Meta-Search Engine based on Semantic Similarity for E-Learning Application

Harini Natarajan, Anita C.S.
Department of Computer Science and Engineering
R.M.D. Engineering College,Chennai, India
nharinirmd@gmail.com, csa.cse@rmd.ac.in

***Abstract*— With the sharp growth in internet usage, the amount of data stored in web is outsized. The web users simply query their needs to get appropriate results. This is enabled through search engines. Search engines are software, particularly designed to search information from World Wide Web (WWW).Such traditional engines go in vain while searching the semantic web. Thus metasearch engines were developed, which aims in utilizing the traditional search engines to enhance the search results to retrieve links that are highly related to user query. A semantic metasearch engine is proposed which includes semantic analysis of user queries and Meta search engine benefits. The proposed system uses WordNet tool to mine synonymous words for the given user query. Then these words are passed to the Google search engine to retrieve top 10 links. The tag of each link is pre-processed to obtain the keywords and ontology is created. The genetic algorithm is used to rank the links which has high word fitness to the user query. Finally after harmonization, the keywords are passed to Google to get significant results. The resultant links prove to be greatly related to user query and are also mostly active links. This metasearch engine can be implemented in E-learning domain to aid the students and researchers.**

***Keywords*—semantic web; Meta search engine; web search; wordnet; natural language processing**

I. INTRODUCTION

A web service is any piece of software that is available over the internet and uses a standard XML messaging system. In order to encode the communications to a web service, XML is used. As all the communications are in XML, web services are not restricted to any one operating system or programming language. Web Services are distributed, modular, dynamic, self-contained applications that can be located, invoked, described or published over the network to build processes, products, and supply chains. Web service allows programs submit requests to erstwhile programs over the Internet via open protocols and standards. Many traditional search engines like Google are boosting their traffic through Web service APIs. Many different Web services can be invoked by a single internet application - for example, the meta search engine WebSifter [2] uses several online Ontologies to improve a user's query request into a much more meaningful query and then submits that query to various search engines. Such applications are called composite Web service.

Thus meta search engines aims in utilizing the traditional search engines to enhance the search results to retrieve links that are highly related to user query. The traditional engines have a drawback of using keyword matching techniques rather than semantic techniques. When the meta search engine uses such keyword based search engines they also suffer from the same drawbacks. Incorporating semantic technologies into meta search engines can yield highly relevant search results to users. The user may be a naïve user or an expert user. In case of naïve user, the meta search engine must be able to understand the expectations of user from the query terms. The search engine must not only search the exact query terms given by the user but also its proximity terms. This is possible only through semantic techniques. The semantic technique involves extraction of the most critical words or keywords from the user query and to find its synonymous terms and search by relevance to those keywords. For example the sentence1: "This car is grey in colour" is similar to the sentence2: "This automobile is grey in colour". Both sentence1 and sentence 2 gives the same meaning but they differ in usage of terms. The term "car" is synonymous to the term "automobile". They can be used interchangeably. Such terms can greatly influence the search results.

The existing traditional search engines do not retrieve the search results based on user expectations. The metasearch engine combines the search results of traditional search engines to produce the final search results page. Examples are Dogpile, WebCrawler, etc. But they do not interpret the user query semantically. The Meta search engines may need to decode query forms. Its translation of query syntax and fields are not exact. Other demerits include vulnerability to search spam and it lacks ways of comparing relevance scores. They refer to keyword based search engines thus does not include semantic relevance to the user query.

The results are unlimited thus time consuming in search of relevant searches. They display many older (visited most) links than appropriate links. All the links provided as search results may not be active currently. The objective of this work is to devise a metasearch engine for E-learning domain which utilises Google search engine to retrieve user query specific results by combining the techniques of ontology extraction and semantic analysis using WordNet tool to enable the students and researchers to browse the web with ease.

The rest of the paper is organised as follows. Section 2 contains the related works regarding the different meta search engines. Section 3 briefs about the existing systems and its disadvantages. Section 4 describes the proposed system and its architecture. Section 5 consists of the description of the Genetic Algorithm for the proposed system. Section 6 contains implementation results and

Section 7 contains the conclusion.

II. RELATED WORKS

This survey of various papers related to the proposed work gives more insight into the techniques and their usage. This survey also gives ideas that can be incorporated to the proposed to get valid and fail safe results. In the paper [3], they have proposed a metasearch engine in which the query given by the user is input to Wordnet ontology to obtain the neighbour keywords. Then by using semantic similarity measure refining of the input query is done and given to different search engine like Bing, Yahoo and Google. After getting the resultant web pages from the web, those pages are ranked using the ranking measure.

