Investigation and Analysis of New Approach of Intelligent Semantic Web Search Engines

Ritu Khatri, Kanwalvir Singh Dhindsa, Vishal Khatri

Abstract- As we know that WWW is allowing peoples to share the huge information globally from the big database repositories. The amount of information grows billions of databases. Hence to search particular information from these huge databases we need the specialized mechanism which helps to retrieve that information efficiently. Now days various types of search engines are available which makes information retrieving is difficult. But to provide the better solution to this problem, semantic web search engines are playing vital role. Basically main aim of this kind of search engines is to providing the required information is small time with maximum accuracy. But the problem with semantic search engines is that those are vulnerable while answering the intelligent queries. These kinds of search engines don't have much efficiency as per expectations by end users, as most of time they are providing the inaccurate information's. Thus in this paper we are presenting the new approach for semantic search engines which will answer the intelligent queries also more efficiently and accurately. With the keywords based searches they usually provide results from blogs or other discussion boards. The user cannot have a satisfaction with these results due to lack of trusts on blogs etc. To get the trusted results search engines require searching for pages that maintain such information at some place. Here propose the Intelligent Semantic web based search engine. We use the power of XML meta-tags deployed on the web page to search the queried information. The XML page will be consisted of built-in and user defined tags. The metadata information of the pages is extracted from this XML into RDF. Our practical results showing that proposed approach taking very less time to answer the queries while providing more accurate information.

Index Terms—Information retrieval, Intelligent Search, Search Engine, Semantic web, XML, RDF.

Manuscript received on 25-March-2012

Ms. Ritu Khatri Assistant Prof in BPIT,Rohini (IP Univ)email id ritu.demla@gmail.com
Mr.Kanwalvir Singh DhindsaAssociate Prof in BBSBEC, Fatehgarh Sahib, (email id kdhindsa@gmail.com)
Mr.Vishal Khatri, Assistant Prof in SIET Greater Noida, (email id vishalvce@gmail.com)

I. Introduction

The Semantic Web is an extension of the current Web that allows the meaning of information to be precisely described in terms of well-defined vocabularies that are understood by people and computers. On the Semantic Web information is described using a new W3C standard called the Resource Description Framework (RDF). Semantic Web Search is a search engine for the Semantic Web. Current Web sites can be used by both people and computers to precisely locate and gather information published on the Semantic Web. Ontology is one of the most important concepts used in the semantic web infrastructure, and RDF(S) (Resource Description Framework/Schema) and OWL (Web Ontology Languages) are two W3C recommended data representation models which are used to represent ontologies. The Semantic Web will support more efficient discovery, automation, integration and reuse of data and provide support for interoperability problem which cannot be resolved with current web technologies. Currently research on semantic web search engines are in the beginning stage, as the traditional search engines such as *Google*, *Yahoo, and Bing (MSN)* and so forth still dominate the present markets of search engines [1] [4].

First, they do not provide the factor of reliability as the user demands. For example, when a particular user issue any query like "Which is the best University in my city?" the search engine although provides thousand of result to the user but it's difficult for the user to find out which source is reliable. The user has to sift through all the retrieved pages to find only the reliable results. Secondly, the relevancy of provided results is not up to the mark. The results against the previous query provides some of the results like" Scholarships in best University of my city?" or "Admissions in best University of my city" etc.

In this paper, we propose the semantic web based search engine which is also called as Intelligent Semantic Web Search Engines. We use the power of XML meta-tags deployed on the web page to search the queried information. The XML page will be consisted of built-in and user defined tags. The metadata information of the pages is extracted from this XML into RDF. The RDF graphs are populated by inputting through XForms. These tags will help the system in getting answers from reliable sources. For relevancy factor, we use the power of ontology in order to group the domain information of our interest.

II. Background

Information retrieval by searching information on the web is not a fresh idea but has different challenges when it is compared to general information retrieval. Different search engines return different search results due to the variation in indexing and search process. Google, Yahoo, and Bing have been out there which handles the queries after processing the keywords. They only search information given on the web page, recently, some research group's start delivering results from their semantics based search engines, and however most of them are in their initial stages. Till none of the search engines come to close indexing the entire web content, much less the entire Internet.

Presently web having lack of semantic structure which makes it is very difficult task for machine while understanding information which is provided by end user. When the information was distributed in web, we have two kinds of research problems in search engine i.e.

1. How can a search engine map a query to documents where information is available but does not retrieve in intelligent and meaning full information?
2. 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?

