International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 3 Issue 6 ǁ June 2015 ǁ PP.01-05

Semantic Search Engine using Ontologies

Sumedh Pundkar[1], Kapil Baheti[2]
1(Computer, Usha Mittal Institute of Technology/ SNDT University, India)
2(Computer, Mukesh Patel School of Technology Management and Engineering/NMIMS University, India)

ABSTRACT: *Nowadays the volume of the information on the Web is increasing dramatically. Facilitating users to get useful information has become more and more important to information retrieval systems. While information retrieval technologies have been improved to some extent, users are not satisfied with the low precision and recall. With the emergence of the Semantic Web, this situation can be remarkably improved if machines could "understand" the content of web pages. The existing information retrieval technologies can be classified mainly into three classes.The traditional information retrieval technologies mostly based on the occurrence of words in documents. It is only limited to string matching. However, these technologies are of no use when a search is based on the meaning of words, rather than onwards themselves.Search engines limited to string matching and link analysis. The most widely used algorithms are the PageRank algorithm and the HITS algorithm. The PageRank algorithm is based on the number of other pages pointing to the Web page and the value of the pages pointing to it. Search engines like Google combine information retrieval techniques with PageRank. In contrast to the PageRank algorithm, the HITS algorithm employs a query dependent ranking technique. In addition to this, the HITS algorithm produces the authority and the hub score. The widespread availability of machine understandable information on the Semantic Web offers which some opportunities to improve traditional search. If machines could "understand" the content of web pages, searches with high precision and recall would be possible.*

Keywords <*ranking, search engine, searching algorithm, semantic search, time rank* >

I. INTRODUCTION

All over the world, people use search engines for some or the other work. Searching the web has become the part of our daily life. This includes everything from searching a food recipe to searching the latest trends in different technologies.

Though, searching the internet and user queries have increased but the satisfaction level of the users is still not up to the mark. Users still struggle to get the appropriate information on the internet. Getting the most accurate result for the searched query is a difficult task. Adding to the problem of the user, the number of results returned by the search engine are very large. It is practically impossible to go through all the links and get the answer.

The basic problems of the users include:

- Displaying the results which are not relevant
- Large number of results making difficult for the user to browse
- Fetching the results which are not authorized
- User is unaware of the logic used to fetch the results for the query making it difficult for user to analyze the results
- Low Precision
- Low Recall

These problems can be observed on any search engine.

For example, if the search query is technical related to programming, then the top results are some blogging website. There are several problems with information seeking on the Web. First, the Web is an open system which is constantly changing: new sites appear, old ones change or disappear, and in general the content is always randomly growing rather than planned. This implies that results are not stable and that users may need to vary their strategy over time to satisfy similar needs. Secondly, the quality of information on the Web is extremely variable and the user has to make a judgment. For example, if you submit the query "search engine tutorial" to any of the major search engines you will get many thousands of results. Even if you restrict yourselves to the top 10 ranked tutorials, the ranking provided by the search engine does not necessarily correlate with quality, since the presented tutorials may not have been peer reviewed by experts in a proper manner.

Thirdly, factual knowledge on the Web is not objective, so if the query is "who is the president of the United States" result may be several answers. In this case user may trust the White House web site to give a correct answer but other sites may not be so trustworthy. Finally, since the scope of the Web is not fixed, in

many cases we do not know in advance if the information is out there. There is uncertainty hanging in the air that does not diminish after we do not find what we are looking for during the first search attempt. For example, if user is looking for a book which may be out of print, can try several online second-hand book stores and possibly try and locate the publisher if their web site can be found. As another example, user may be looking for a research report and not know if it has been published on the Web. In this case, he/she may look up the author's home page, or the name of the report if it is known to, or alternatively, may try and find out if the institution at which the report was published maintains an online copy. In such cases, there will have to combine several strategies and several search sessions before user finds what he was looking for, or simply give up. The choice of appropriate strategy also depends on the user's expertise. Novice users often prefer to navigate from web portals which provide them with directory services. One reason for this is that their home page is by default set to that of their service provider, and being unaware of other search services they are content to use the portal put in front of them. Once a user learns to use search engines and becomes web savvy, he or she can mix the various strategies. As search engines make it easier to find web sites and pages with relevant information, more users are turning to web search engines as their primary information-seeking strategy. One interesting by-product of this shift to search engines is that users spend less time navigating within web sites and tend to jump from one site to another using the search results list as a guide from which to choose the next site to jump to.This makes the result less trust worthy for the user. Also, sometimes, user is not aware of the exact term needed to search. Thus, if exact keyword is not matching then the result may not be very accurate.Search engines must not restrict themselves to keyword match only. The semantics of the words must also be taken into consideration. The logic should be fuzzy. This paper compares existing systems providing such features and also proposes a developing mechanism for the search engine which will connect the information on existing web page with background ontological knowledge.The main aim of our Search Engine is to optimize the Information Retrieval system. Since there are Search Engines which retrieve redundant and unnecessary result for a given keyword; This Search Engine extracts synonyms of the entered query and displays the result related to only that query. The users need not to perform searching in an ad-hoc manner and waste time. It understands the user query and based on that shows the desired pages. It improves the searching mechanism and gives only relevant results to the user.

