Semantic and Keyword Based Web Techniques in Information Retrieval

Vajenti Mala
School of Computer and Systems Sciences
Jawaharlal Nehru University
New Delhi, India
er.vajenti@gmail.com

D.K.Lobiyal
School of Computer and Systems Sciences
Jawaharlal Nehru University
New Delhi, India
lobiyal@gmail.com

***Abstract*—Semantic and keyword web based technique is becoming a generic issue in an application of Information Retrieval (IR). Most of the researchers used different web techniques for finding relevant information and find the keyword based search, which are not able to fetch the relevant search result because they do not know the actual meaning of the term or expression and relationship between them in the web search. In this paper, semantic and keyword based web search method have been applied on the different web search engines. The selected search engines such as semantic search engines (Google, Yahoo, Wikipedia) and keyword search engines (Hakia, Bing, DuckDuckGo). Performance is based on their precision ratio and natural language queries. Various queries was input on different search engine and output of the documents was classified a relevant documents and non-relevant documents. Precision ratios were calculated in the final retrieved documents on each web search engines. Also defined some popular semantic and keyword search engine features.**

Keywords—,Information Retrieval, Semantic Web search, keyword web search, web search engine

I. INTRODUCTION

There are many techniques in Information Retrieval (IR) to retrieve information from documents but IR techniques are responsible for tackling annotation in semantic and keyword web languages. With the huge amount of information available on web which may be in form of structured, unstructured or semi structure. Therefore, it is difficult to find out of identifying the relevant information from search engine. Search engine has greatly impacted in the area of information retrieval, moreover, most of the web users cannot be search the results which they need. Normal keyword based web is not in the position to provide the exactly search result to the user. In this situation, we need semantically web search engine.

A. Semantic Based Web Search Method s (SBWSM)

Semantic web is a web where information represented in the process of machine learning [1]. The documents on the web are represented as HTML form, RDF (Resources Description Framwork), and OWL (Web Ontology Language) is used for semantic web based documents. it can be search accuracy as well as understanding the trems as they appears in the searchable databases such that media objects(web pages ,images and audio films). Moreover, semantic web contains single kind of relationships (hyperlinks) between the resources and also different kinds of other resources which is mentioned in [14]. Semantic search engines are Hakia,DuckDuckGo etc. Semantic web search store all information in semantically form it solve the complex queries on the web.

B. Keyword Based Web Search Method (KBWSM)

Keyword web search engine is very helpful for finding information on the internet. It suffers the m,eaning of some terms and expression which is used in the webpages. Currently keyword web based approach has reached a plateau. In the literature surveys 25% of web searches do not give the accurate results because they return the results in the first URL_s and daily sixty-terabyte increase in the size of the web [15]. This search approach, queries are very sensitive and search words often have multiple meanings.

In this paper, we have divided into different sections which are mentioned given below. The remaining section is determined by the related work. Section III is discussed about the methodology that how the work has been performed with the keyword based and semantic based approaches. Section IV is discussed about the experiments performed by MATLAB 2010b. In the final section, conclusion points such as limitation and advantage as well as future work.

II. RELATED WORK

A good search engine can be selected by their performance and effective results and effective results can be measured by their precision and recall. Here we have explored the different literature survey related to the field.

Albertoni R et al [4] explained about the semantic web and visualization of information.

D. Tumer et al [5] has determined an empirical evaluation on Semantic approach and search performance of Keyword-

ISBN: 978-1-5090-1666-2/16/$31.00 ©2016 IEEE

Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia etc.

M. Andago et al [6], selected an evaluation of a Semantic Search Engine and compared with a Keyword Search Engine by using pre-defined formula "first 20 Precision". He has selected 30 queries and entered into the search engines and calculated precision ratios. The Google outperforms with Hakia. Because it has higher precision at 0.64 compared with Hakia. It was clear that Google is superior to Hakia.

III. METHODOLOGY

When work out an evaluation of web based environment, we have taken six web search engines, three are keyword based named Google, Yahoo, Wikipedia while others are semantic web based named Bing, Hakia and DuckDuckGo. We have selected these web engines for relevant information by using precision ratio formula which is the most important for effective measurements. However, ten queries have been taken from various subjects contains one or two terms randomly for determine effective measurements shown in the Table I.