In the paper [7] they have devised a fully functional metasearch engine that creates personal user search spaces. Users can also define topics of interest. Personalization techniques used in this not only deal with the presentation style but also with the retrieval model and the ranking of the retrieved pages. Process can be increased to get precise results in shorter amount of time. It also gives details about the use of ontology in Semantic Web and ontology based retrieval techniques. It gives account of different information retrieval techniques used and also about the role of ontology in semantic web.

III. EXISTING SYSTEM

The existing metasearch engines combine the search results of traditional search engines to produce the final search results page. The working of Meta search engine is shown in figure 1 and described as follows. The user gives the input to the Meta search engine. The Meta search engine generally queries one or more search engines simultaneously for the user request. These search engines process the request from Meta search engine by utilizing their own processing techniques. The search results from these search engines are returned to the Meta search engine. The Meta search engine finally combines all the results returned and sorts them as required. Finally the resultant links are displayed to the user.

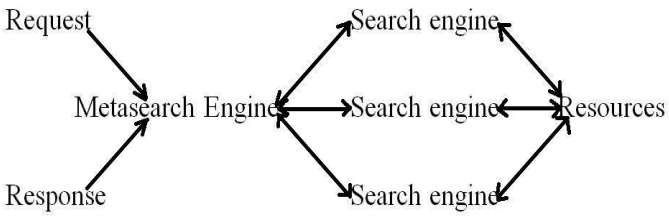

Fig. 1. The working of Meta search engine

In the paper [9], they have designed a semantic search engine- SIEU (Semantic Information Extraction in University Domain) confined to the university domain. The knowledge base for SIEU is constructed using the ontology. The query is analyzed both syntactically and semantically. It classifies that if it is a location based user query, if so, analysis is carried out based on specific keywords from the input query.

The paper [1], deals with SAVVYSEARCH, a meta search engine that learns to identify which search engines are most appropriate for certain queries, it reasons about the resources needed, and follows an iterative parallel search strategy as a simple plan. It maximizes the likelihood of returning good links. It minimizes computational and web resource consumption.

The paper [5] proposes a cooperative distributed multiagent system that locates and semantically integrates the access to heterogeneous distributed data sources, using Dublin Core as the metadata model.

The paper [8], provides a brief survey on the concept of Meta search engines, and is focused on the different architecture and their important technologies involved in it. Illustrates the key technologies of Meta search engines through which the performance of a Meta search engine can be determined. It summarizes some important Meta search engines that are developed.The paper [10], presents a brief survey on Information Retrieval (IR) technology and the evaluation of IR. Along with this, paper also deals with various approaches, how the Information Retrieval (IR)

The links extracted from the other search engines are by means of ranking algorithms. Different metasearch engines utilize different ranking algorithms to get the final result page. Most of them are based on the number of times the link has been visited. These metasearch engines also have a drawback of not using semantic analysis in processing the user query. The user query is matched with the tags of the resultant links and the links which matches most of the queried terms are returned back. Moreover the links retrieved by them may or may not be active currently. Thus it discomforts the users to browse many result pages to find an optimal link.

DISADVANTAGES

1. Most of the metasearch engines do not apply semantic analysis.
2. The results are based on the number of times the links have been visited.
3. The search results generated are unlimited and they are time consuming to browse all links to find a best link

IV. PROPOSED SYSTEM

A query specific meta-search engine is proposed for providing the most appropriate results for the user. The proposed method utilizes a set of queries instead of a single query. It is done with the help of WordNet ontology. The ontology is used for extracting the similar words for the user query. When the query set is created, then it is subjected to search in Google search engine and the top results are selected. The tags of those links are pre-processed to remove unwanted tags and stopwords and the keywords are extracted. The keywords are then ranked using Genetic algorithm and the terms with highest fitness are the ultimate keywords for searching. The final keywords are then passed to Google search engine to get the end results. The key offerings of this system are,

- A complete methodology is defined for automatic

knowledge extraction, in the form of ontological concepts, from a knowledge base of heterogeneous documents with the help of WordNet tool.

- The proposed method is implemented as an integrated system for E-learning purpose.

