web as the Semantic Web, where the focus is on the role of semantics for automated approaches to exploiting Web resources. 'Semantic' also indicates that the meaning of data on the web can be discovered not just by people, but also by computers. Then the Semantic Web was created to extend the web and make data easy to reuse everywhere. Semantic web is being developed to overcome the following main limitations of the current Web. The web content lacks a proper structure regarding the representation of information. Ambiguity of information resulting from poor interconnection of information. Automatic information transfer is lacking. Unable to deal with enormous number of users and content ensuring trust at all levels. Incapability of machines to understand the provided information due to lack of a Universal format. Currently many of semantic search engines are developed and implemented in different working environments and these mechanisms can be put into use to realize present search engines. Alcides Calsavara and Glauco Schmidt propose and define a novel kind of service for the semantic search engine.

A semantic search engine stores semantic information about Web resources and is able to solve complex queries, considering as well the context where the Web resource is targeted, and how a semantic search engine may be employed in order to permit clients obtain information about commercial products and services, as well as about sellers and service providers which can be hierarchically organized. Semantic search engines may seriously contribute to the development of electronic business applications since it is based on strong theory and widely accepted standards. Sara Cohen Jonathan Mamou et al presented a semantic search engine for XML (XSEarch)[9].It has a simple query language, suitable for a naïve user. It returns semantically related document fragments that satisfy the user's query. Query answers are ranked using extended information-retrieval techniques and are generated in an order similar to the ranking. Advanced indexing techniques were developed to facilitate efficient implementation of

XSEarch[9]. The performance of the different techniques as well as the recall and the precision were measured experimentally. These experiments indicate that XSEarch is efficient, scalable and ranks quality results highly. Bhagwat and Polyzotis propose a Semantic-based file system search engine- Eureka, which uses an inference model to build the links between files and a File Rank metric to rank the files according to their semantic importance. Eureka has two main parts:

a. Crawler which extracts file from file system and generates two kinds of indices keywords' indices that record the keywords from crawled files, and rank index that records the File Rank metrics of the files;
b. When search terms are entered, the query engine will match the search terms with Keywords Indices, and determine the matched file sets and their ranking order by an information retrieval based metrics and File Rank metrics. Project a semantic search methodology to retrieve information from normal tables, which has three main steps:

Identifying semantic relationships between table cells; converting tables into data in the form of database; retrieving objective data by query languages. The research objective defined by the authors is how to use a given table and a given domain knowledge to convert a table into a database table with semantics.

V. SOME COMMON ISSUES

We have discussed a preliminary survey of the existing and dynamic area in intelligent semantic search engines and methods. Although we have not claimed this survey is comprehensive, some common issues in the current semantic search engines and methods are concluded as follows:

a. Low precision and high recall Some Intelligent semantic search engines cannot show their significant performance in improving precision and lowering recall. In Ding's semantic flash search engine, the resource of the search engine is based on the top-50 returned results from Google that is not a semantic search engine, which could be low precision and high recall[10].
b. Identity intention of the user intention identification plays an important role in the intelligent semantic search engine.For example, in chiung-Hon leon lee introduced a method for analyzing the request terms to Fit user intention, so that the service provided will be more suitable for the user[11].
c. Individual user patterns can be extrapolated to global users. In early search engine that offered disambiguation to search terms. A user could enter in a search term that was ambiguous (e.g., Java) and the search engine would return a list of alternatives (coffee, programming language, island in the South Seas).
d. Inaccurate queries. We have user typically domain specific knowledge. And users don't include all potential Synonyms and variations in the query, actually user have a problem but aren't sure how to phrase.

VI. CONCLUSIONS

The crawler based search engines are more appropriate and suitable for web search and an effort has been made to design a powerful search engine which is more suitable for intranet search. The developed An Efficient Semantic web Search Engine is efficient search engine which is best suited for intra Organizational search process. The best feature in this search engine is that it can produce the faster results of each and every aspect of the Organization provided information is preprocessed initially. Another important aspect of this search engine is that it provides the best qualitative results with more user friendly interaction.

VII. REFERENCES

[1]. G. Antoniou and F. van Harmelen, A Semantic Web Primer, (Cooperative Information Systems). 2nd ed. 2008: The MIT Press.

[2]. "Google Search Engine".http://www.google.com

[3]. "Yahoo Search Engine".http://www.yahoo.com

[4]. "Bing Search Engine".http://www.bing.com

[5]. D. Tümer, M. A. Shah, and Y. Bitirim, An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia, 2009 4th International Conference on Internet Monitoring and Protection (ICIMP'09) 2009.

[6]. "Top5Semantic Search Engines".http://www.pandia.com/.

[7]. Fu-Ming Huang et al. "Intelligent Search Engine with Semantic Technologies"K. Elissa, "Title of paper if known," unpublished.

[8]. O' REILLY, Head First to JSP, Phil Hanna, 3rd Edition, 2001

[9]. Faizan Shaikh, Usman A. Siddiqui, Iram Shahzadi Department of Computer Science,National University of Computer & Emerging Sciences,Karachi, Pakistan:2010 IEEE 978-1-4244-8003-6/10 SWISE: Semantic Web based Intelligent Search Engine

[10]. D. Ding, J. Yang, Q. Li, L. Wang, and W. Liu, "Towards a flash search engine based on expressive semantics," in Proceedings of WWW Alt.'04 New York, 2004, pp. 472-473.

[11]. Chiung-Hon Leon Lee, Alan Liu, "Toward Intention Aware Semantic Web Service Systems," scc, vol. 1, pp.69-76, 2005 IEEE International Conference on Services Computing (SCC'05) Vol-1, 2005.