WEB PAGE RETRIEVAL USING SEMANTIC BASED SEARCH ENGINE

D.Magdalin Diviya[1], D.Praveena[2], C.Ranjani[3]

[1]*(Department of CSE, PG scholar, Sri Ramakrishna Engineering College, Coimbatore)*

[2]*(Department Of CSE, PG scholar, Sri Ramakrishna Engineering College, Coimbatore)*

[3]*(Department Of CSE, Assistant Professor, Sri Ramakrishna Engineering College, Coimbatore)*

***Abstract* -- Search engines have become one of the most powerful tools for retrieving useful information from the Internet. However, the machine on the current Web cares about the location and display of information rather than caring about the semantics of the information. Because of this shortfall of the current Web, the search results given by the most popular search engines cannot produce accurate results. In current web, the relevant information for the user's query cannot be provided by the search engines completely. Content structure in the web and methodologies followed by the web search engines are significant reasons behind this. Keyword searching is the technique used in traditional search engines. Semantics of user's query are not considered by these traditional search engines. The goal of the proposed work is to highlight the usage of semantic web and to overcome the cons of conventional search engines using semantic web concept. In this work, we present the design, development and implementation of a search engine which uses semantic web. Here searching is not just based on keyword search but also on the semantics of the keyword. In semantic web layered architecture, Ontology is used to provide more relevant search to fulfill the user's requirements. Semantic web search examine the user's query semantically to provide better search results**.

***Index Terms* –Semantic Web, Relevant Search, Semantic Search, Web Crawling, Semantic Relevance, Ontology**

I. INTRODUCTION

Current web is being a source of a huge amount of information. Mostly the information published on web is in HTML file format. There are many search engines used for searching through current web efficiently. Due to the presence of large amount of data, accurate searching is not always possible. Meaningless and unstructured information in HTML file format is the main reason behind this. Although, HTML files are useful to the user in some context but it fails to provide the meaning of data. HTML is unable to provide description and meaning of data. Semantic Web is a concept based and it provides meaning of data. Only keyword based search is another important problem with traditional search engines. Traditional search engine doesn't use the description and meaning of data in the process. So, it fails to provide the relevant and most accurate search results to user. Semantic web[5] is web technology which provides both the description and meaning of data. Semantic web can be considered as next generation of existing web. Current web content is human understandable but are not machine understandable.

A. Semantic Web:

Semantic web data is both human and machine understandable which makes it more efficient standard for data representation on the web. Ontology is an important component of semantic web layered structure and it provides knowledge base to semantic web data. It is used to provide meaning, intelligence and description to data and it also describes relations among the concepts. Ontology is mainly domain specific as: Medical, travel, education etc. Web Ontology Language (OWL) is used for constructing ontology[6]. There are still many researches going on for the improvement of existing search engines. Search engines have made a drastic improvement but the web data amount is increasing at a rapid rate than the technology which makes the issues like the same. In traditional search even if search is successful, user has to go through all the resulted web pages to extract the information which is more relevant to his/her query. This task is very time consuming. There are various applications that can make keyword based search but when it comes to relevant search, meaning of query and data is significant, traditional search engines fails to do this. Aim of our proposed work is to design and implement a search based on semantic web that can examine the semantic of user query and use the properties of semantic web structure to provide most appropriate results to user by using ontology for concept retrieval. Semantic web based search is performed on semantic web and it retrieves the most relevant results for user query that belongs to one specific domain. Semantic web is next generation of the current web. Semantic web pages are machine understandable and more structured. Ontology plays a significant role in semantic search as it provides the knowledge base for web. This knowledge base helps to make query results more accurate in context of users.