ADVANTAGES

1. The metasearch engine proposed results in retrieval of relevant results for user queries.

2. It does semantic analysis by using WordNet tool and ontologies are extracted.

3. The results are limited and highly appropriate.

4. The ranking algorithm is based on user specific query for the links to be retrieved.

5. The possibilities of dead links are minimized.

The architecture of the proposed system is given in the figure 2. The users provide their search input in the search box provided in the home page and get their resultant links in this page after the processing of the user‟s search query.

The search query of the user is passed to the WordNet tool to perform semantic analysis.The goal of using WordNet was to develop a system that would be consistent with the knowledge acquired over the years about how human beings process language. The WordNet tool finds the synonymous terms for each term of the user query.

For example : If the user has searched for a word „Query‟ it could give its synonyms „inquiry‟, „question‟. This helps to search the user query in broader sense to get apt results. The query terms are passed to the Google search engine. The top results of the search are retrieved. The tags and texts of the resultant links are extracted. Task in this phase is the pruning of irrelevant concepts from the extracted ontologies. A pruning strategy is adopted which advocates that frequent terms in a text denote domain concepts, while less frequent ones lead to concepts that can be safely eliminated from the ontology. The documents are converted from the original format to a more suitable one. It is the process of reducing a term of the analyzed document to its stem or root form. However, the stem does not need to be alike the morphological root of the term; it is usually sufficient that related words map to the same stem.

The Genetic Algorithm is used to find the relationship between the extracted terms from the links and the user query using the genetic algorithm. In a genetic algorithm, a population of candidate solutions to an optimization problem is evolved toward optimal solutions.

Each feasible candidate solution contains a set of properties which can be altered and mutated. The evolution generally begins from a population of randomly generated individuals, and is an iterative process. The population in the each iteration is called as generation. In each generation, every individual‟s fitness in the population is evaluated; the fitness is usually the value of the objective function in the optimization problem. The individuals which fit well are stochastically selected from the current population to form the new generation.

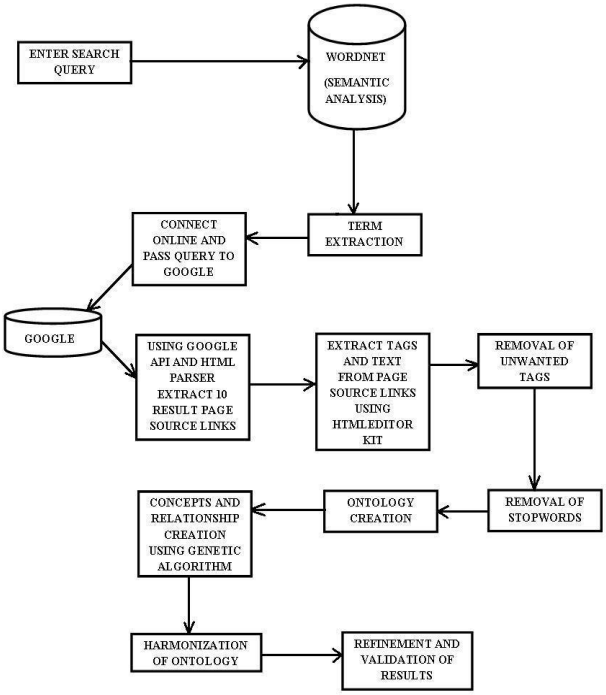

Fig.2. The architecture of the proposed system

The new generation formed is then used in the next iteration of the algorithm. Usually, the algorithm end when either a maximum number of generations has been reached, or an adequate fitness level has been produced for the population. The refinement phase deals with the tuning of the target resultant links from the terms generated by the Genetic algorithm. This refinement allows only links with high relative-ness to the user search query to be displayed as final output.

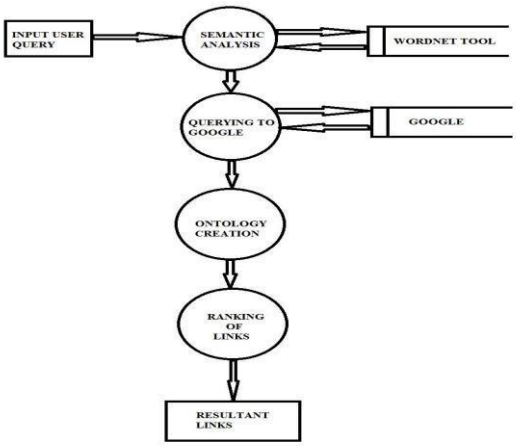

Fig. 3. The Data flow diagram for the proposed system

The generated terms are queried again in Google search engine to get the final result of links. These links are displayed as results to the user. A data-flow diagram (DFD) is a graphical representation of the "flow" of data through an information system. The DFD for the proposed system is shown in the figure 3. The WordNet tool and Google are the data stores of the system. The flow of data from the user input query to different process till the final result is visually shown.

