Volume 4, Issue 12, December 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Semantic Web Crawler for More Relevant Search Using Ontology

Amit Upadhyay, Amit Paul, Pijush K. Dutta Pramanik

Computer Science Engg., B.T.K.I.T, Dwarahat,

Almora, Uttarakhand, India

Abstract: In current web scenario, search engines are not able to provide the relevant information for user's query to full extent. Structure of contents in web and strategies followed by web search engines are crucial reasons behind this. Conventional search engines are mainly based on keyword searching. Semantics of user's query are not used by these search engines. The objective of the presented work is to highlight the property of semantic web and to overcome the drawbacks of conventional search engines using semantic web. This work presents the design, development and implementation of a semantic web search engine. Here searching is not just based on keyword search. Ontology which has its importance in semantic web layered architecture is used to provide more relevant search to fulfil the user's requirements. Semantic web search analyse the user's query semantically for better search.

Keywords: Semantic Web; Semantic Search; Web Crawling; Relevant Search; Ontology; Semantic Relevance.

I. INTRODUCTION

Current web is source of a very huge amount of information. Most of information published on web is in HTML file format. There are also many good search engines, which are used for searching through current web. But this huge amount of data creates problems in accurate searching. Unstructured and meaningless information in HTML file format is main reason behind this. Although, HTML files are useful to user in some context but at the same time it fails to provide the meaning of data. HTML is unable to provide description as well as meaning of data. Semantic web technology is concept based and it provides meaning of data. Only keyword based search is another major problem with conventional search engines. Conventional search engine doesn't use the description and meaning of data in the process. So, it fails to provide the relevant search results to user.

Semantic web is web technology which provides description and meaning of data. Semantic web can be considered as next generation of current web. Current web content is human understandable but fails with machines. Semantic web data is human understandable as well as machine understandable which makes it more efficient standard for data representation on web.

Ontology is a crucial component of semantic web layered structure and it provides knowledge base to semantic web data. It is used to provide meaning, description and intelligence to data and it also describes relations among concepts. Ontology is mainly domain specific as: Travel, education, medical etc. Web Ontology Language (OWL) is used for construction of ontology.

There are still many researches that are going on for the improvement of search engines. In the last few years search engines have made a drastic improvement but the amount of web data is increasing at a greater rate than technology which makes the problems like the same. In conventional search even if search is successful, user has to go through all the pages to extract the information relevant to his/her query. This is very time consuming task. Of course there are many various applications that can make search based on keyword but when it comes to relevant search, meaning of query and data is important, conventional search engines fails at this point.

Aim of our presented work is to design and implement a semantic web search that can analyse the semantic of query and use the properties of semantic web structure to provide more accurate results to user by utilizing ontology for concept retrieval. Semantic web search is performed on semantic web and it retrieves the most relevant results for query that belongs to a specific domain. Semantic web is next level of current web. Semantic web pages are more structured and machine understandable. Ontology plays an important role in semantic search as it provides knowledge base for web. This knowledge base helps to make query results more relevant in context of users.

II. LITERATURE SURVEY

Semantic web search is an upcoming trend in World Wide Web. There are still many researches that are going on in this field. In this section we present a survey of some of the research works that are done in semantic web.

In this survey [1] authors describe growing size and complex structure of current web as the main requirement for evolution of semantic web mining. This survey defines semantic web mining as combination of two different areas: Semantic Web and Web Mining. It also presents a view of what is web, what it present and how it can be represented and analysed using the concepts of semantic web mining. Benchmarking RDF schemas for the semantic web [8] not only describes the requirement of semantic web but also its evolution. Here the use of RDF (Resource Description Framework) and RDFS (Resource Description Framework/Schema) is presented. This work describes RDF and RDFS as the better means for semantic web creation and the representation of web data.

In the study [2], authors have designed a new searching system, which is a multi agent based system. All these agents are used to do the different tasks: processing, recognition, extraction, extension and matching of contents. In this study RDFs are used to analyse and determine the semantic of contents. A new algorithm for extraction and matching has been discussed by authors. A new agent based approach of content extraction is defined which can adapt the user's behaviour and share the information between different users to provide much more relevant search results for user's query.

In the study [6] Authors have used ontology for query expansion for better search. This study is about the used WordNet lexical dictionary for query expansion and has introduced the study about integration of ontology knowledge base for information retrieval, image retrieval and text recognition. Ontology for education domain has been constructed for specific University. Usage of Protégé framework for ontology constructor has been defined by authors in this study. The main objective of this study is to illustrate the construction process of ontology and to provide guidelines to work within it.