B. Semantic Analysis:

Search engines have become the most powerful tool in the Internet for information retrieval. However, the search results of the most popular search engines are not satisfactory. It is surprising that the web pages resulted by the search engine do matches with the user's input keyword and yet, most of those results are useless. This problem can be depicted using the following three keywords, “Mumbai,” “Five star” and “Hotel” in the popular search engine Google [2], and submitted it. Even in the first page of the Google search results, we found only one hyperlink directly linking to the homepage of a Five Star hotel located in Mumbai .We did find another Five Star hotel, the Le Meridian Hotel, on this page. However, the hotel is located in Delhi, not in Mumbai. The search engine returned this page just because the page involves the keyword “Mumbai.” When we examined this page carefully, we understood why the keyword “Mumbai” appears on this page: It is just because the “Mumbai Coromandel Restaurant” advertisement was displayed in that page. In fact, the relations between them were erased because there is no way to record the relations between entities under the system architecture of the current Web. Thus, the search engine cannot return the pages we want. Relations lost—this is the key of the whole problem!

For example, with regard to “Mumbai” and “hotel,” one of the relations between them is “Located In.” “Hotel” also relates to “Five Star;” the relation between them is “has Rating.” Together, the relations form the semantics of “Hotel” in this context: The hotel is located in Mumbai and it has been ranked as a “five star” hotel. In fact, the semantics of an entity exist in the relationship between the entity and others. The relationships between entities have to be defined before the machines can understand the semantics of each other for communications between the machines. The next generation Web [10], [11], Semantic Web, offers a solution to this problem in the system architecture level. In Semantic Web, the semantics information is presented by the relation with others and is recorded by RDF

[9]. Then, the relation is interpreted by OWL [1].

II. RELATED WORK

Semantic Web is literally an intelligent web. Pages in Semantic Web are very well structured, which makes it understandable by the machine and its ontology adds intelligence. Ontology plays a vital role in providing a knowledge base for the concepts. It provides a common understanding of a term and also its relationship with other terms. Using their relationships and concepts a hierarchy can be formed. Considering this domain, each entity denotes a class and class definitions. The entities are used to get their properties and related terms from the ontology which is constructed. It is the next generation web which overcomes the drawbacks of current web. Here we have proposed a Semantic Web based crawler that can automatically discover web pages to our domain with greater relevancy and checks for page relevancy for more accurate search results over semantic web. Figure 1 represents the low level design of proposed SWC.

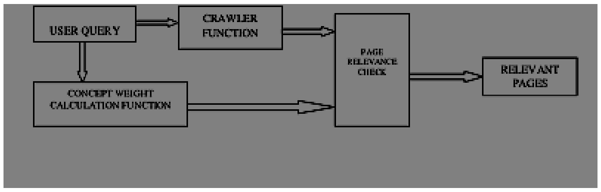

Fig. 1. Low Level Design of Proposed SWC

The three major components of our proposed SWC system are as follows:

A. Domain Ontology Construction

B. Query Expansion and Concept Retrieval

C. Relevance Calculation of Retrieved Web Pages

A. Domain Ontology Construction:

This component describes the ontology construction which forms the knowledge base of the SWC system. The concepts related to Animal domain are gathered from several websites and from various other information sources. These concepts are well structured in a hierarchical form in ontology which acts as database for the entities related to Animal domain. The ontology in our proposed system is mainly highlighted with Animal domain. Figure 2 shows the part of the ontology which is constructed. In this part of ontology Animal class have three child classes, mammals is one of them as described in figure 2. Our domain ontology is denoted in Web Ontology Language (OWL). OWL describes classes, characteristics and relations among these concepts so that makes the semantic web content machine understandable [4].

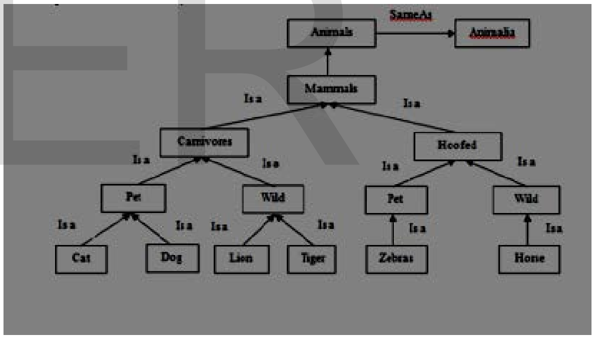

Fig. 2. Part of the domain ontology

B. Query Expansion and Concept Retrieval:

This component is used for query expansion. User query is expanded to provide a much better search result. In this component the concepts related for the keywords that are contained in the user query are retrieved[3]. The domain keywords that are related semantically to the user query are extracted from the constructed domain ontology. This results in the retrieval of large number of words which are semantically related. These refined queries are the queries with expanded keywords that has more semantic relevance and it has the semantic information about these keywords.