V. GENETIC ALGORITHM

A Genetic algorithm (GA) is great for finding solutions for complex problems. They are used to schedule tasks, to design computer algorithms, and to solve other optimization problems. Here Genetic algorithm is used to find frequently used words in the retrieved link. It follows a tree based search. It enables to perform fast search and to find frequency of words. The genetic algorithm uses following parameters,

1. Initialization

This parameter creates an initial ontology. The words are usually generated randomly. The words can be of any desired size, from a few individuals to thousands.

2. Evaluation

Each word of the ontology is then evaluated to calculate a 'fitness' for that individual. A fitness function is a particular type of objective function that prescribes the optimality of a solution (keywords/terms) in a genetic algorithm, so that those particular terms may be ranked against all the other terms. Optimal terms, or at least terms which are more optimal, are allowed to crossover, producing a new generation that will be even better.

3. Selection

Overall fitness of the ontology is to be constantly improved. To attain this, selection helps in discarding bad terms and only keeping the best terms.

4. Crossover

During crossover new individuals are created by creating crosses of selected individuals called as parent terms. The idea is that, the combination of these parent terms will create an even 'fitter' offspring for the ontology which inherits the best bits of both individuals. Genetic Algorithm Pseudo code is given below,

Begin EA

t := 0; Init word P(t);

Evaluate P(t);

while not done do

t := t + 1;

P' := Selectparents P

(t); Recombine P' (t);

Mutate P' (t);

Evaluate P' (t);

P := Survive P,P' (t);

End

VI. IMPLEMENTATION RESULTS

This concept implemented requires operating system Windows 7. The Front-End Tool is JAVA (Jsp/Servlets) using IDE NetBeans 7.3. The Back-End Tool is MYSQL 5.5 and WordNet 2.1. The Web Server used is Apache Tomcat 7.0. The sample query term “Ontology” is given by the user. Its relevant words are retrieved from the WordNet tool. Then the links are retrieved from Google search engine as shown in the screenshot given in figure 4. After the retrieval of links, the HTML tags are removed. Later the Genetic Algorithm is used to find the most relevant query terms from the links acquired as shown in figure 5. Then those are queried in Google and resultant links are obtained which is shown in figure 6.

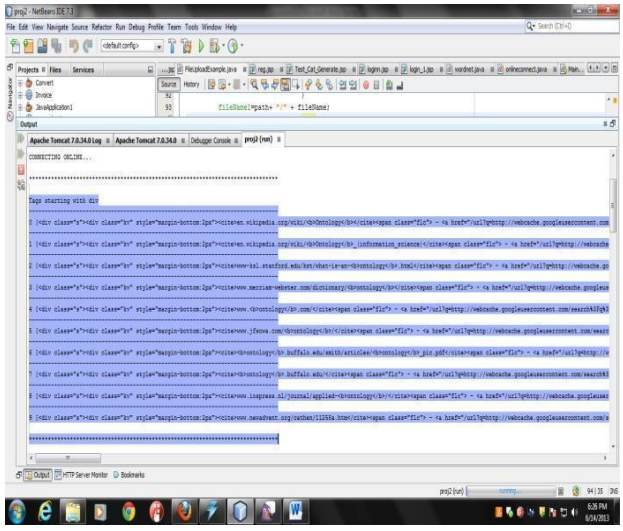

Fig. 4. Initial retrieval of links from Google.

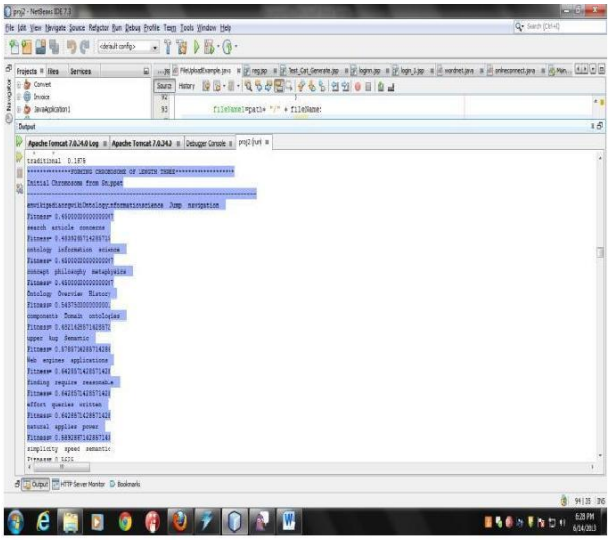

Fig. 5. Query term Optimization using Genetic Algorithm.