In study semantic information retrieval system [9], a framework for information retrieved for web has been introduced. In this study ontology and SPARQL query language has been used for information retrieval. Ontology used in this is specific to Sports domain. SPARQL is used to query ontology for concept discovery. All these discussed study and works are related to semantic web mining and gives an overview of semantic web, its requirement and its usage.

III. THE PROPOSED WORK

Semantic Web is an intelligent web. Pages in Semantic Web are well structured, that makes it machine understandable and its ontology makes it more intelligent. Ontology plays an important role in providing a knowledge base for concepts. It provides a common understanding of a term and also its relationship with other terms. Using these concepts and their relationship a hierarchy can be formed. Considering our domain, each entity will express a class and class definitions. The entities can be used to get their related terms and properties from the constructed ontology.

Semantic Web is the next generation web. It provides many advantages over today's web. Here we have proposed an effective Semantic Web based crawler that can automatically discover web pages relevant to our domain and also check for page relevancy for more relevant results over semantic web. Figure 1shows the low level design of proposed SWC.

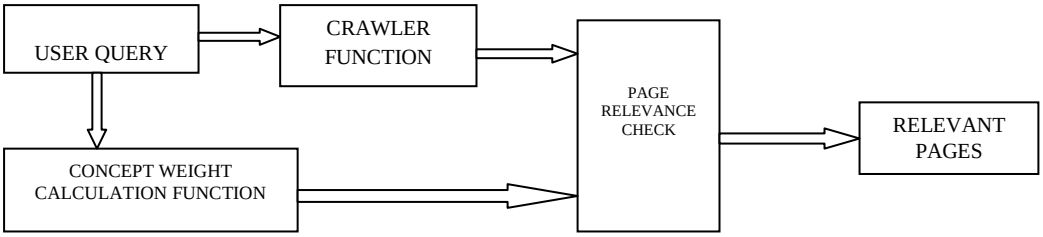

Fig. 1: Low Level Design of Proposed SWC

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

1. Domain Ontology Construction
2. Query Expansion and Concept Retrieval
3. Relevance Calculation of Retrieved Web Pages

3.1 Domain Ontology Construction

This component describes the construction of ontology which forms the knowledge base of SWC system. The concepts related to Animal domain are gathered from various web sites and from various other sources such as Wikipedia. These concepts are structured in a hierarchical form in ontology which serves a as database for the entities related to Animal domain. The ontology in our system is focused on Animal domain. Figure 2 shows the part of the constructed ontology. 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 represented in Web Ontology Language (OWL). OWL describes classes, properties and relations among these concepts in such a way that makes it machine understandable semantic web content.

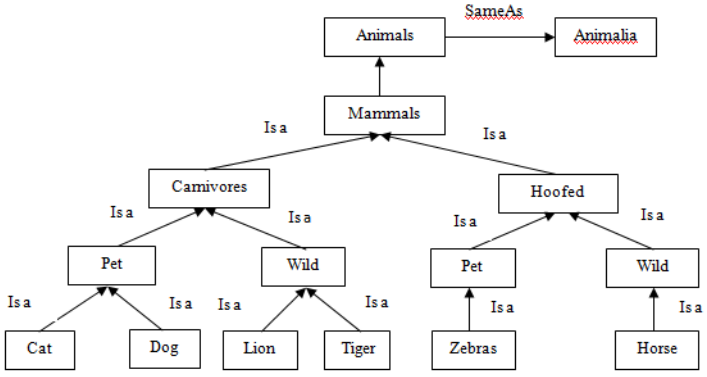

Fig. 2: Part of the domain ontology

3.2 Query Expansion and Concept Retrieval

This component is responsible for query expansion. Query given by the user is expanded to provide a much better search result. In this component the related concepts for keywords that are contained in the query are retrieved. The domain keywords that are semantically related to the query are extracted from the domain ontology. This step results in the retrieval of more number of semantically related words. These refined queries are queries with expanded keywords that has more semantic relevance involved and it contains the semantic information about these keywords.

3.3 Relevance Calculation of Retrieved Web Pages

This component produces a wide set of web links of semantic web pages that are semantically relevant to the user query. The expanded query serves as input for the web search and performs crawling functions. After the completion of crawling function system produces web links of pages related to user query. In this step all these relevance score is calculated for all related web links. This relevance score is then used to check the rank of web pages. All the pages that have relevance score greater than a specified threshold value are considered relevant for user query. The web links are ranked based on their semantic relevance which is attained by the means of extracted domain keywords from the ontology.