First problem is solved by semantic web by providing the semantic annotations in order to provide the intelligent and meaningful information [4] [5]. The Semantic web would require solving extraordinarily difficult problems in the areas of knowledge representation, natural language understanding. Following figure 1 is showing the framework for semantic search engine.

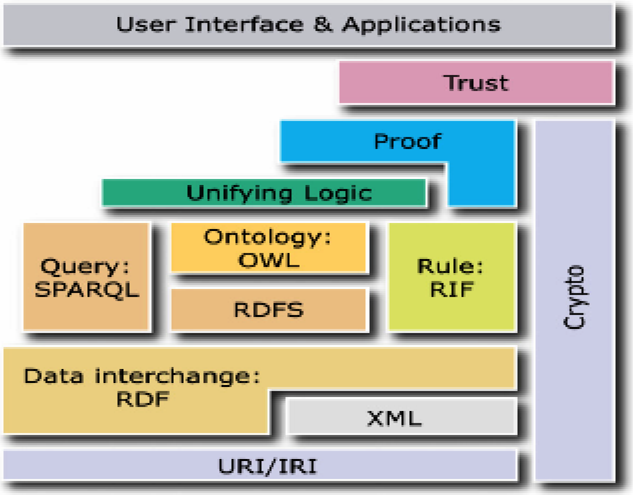


Figure 1: Semantic Web Frame Work

III. Existing Works and Limitations

The existing WWW is big global database which lacks the existence of a semantic structure and this makes 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.

- 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.
- Usability 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.

Hakia is a general purpose semantic search engine that search structured text like Wikipedia. Hakia calls itself a "meaning-based (semantic) search engine". 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.

It can process any kind of digital artefact by its Semantic Rank technology using third party API feeds. A single query by the user brings results from any repository including Web, News, Blogs, Video, Images, and Hakia Galleries and also from Credible Sources. For short queries the site displays results in categories, instead of a standard list as shown in current search engines. For longer queries, Hakia highlights relevant phrases or sentences. The results are somehow relevant and reliable but Hakia does not reveal it's inside technology. Hakia take the searched query and find the results in many categories for example from galleries, videos etc. so it took more time than the usual search engines in the retrieval of results.

SenseBot represents a new type of search engine that prepares a text summary in response to the user's search query. SenseBot extracts the most relevant results using Semantic Web technologies from the Web. It then summarizes the results together for the user as per topic. It uses text mining algorithms to parse (human readable) Web pages which lead to identification of key semantic concepts. The coherent summary is then performed from multidocuments that are retrieved. This summary itself becomes the main result of the search. Although the search results are still not relevant because the summarized result may divert the results from actual demands of the user.

The sources from which the results are coming are usually the news agencies so reliability is also somehow missing. Powerset does not search simply on keywords alone, but also try to understand the semantic meaning behind the search phrase as a whole. Powerset's first product is a search and discovery experience for Wikipedia. It attempts to use natural language processing to understand the nature of the question and return pages containing the answer. It gives more accurate results, and aggregates information from across multiple articles. The results returned by Powerset are most reliable and relevant than all the other semantic search engines however its scope is only limited to articles of Wikipedia.

DeepDyve is a powerful, professional research engine that lets users access expert content from the "Deep Web" the part of the internet that is not indexed by traditional search engines. It indexes every word in a document, but also computes the factorial combination of words and phrases in the document and uses some industrial strength statistical techniques to assess the "informational impact" of these combinations. The presentation of search results is very complex. It presents the users with many advanced options for refining, sorting or saving the search. The search results are however relatively easy to navigate. The results presented are only for the paid customers, they are not available for the general public.

IV. PROPOSED APPROACH

The problems described in previous sub-sections related to the semantic web search engines 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.

Proposed Approach for Semantic Web Search Engine

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 interoperability issues can be resolved by using W3C compliant tools. For representing domain knowledge, W3C proposes Ontologies in OWL while metadata can be represented in Graphs as RDF Triples. This approach will ensure heterogeneity at data, schema and device level.

The important part in our semantic web documents block is the instances of ontology. These instances are represented as metadata that contains information about the target web pages. We used W3C based tools to ensure semantic interoperability. In this regard we used OWL / RDF to represent metadata in a graph based structure. This tool represent information as 'subject', 'predicate' and 'object' analogous to English language grammar structure. Subject and object is an instance of classes while predicate defines the relationship or property between them. With respect to mapping, the set of subjects are domain that is mapped to set of objects as range through relationships.