When run a query, top twenty documents will be retrieved and were evaluated using human relevance judgment. After that every document may classified by "Relevant" and "non-Relevant documents. In this study, we have collected queries from medical domain such as cancer, eye disease and blood sugar etc. shown in the table I query list. While an evaluation of web search engine by user's effort measures. The input queries entered into keyword and semantic search engines. After that precision was determined by pre-defined formula. The calculation was performed on six search engines, Google, Yahoo, Wikipedia, Bing and DuckDuckGo. The performance of Google which is keyword search engine has higher mean precision compared with Hakia which is semantic engine in terms of the first 20 precision.

The overall Hakia performance is 75%age which is lowest than Google. Google performance is higher 82.8 %age.by comparing different search engine keyword search has best result provided. In this paper we have different ten queries which provide the performance comparison of each search engine. The result shows that Google has efficient result outcome when compared to the other search engines. On the other side semantic search engine the Hakia is lowest result than other semantic search engine. Bing stayed second after DuckduckGo. This experiment shows that keyword base search engine produced efficient result when compared with semantic search based engine.

A. Measuring Search Effectiveness

After finishing a search process and found most relevant documents we check the effectiveness measures by using Recall and Precision techniques. They are the basic measures through which we can determine the search strategies. Results can be measures by relevant and irrelevant Documents as shown in equation: 1 and 2.

- *Precision:* - Precision is the ratio of Documents with respect to total relevant documents retrieved to the irrelevant documents.

$$\text{Precision} = \frac{\text{Doc}_{\text{Total\_Relevent}}}{\text{Doc}_{\text{Total\_Retrieved}}} \ldots(1)$$

- *Recall:* - Recall is the ratio of documents with respect to retrieved relevant document to the possible relevant documents.

$$\text{Recall} = \frac{\text{Doc}_{\text{Retrieved\_Relevent}}}{\text{Doc}_{\text{Possible\_Relevent}}} \ldots(2)$$

Following formula is pre-defined for first 20 precision.

$$\text{First\_20\_Precision} = \frac{\sum_{i=1}^{n=1} Score_i}{Doc_{\text{Rel}} \times SE_N} \ldots(3)$$

The above pre-defined formula Eq. (3) has been performed in the final result of relevant retrieved documents.

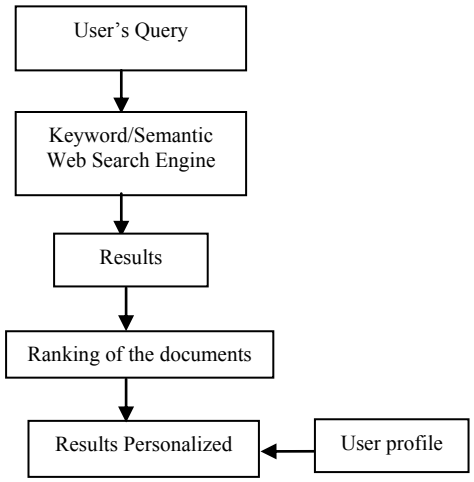

Fig: - 1 Process of crawling and Ranking

IV. EXPERIMENTAL RESULTS

For checking effectiveness efficiency of the approaches, we simulated in MATLAB 2010b. In this experiment, we have 6 search engines, 3 are keyword based and 3 are semantic based having 20 documents of each search engines. 10 queries were taking from medical domain. The simulation result shows in fig 3 and 4 that Google which is keyword based web has maximum no: of relevant documents retrieved then Hakia which is semantic based

web. All search engines provide the result with respect to the user query which may occur on more than two terms.

Table I: Input query list from Medical domain

Query no	Query/Keyword	Query no	Query/Keyword
q1	Cancer	q6	Brain Tumor
q2	Eye Diseases	q7	Animal Disease
q3	Blood Sugar	q8	Heart Attack
q4	Breast Cancer	q9	X-Ray
q5	Dengue	q10	Root Canal treatment

Query/SE	G	Y	W	B	H	D
q1	17	16	20	10	16	17
q2	17	13	5	17	14	16
q3	14	17	12	12	9	17
q4	14	19	13	10	8	12
q5	11	8	15	13	11	7
q6	15	9	11	8	10	9
q7	10	12	17	14	12	14
q8	6	12	18	16	15	14
q9	20	19	19	20	19	20
q10	14	12	10	15	11	13
Total	138	137	140	135	125	139
Average	82.8	82.2	84	81	75	83.4