IV. DETAILED ARCHITECTURE

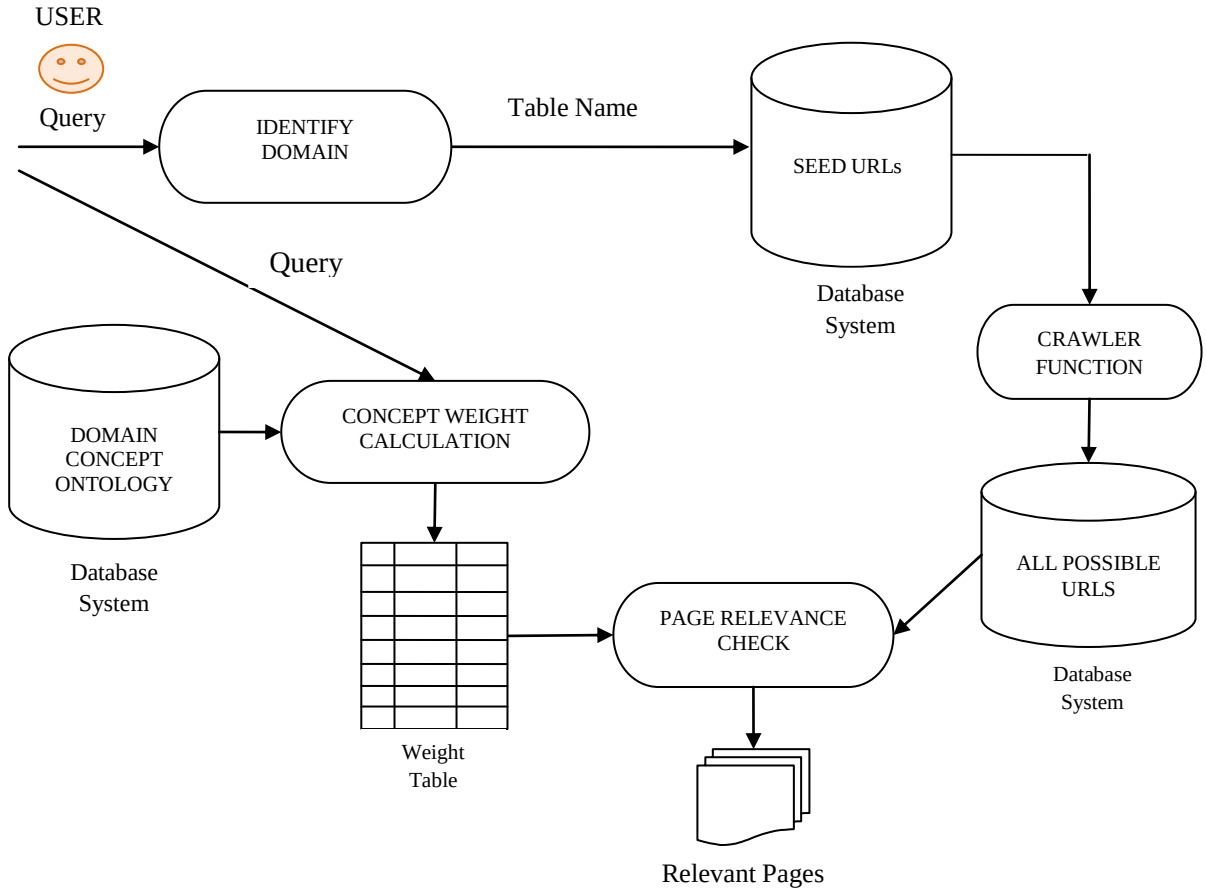

Fig. 3: Complete Architecture of Proposed SWC

4.1 Ontology Construction

Ontology, a formal representation of knowledge as a set of concepts within a domain, forms the knowledge base for our project that is constructed based on the concepts related to the Animal domain. The concepts related to the Animal domain are gathered from various web sites and other sources such as Wikipedia. These concepts are structured in a hierarchical form in ontology which serves as a database for the entities related to the Animal domain.

4.2 User Input

The user of the system enters a query related to the Animal domain in natural language. The expected outputs of this query are semantically relevant web links. The irrelevant links are filtered out.

Input: User query (Q), Concept Ontology (Co)
Output: Concept weight table (Wt)
Algorithm:

Get graph (Gc) of Co and parse Q
K_i = Key terms in Q


```
for( each Ki )
        do
        Traverse graph Gc for Ki
        T = Conceptually related term of Ki
        for( each Ti )
                do
                matchPoperty(Ti)
                Wi = getWeight(Ti)
                Save Ti and Wi inWt
                end do
        end for
        end do
end for
```

Algorithm for Weight Table Construction

4.3 Concept Extraction from Ontology

This process plays a very important role in SWC because in this step semantically related terms for user query are retrieved from domain ontology. Initial query given by user is used to query ontology and this query is matched with the concepts contained in the ontology to get more semantically related concepts. At the end of this process we get a collection of concepts and properties that are semantically related to user query.