II. EXISTING METHODOLOGIES

A semantic web search engine implementation needs to deal with the following aspects: developing a fuzzy ontology, natural language processing and crawler. In the paper "Developing a Fuzzy Search Engine Based on Fuzzy Ontology and Semantic Search", the author discusses about constructing a two-layered fuzzyontology to organize terms that are elicited from WordNet[1]. WordNet is a large lexical database of English, built by Princeton University. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms, each expressing a distinct concept. In the two-layered fuzzy ontology, the first layer forms a domain hierarchy. Each domain contains a term lattice in the second layer.

Another technique that can be used to enhance semantic search is Natural Language Processing. The author "The HWS hybrid web search" discusses the use of an agent to process the natural language questions [2]. The agent employs LR (L means that the parser reads input text in one direction without backing up; that direction is typically Left to right within each line, and top to bottom across the lines of the full input file. The R means that the parser produces a reversed rightmost derivation) algorithm to complete the grammar parse for a given question. A given question is parsed to create a grammar tree which is then submitted to slot-filling. In contrast to the slot-filling of some nature languages, the agent employs slot-filling with grammar structure.

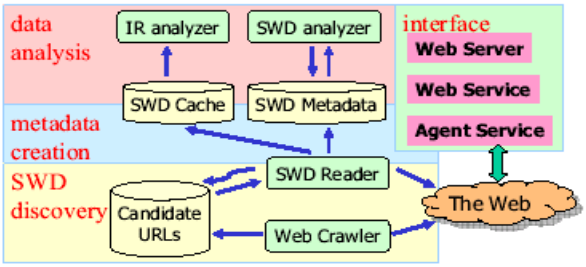

Fig. 2: Swoogle Search Approach

Thus, in addition to pattern match, a given question can be processed based on grammar parser. The agent then employs Brill's part-of-speech tagger to analyse the words in a given question. The agent deletes certain frequent words, acquired from the widely used WordNet, such as punctuation, preposition, article, or conjunction. It treats the rest of the question as keywords. In the meantime, the agent also employs WordNet to

identify phrases as keywords [2].Another algorithm is RK (RungeKutta) algorithm to find the words that are closely related to the entered keywords and their synonyms [2].

An important pre-processing step to searching is crawling. Crawler is a program that searches a World Wide Web typically in order to create an index of data. In [3], the authors begin the discussion of the first component required for building the index, and thus for retrieving the raw RDF documents from the Web: that is, the crawler [3].

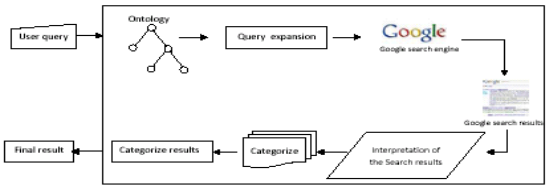

Fig.3: Semoogle Search Approach

The crawler starts with a set of seed URIs, retrieves the content of URIs, parses and writes content to disk in the form of quads, and recursively extracts new URIs for crawling. In this paper, the author discusses the architecture and implementation of the Semantic Web Search Engine (SWSE). Following traditional search engine architecture, SWSE consists of crawling, data enhancing, indexing and a user interface for search, browsing and retrieval of information; unlike traditional search engines, SWSE operates over RDF Web data – loosely also known as Linked Data – which implies unique challenges for the system design, architecture, algorithms, implementation and user interface.

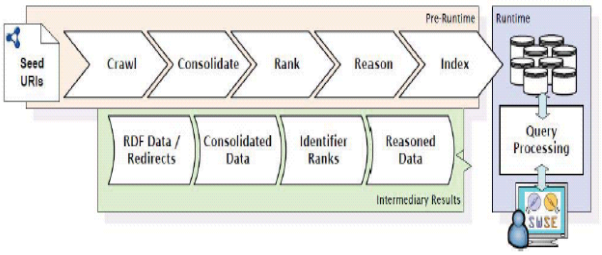

Fig.4: SWSE Search Approach

III. PROPOSED SYSTEM FOR SEMANTIC SEARCH ENGINE USING ONTOLOGIES

Search Engine that understands the meaning of the user query and relatively reasons him with the appropriate result is proposed. Not only the user entered keyword based pages would be returned but also the pages that is appropriate enough with the meaning of the user entered keyword.

User will be provided with facility to mark pages which would be displayed first the next time user enters the related term.