Following figure 2 is showing the detail model for proposed approach.

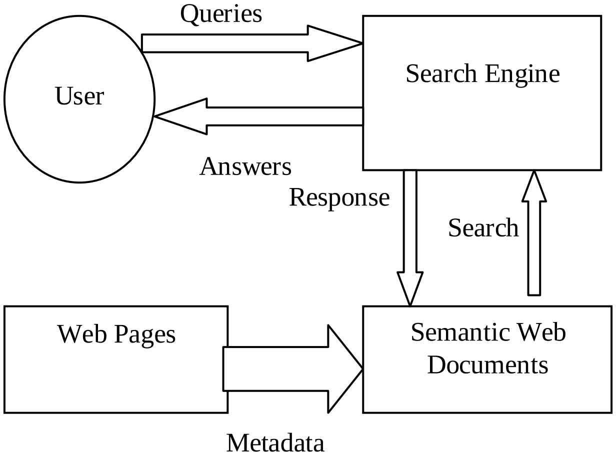


Figure 2. Design Architecture of Intelligent Semantic Web Search Engines

V. WORK DONE

In this work we have implemented the proposed semantic search engine with the proposed approach for the discovering personal name aliases from the semantic web. Our proposed approach claims that this new approach for semantic web as well as personal name aliases discovering having more effective and efficient results as compare to all the existing methods. We had done our implementation using .Net framework. We used Sql server database as backend processes.

VI. WORK DONE

For existing search engines which we discussed previously existing semantic search engines and methods. Here I present the some of the common issues in this existing intelligent search engines as concluded below:

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

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.

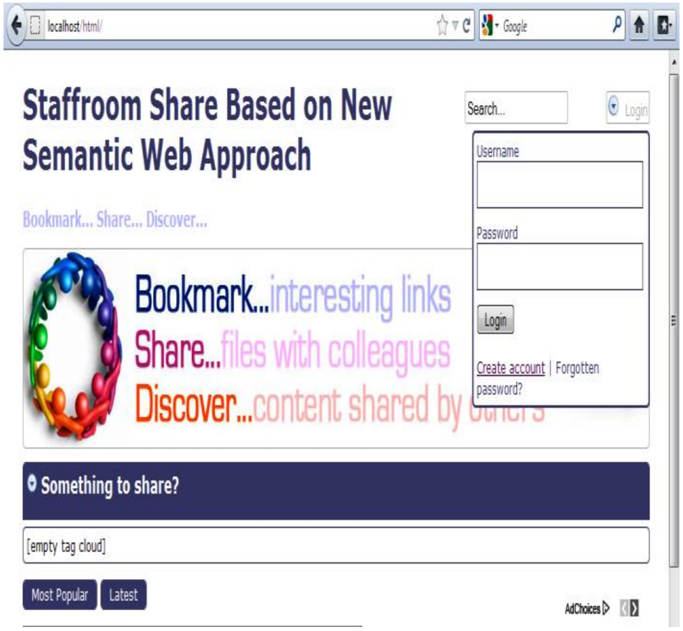
In early search engine that offered disambiguation to search terms. A user could enter in a search term that was ambiguous (e.g., Java) and the search engine would return a list of alternatives (coffee, programming language, island in the South Seas). [11]

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. [11]

VII. Experiment and Result

Here we implemented the proposed approach with the e-learning application for sharing information. Our search engine shows the better performance in case accuracy and efficiency. Following figure 3 shows the home page for this application:



Staffroom Share Based on New Semantic Web Approach

Bookmark... Share... Discover...

Bookmark...interesting links
Share...files with colleagues
Discover...content shared by others

Something to share?

[empty tag cloud]

Figure 3: Proposed Application Home Page

Following charts shows the results for precision rate, recall rate for this proposed approach as compared to existing search engines.

Precision Rate: precision is the fraction of retrieved documents that are relevant to the search.