4.4 Query Expansion

Extracted semantic concepts for user query are used to make query more relevant for searching. Concepts are added with user query and this query is now used for searching. This step is used to give a semantic meaning to user query.

Input: Web pages (Wp) , Weight table (Wt) , Threshold Relevance Score (Th_{rs})

Output: Relevance score of semantic web pages (Rs)

Algorithm:

```
Set Relevance Score Rs = 0
for(i =1 to N)  // N= No. of crawled web pages
        do
        for(j = 1 to M)   // M= Concept count in weight table
                do
                Count concept (Cj) frequency in Wpi
                Rsi = Rsi + (Cf * Wti)
                if(Rs > Thrs)
                Save Wpi and Rsi
                else
                Discard Wpi
                end do
        end for
        end do
end for
```

Algorithm for Calculation of Relevance Score for the Web Pages

4.5 Relevance Calculation of Retrieved Web Pages

In this process expanded query is used to search the web pages relevant to user query. Using expanded query, web links are checked for their relevancy to user requirements. After calculation of relevance score for each web page these pages are ranked according to their relevance score. If relevance score for any webpage is lower than defined threshold relevance score than that page is marked as irrelevant for user and is filtered out. Relevance calculation is applied to all web links obtained from all possible queries. On applying relevance check, the web links are ranked in the appropriate order of their semantic relatedness.

Input: Seed URL Table (St), Threshold Relevance Score (Th_{rs})

Output: Relevant pages for user query

Algorithm:

```
Get Csu = Count of seeds in St
for( i = 1 to Csu)
        do
        Get Wp = Web page for Sui     // Su = Seed URL
        Perform crawling function on Wp
        for( each next level URL )
```



```
                    do
                    Wpn = Web page for Nu_i      // Nu = Next level URL
                    Rs = Relevance score for Wpn
                    if( Rs ≥ Th_rs )
                    Dispaly Wpn
                    Else
                    Discard Wpn
                    end do
            end for
            end do
end for
```

Algorithm for Crawler Based on Domain Ontology Relevance

V. CONCLUSION

We have proposed a design and development process of Semantic Web Crawler for semantic retrieval of web documents from semantic web in certain domain. This work defines advantages of semantic web over current web and exploits its property for a better search. Conventional search engines are not relevant for searching information from web. Structure of web contents of current web is responsible for this. Semantic web is quiet more relevant for web content representation and also have many properties that can make searching more relevant for users. In presented work, semantic properties of Semantic Web have been used to provide a better search application to user.

VI. FUTURE WORK

In this proposed work semantic properties of Semantic Web have been used to provide a better search options. This system uses the ontology for semantic related concept search. But, this proposed system is designed for a specific Animal domain. In future this work can be extended for multiple domains. Semantic related concepts from ontology are used for query expansion and concepts relations are used in the calculation of relevance of pages. In further extension of this work some other improved algorithms can be utilized for query expansion and ranking of the pages.

REFERENCES

[1] Gerd Stumme, G., Hotho, A. and Berendt, B. (2006). "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 (02/02/2006).

[2] Luo, J. and Xue, X. (2010). "Research on Information Retrieval System Based on Semantic Web and Multi-Agent," International Conference on Intelligent Computing and Cognitive Informatics. 978-0-7695-4014-6/10, IEEE (2010).

[3] World Wide Web Consortium (W3C) (2004) OWL, Web Ontology Language (OWL). http://www.w3.org/2004/OWL/ [accessed 05/03/2014].

[4] Su, X. and Ilebrekke, L. "A Comparative Study of Ontology Languages and Tools," Conference onAdvanced Information System Engineering (CAiSE'02), (2002).

[5] Manuel, D., Maria, M., Alfonso, U. L. and Jose, P. (2010). Using WordNet in Multimedia Information Retrieval. CLEF 2009 Workshop, Part II, LNCS 6242, Springer-Verlag Berlin Heidelberg p. 185–188.

[6] 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 2002 (ISWC 2002), 2002.

[7] Magkanaraki, A., Alexaki, S., Christophides, V. and Plexousakis, D. (2002). "Benchmarking RDF Schemas for the Semantic Web," The Semantic Web – ISWC 2002, Vol. 2342, Springer p.132-146.

[8] Noy, N., Sintek, M., Decker, S., Crubezy, M., Fergerson, R. and Musen, M.(2001). "Creating semantic web contents with Prot´eg´e-2000," IEEE Intelligent Systems (2001).

[9] Hongsheng, W., Jiuying, Q. and Hong, S. (2009). "Expansion Model of Semantic Query Based on Ontology," Web Mining and Web-based Application. WMWA '09. IEEE (2009).