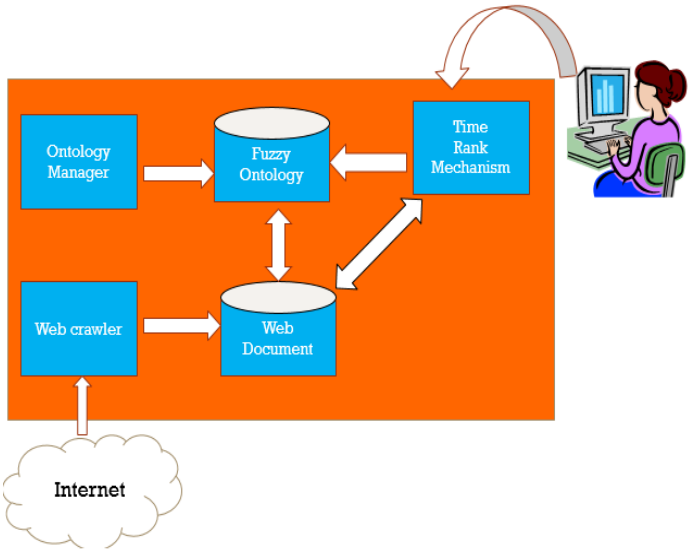

Fig. 2.5: Proposed System

Query Evaluator:

The Query Evaluator reduces each Semantic Web search query in an online step to a sequence of standard Web search queries on standard Web and annotation pages, which are then processed by a standard Web Search Engine, assuming standard Web and annotation pages are appropriately indexed. This block filters out the keywords from the user entered phrases and generates the synonyms to it. The Query Evaluator also collects the results and re-transforms them into a single answer which is returned to the user.

The search engine block takes the keywords from the query evaluator and checks it in the web document for the relevant pages which is returned to the inference system. Also, the annotations are used and algorithms are applied to generate the result.

Inference System:

Using background ontology Inference system adds all properties that can be deduced / induced from the ontology and returned to the web documents for other relevant pages.

Time Rank Mechanism:

A Time Rank Mechanism for ranking the pages which user searches can be implemented. This is a simple mechanism which ranks the pages based on the amount of time user has stayed on it previously. Higher the time, higher would be the rank of the page.

This mechanism would contain a database to store the time taken by the user to stay on the page.

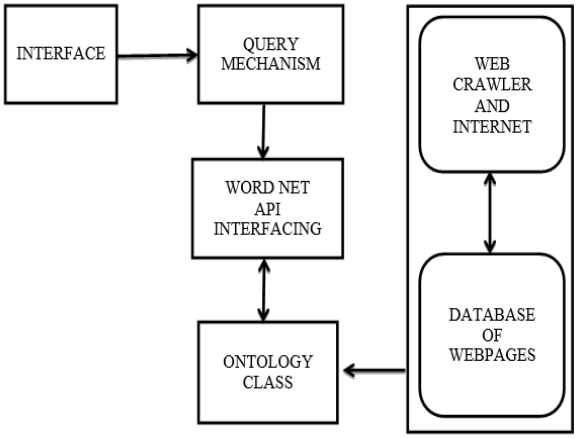

Fig. 2.6: Architecture of Proposed System

We have used Protégé to generate ontologies. Protégé is a free, open-source platform that provides a growing user community with a suite of tools to construct domain models and knowledge-based applications with ontologies.

Protégé Desktop is a feature rich ontology editing environment with full support for the OWL 2 Web Ontology Language, and direct in-memory connections to description logic reasoners.

Crawler 4j, an open source crawler was used for demonstration purpose.

IV. CONCLUSION

Semantic search on the Web, where standard Web pages are combined with background ontologies, on top of standard Web search engines and ontological inference technologies.

There is a formal model behind this approach. Generalized PageRank technique is used. Technique for processing semantic search queries [5] for the Web, consisting of an offline ontological inference step and an online reduction to standard Web search queries. Implementation in desktop search along with very promising experimental results is expected. This search engine with time rank algorithm will be implemented. This mechanism will not only rank the pages based on its importance, but also on the basis of time user spends on each page. As of now, for demonstration purposes, the entire search engine is offline with limited ontology and limited webpages obtained from a crawler in limited time. But on a greater scale, ontologies can be automated and crawler can be constantly updating the list of webpages adding new entries every second.

REFERENCES

[1] Lien-Fu Lai, Chao-Chin Wu, Pei-Ying Lin, "*Developing a Fuzzy Search Engine Based on Fuzzy Ontology and Semantic Search*", FUZZ-IEEE 2011,pp 2684-2689

[2] Lixin Han, Guihai Chen "*The HWS hybrid web search*", Information and Software Technology 48 (2006) 687–695

[3] Aidan Hogan , Andreas Harth , Jürgen Umbrich , Sheila Kinsella , Axel Polleres , Stefan Decker, "*Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine*", JWS, Volume 9, Issue 4, December 2011, Pages 365–401.
[4] Dwi H. Widyantoro, John Yen, "*A Fuzzy Ontology-based Abstract Search Engine and Its User Studies*" The 10th IEEE International Conference on Fuzzy Systems, 2001, pp 1291-1294.
[5] Thomas Lukasiewicz, "*Ontology Based Semantic Search on the Web*". Annals of Mathematics and Artificial Intelligence Volume 65, Issue 2-3 , pp 83-121
[6] Sumedh Pundkar, Kapil Baheti "*Survey on Semantic Search Engines using Ontologies*" IJARET Vol. 3, Issue 3, March 2015