Recall Rate: Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved

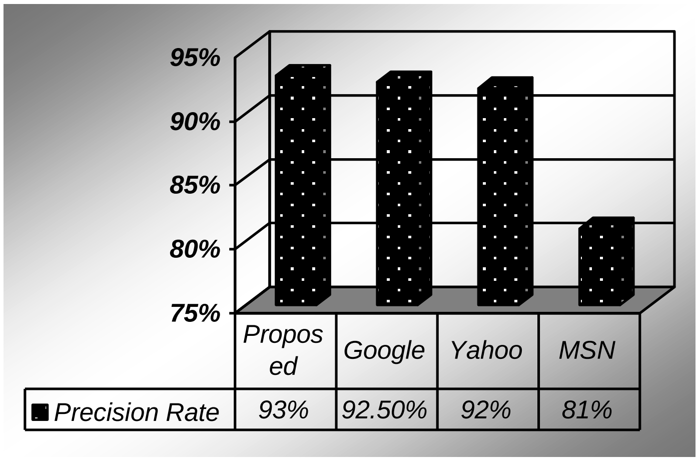


Chart 1: Precision Rate

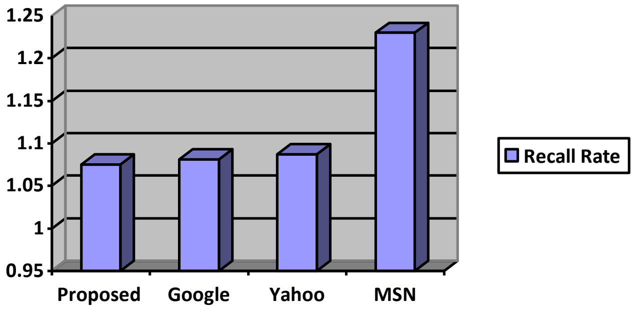


Chart 2: Recall Rate

VIII. Conclusion

The issues within the reviewed intelligent semantic search methods and engines are concluded based on four perspectives differentiations between designers and users' perceptions, static knowledge structure, low precision and high recall and lack of experimental tests.

Here we presented the use of Semantic Web tools for searching the semantic information on the pages using W3C compliant tools. This enables our system to work on any platform. There are many extensions possible in this system. Information from other domains can be included by proposing Ontologies. Currently metadata feeds is the manual process using XForms. We are currently working towards automation. There are two possible solutions for this problem. One is Semantic Crawlers while second is the use of RSS feeds. The queries are required to be parsed according to Natural Language Processing framework. Complex set of algorithms are required to be implemented.

Above results claims that proposed approach is showing the improvements in precision rate with lower recall rate as compared existing search engines.

IX. REFERENCE

[1] Berners-Lee, T., Hendler, J. and Lassila, O. “The Semantic Web”, Scientific American, May 2001.

[2] Deborah L. McGuinness. “Ontologies Come of Age”. In Dieter Fensel, J im Hendler, Henry Lieberman, and Wolfgang Wahlster, editors. Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. MIT Press, 2002.

[3] Ramprakash et al “Role of Search Engines in Intelligent Information Retrieval on Web”, Proceedings of the 2nd National Conference; INDIACom-2008.

[4] R. Bekkerman and A. McCallum, “Disambiguating Web Appearances of People in a Social Network,” Proc. Int’l World Wide Web Conf. (WWW ’05), pp. 463-470, 2005.

[5] G. Salton and M. McGill, Introduction to Modern Information Retreival. McGraw-Hill Inc., 1986.

[6] M. Mitra, A. Singhal, and C. Buckley, “Improving Automatic Query Expansion,” Proc. SIGIR ’98, pp. 206-214, 1998.

[7] P. Cimano, S. Handschuh, and S. Staab, “Towards the Self-Annotating Web,” Proc. Int’l World Wide Web Conf. (WWW ’04),2004.

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

[9]. F. Manola, E. Miller, and B. McBride, RDF primer. *W3C recommendation*, Vol. 10, No., 2004.

[10]. "World Wide Web Consortium (W3C)".http://www.W3C.org

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

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

[13] “SWISE: Semantic Web based Intelligent Search Engine” Faizan Shaikh, Usman A. Siddiqui, Iram Shahzadi Department of Computer Science, National University of Computer & Emerging Sciences Karachi, Pakistan,2010,IEEE.

[14] “Automatic Discovery of Personal Name Aliases from the Web”, Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka, Member, IEEE, June 2011.

[15] *Dan Meng, Xu Huang* “An Interactive Intelligent Search Engine Model Research Based on User Information Preference”, 9th International Conference on Computer Science and Informatics, 2006 Proceedings, ISBN 978-90-78677-01-7.

[16] Xiajiong Shen Yan Xu Junyang Yu Ke Zhang “Intelligent Search Engine Based on Formal Concept Analysis” IEEE International Conference on Granular Computing, pp 669, 2-4 Nov, 2007.