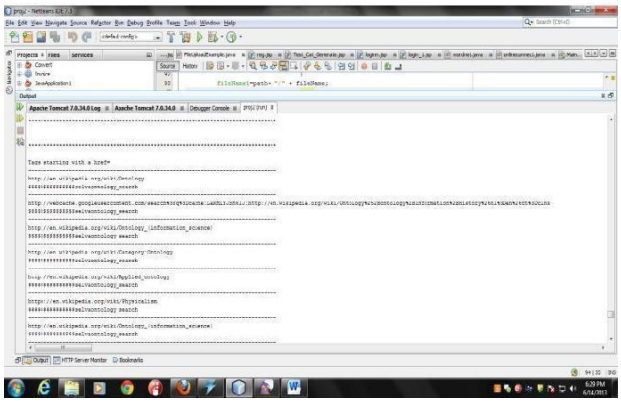

Fig. 6. The final result of links for the query term "Ontology".

VII. CONCLUSION

The Meta search engines have disadvantages of browsing several numbers of pages for a single relevant link due to unlimited search result page. This not only consumes the time resource of the user but also misleads the user from retrieving optimal results. Moreover these search engines do not implement semantic analysis of user query. Thus their results are not efficient as expected. The proposed Meta search engine for E-learning domain highly understands the requirements of the user by doing semantic learning using WordNet tool and by constructing ontology for the keywords. Thus the proposed system aims at satisfying the needs of the users by incorporating semantic analysis rather than syntactic analysis and produces limited relevant search results for the user convenience. The future enhancement for this work can include designing a better innovative user interface for the best use of its semantic search technology in Meta search engines.

References

[1] Adele E. Howe and Daniel Dreilinger, "SAVVYSEARCH: A Metasearch Engine That Learns Which Search Engines to Query" AI Magazine, American Association for Artificial Intelligence, Volume 18, Number 2, 1997, pp: 19-25

[2] Anthony Scime, Larry Kerschberg , "WebSifter: An Ontological Web-Mining Agent for E-Business" in Semantic Issues in E-Commerce Systems IFIP - The International Federation for Information Processing Volume 111, 2003, pp: 187-201.

[3] A.K.Mariappan, Dr.R.M.Suresh And Dr.V.Subbiah Bharathi, "Semantic Meta Search Engine Using Semantic Similarity Measure", Journal Of Theoretical And Applied Information Technology. Vol. 50, No.3, 2013.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[4] M. Arrigo, M. Gentile, D. Taibi and O. Di Giuseppe, "Specialized Search Engines for E-learning", Via Ugo la Malfa, 153 - 90146 Palermo, ITALY, 2005,pp: 1-5

[5] David F. Barrero, M. Dolores R-Moreno, Oscar García, Angel Moreno, "SEARCHY: A Metasearch Engine for Heterogeneous Sources in Distributed Environments" Proceedings of International Conference on Dublin Core and Metadata Applications, 2005, pp: 235-238.

[6] Manuel Rojas "A Semantic Association Page Rank Algorithm for Web Search Engines" Journal of Computing Research Repository, 2012, pp: 129-138.

[7] Stefanos Souldatos ,Theodore Dalamagas, Timos Sellis , " Sailing the Web with Captain Nemo: a Personalized Metasearch Engine" in W4: Learning in Web Search, at the 22 nd International Conference on Machine Learning, Bonn, Germany, 2005.

[8] G.Sudeepthi, Prof. M.Surendra Prasad Babu, "A Survey on Meta Search Engine in Semantic Web" International Journal of Computer Technology and Applications, Vol 2 (6), 2011, pp-3051-3055.

[9] Swathi Rajasurya , Tamizhamudhu Muralidharan , Sandhiya Devi, Dr.S.Swamynathan, " Semantic Information Retrieval Using Ontology In University Domain", International Journal of Web & Semantic Technology (IJWesT) Vol.3, No.4, October 2012.

[10] Vishal Jain, Dr. Mayank Singh, "Ontology Based Information Retrieval in Semantic Web: A Survey" International Journal of Information Technology and Computer Science, Vol no.10, 2013, pp: 62-69.

[11] http://en.wikipedia.org/wiki/Ontology_(information_science)