C. Relevance Calculation of Retrieved Web Pages:

This component produces a wide set of semantically relevant web links of semantic web pages to the user given query. The expanded query in turn serves as an input for searching the web and performs crawling functions. Once the crawling function is completed system produces web links of pages relevant to the user query. In this step all these relevance score is computed for all web links that are related. This relevance score is then used to check the web page rank. All the pages having relevance score greater than a specified threshold value are considered relevant for user query. The web links are ranked on the basis of their semantic relevance which is attained by the means of domain keywords which are extracted from the ontology.

III. PROPOSED SYSTEM

To solve the challenges presented above, we propose a scheme. The detailed architecture of proposed SWC are as follows,

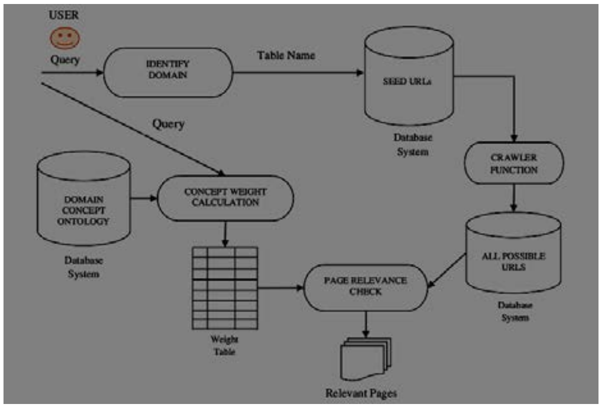

Fig.3. Complete architecture of proposed SWC

The main contributions of this paper include:

A. Ontology Construction:

Ontology is the formal representation of knowledge as a set of concepts under a domain which is the knowledge base constructed based on the concepts related to the any domain. The concepts[7] from a domain are gathered from various web sites and other sources. Ontology tree has the structured representation of these concepts which serves as a database for entities of a particular domain.

B. User Input:

Primarily, the user enters the query in natural language. Then the query is processed and the semantically relevant web links are retrieved as output after filtering out the irrelevant links.

Algorithm:

Input: User query (Q), concept ontology(co)

Output: Concept Weigh Table (Wt)

```
Get the ontology graph (G) and parse the user query Q
Ki=Key terms in the user query Q
for each key term ki
do
Search for the key term ki in the graph G
R=conceptually related term of ki
for each related term ti
do
Match (Ri)
Wi=getWeight (Ri)
Save Ri and Wi in Wt
```

```
        end do
    end for
    end do
end for
```

C. Concept Extraction from Ontology:

This is the very vital process in SWC as here semantically related terms of the user query are retrieved from domain ontology. The user query is used to query the ontology and is matched with the concepts available in the ontology[8] so as to get more semantically related concepts. This process outputs the collection of concepts and properties which are semantically related to the user query.

D. Query Expansion:

The searching process enhances with the help of the expanded query having the collection of semantic concepts and property[13]. This step is used to provide a semantic meaning to the user query.

Algorithm:

Input: Web pages (Wp), Weight table (Wt), Threshold value(Tv)

Output: Relavancy counts for each semantic web pages

```
Set relevancy point Rp=0
for each wp
do
    for j=1 to M // M=Concept count in weight table wt
    do
        Count concept frequency CF in Wp
        Rpi=rpi+(cf*wti)
        if(rpi>tv)
            Save wpi and rpi
        else
            Discard wpi
        end do
    end for
end do
end for
```

E. Relevance Calculation of Retrieved Web Pages:

In this step, the expanded query is used to check for the relevancy in the web pages with respect to the user's query. For each web page, the relevance point is calculated and based on it; those pages are ranked in appropriate of their semantic relatedness. If that value is lower than the predefined threshold, those pages are marked as irrelevant and they are discarded.

Algorithm:

Input: Seed URL Table (St), Threshold Score (Ts)

Output: Relevant pages for user query

```
Get Cs = Count of seeds in St
for(i = 1 to Cs)
do
    Get Wp = Web page for Si // S = Seed URL
     crawl on Wp
    for( each next level URL )
```

```
        do
            Wpn = Web page for N// N = Next level URL
            R = Relevance score for Wpn
            if( R ≥ Ts )
                Dispaly Wpn
            else
                Discard Wpn
        end do
    end for
end do
end for
```

IV. CONCLUSION

We have proposed a design and development process of Semantic Web Crawler for retrieval of web documents semantically from semantic web in a domain. This work defines the merits of semantic web over the current web for an efficient search. Traditional search engines are irrelevant for the information search from web. The Structure of web contents of current web is responsible for this. Semantic web is very relevant for representing the web content and have many properties that makes search more relevant for users. In the presented work, semantic properties of the Semantic Web have been used to provide a better search application to user.

V. FUTURE WORK

In this work semantic properties of the Semantic Web have been used to provide a better search. For semantic related concept search, this system uses the ontology. But, this proposed system is designed for a specific domain. This work can be extended for multiple domains in future. For query expansion, semantic related concepts from ontology are used and concepts relations are used in the calculation of relevance of pages. In further extension of this work some improved and efficient algorithms can be utilized for ranking of the pages and query expansion. Future directions are to improve the complexity flaw and it can be improved, if there is a combination of various other semantic web techniques such as Description logic and information retrieval algorithms.

VI. REFERENCES

[1] World Wide Web Consortium (W3C) (2014) OWL, Web Ontology Language (OWL). http://www.w3.org/2014/OWL/ [accessed 05/03/2014].

[2] The Google.com search engine, http://www.google.com/, 2014.

[3] Luo, J. and Xue, X. (2013). “Research on Information Retrieval System Based on Semantic Web and Multi-Agent,” International Conference on Intelligent Computing and Cognitive Informatics. 978-0-7695-4014-6/13, IEEE (2013).

[4] A. Go´mez-Pe´rez and O. Corcho, “Ontology Languages for the Semantic Web,” IEEE Intelligent Systems, vol. 17, no. 1, pp. 54-60, Jan.-Feb. 2013

[5] Gerd Stumme, G., Hotho, A. and Berendt, B. (2013). “Semantic Web Mining, State of the Art and Future Directions,” Web Semantics: Science, Services and Agents on the World Wide Web 4, p.124–143 (02/02/2013).

[6] Su, X. and Ilebrekke, L. “A Comparative Study of Ontology Languages and Tools,”

Conference onAdvanced Information System Engineering (CAiSE'02), (2013).

[7] Noy, N., Sintek, M., Decker, S., Crubezy, M., Fergerson, R. and Musen, M.(2012). “Creating semantic web contents with Prot´eg´e-2012,” IEEE Intelligent Systems (2012).

[8] Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R. and Wenke, D. “OntoEdit: Collaborative Ontology Engineering for the Semantic Web,” First International Semantic Web Conference 2012 (ISWC 2012), 2012.

[9] Magkanaraki, A., Alexaki, S., Christophides, V. and Plexousakis, D. (2012). “Benchmarking RDF Schemas for the Semantic Web,” The Semantic Web – ISWC 2012, Vol. 2342, Springer p.132-146.

[10] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” Scientific Am., vol. 284, no. 5, pp. 34-43, 2012.

[11] D. Beckett, “RDF/XML Syntax Specification (Revised),” http:// www.w3.org/TR/2011/REC-rdf-syntax-grammar-20110210/, 2011.

[12] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D.L. McGuinness, P.F. Patel-Schneider, and L.A. Stein, “OWL Web Ontology Language Reference,” http://www.w3.org/TR/2011/ REC-owl-ref-20040210/, 2011.

[13] Hongsheng, W., Jiuying, Q. and Hong, S. (2011). “Expansion Model of Semantic Query Based on Ontology,” Web Mining and Web-based Application. WMWA '11. IEEE (2011).