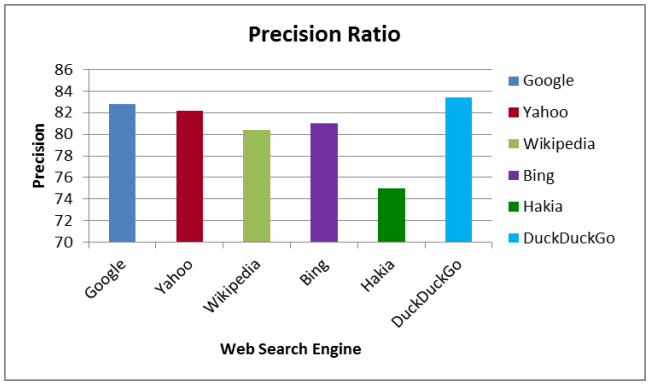

Fig:2 Precision Ratio of different Search Engines

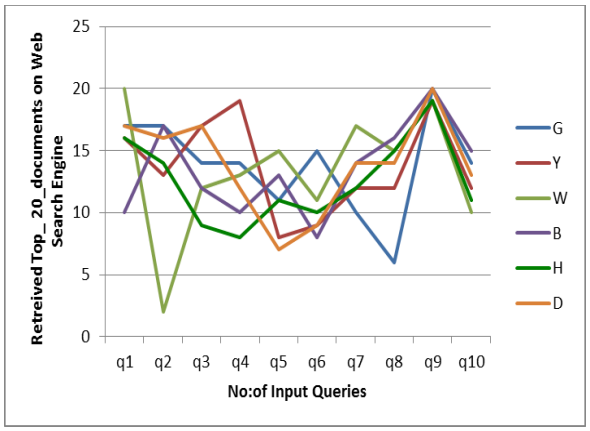

Fig:3 Relevant Retrieved Documents

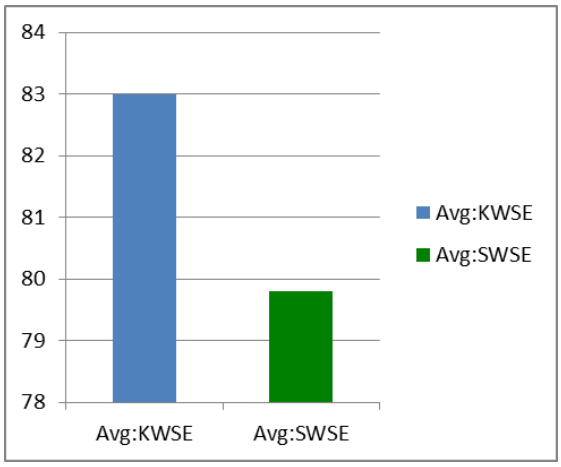

Fig:4 Average of Keyword and Semantic SE

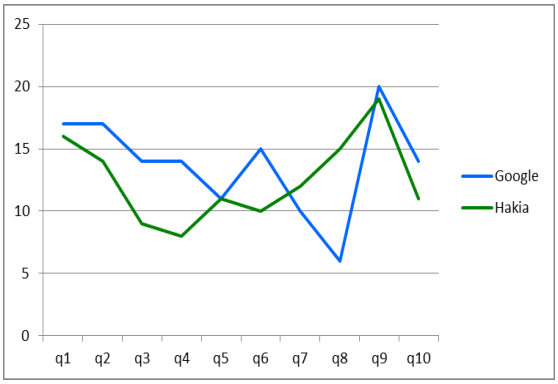

Fig:5 Visual Representation Min and Max ratio of SE

V. CONCLUSION

In information retrieval technique, Semantic web has greatly impacted in the world information technology.

In this paper, we discussed about methods of web search engines, a comparative performance of semantic and keyword. In overall study, Google and DuckDuckGo has perform better result for relevant document and analysis of various features of semantic and keyword search engines. For measuring efficiency, we have used precision and recall

A. Limitations and Advantages

For finding the information from search engines, conventional web is very important for extracting information on the internet and save the time, but they suffer the meaning of the terms and expression on the web pages and the relationships among them, because the URL's does not fetch results in first set of URL'. The following points are limitations.

- Polysemy words which means one word having several meaning such that "BANK" it may be a finance department or river shore.
- Synonymy words which means several words having same meaning such that "BABY" and "INFANT" are treated as synonymy in most of the thesaurus.
- In tradition information retrieval technology, most of the words in a document based on purely on the occurrence of the words in documents. The application of the semantic web is improving the traditional information retrieval web search.

B. Future Work

In information retrieval techniques, web based approaches can be done widely in the area of web search engines.in the future work we study or research on maximum no of queries and experiment will be rich and very helpful in terms of different search engines.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval," ACM Press/Addison Wesley,1999.

[2] T.Berners-Lee,J.Hendler and O.Lassila, "the Semantic Web."Scientific American, 2001.

[3] M. Tang and Y. Sun, "Evaluation of Web-Based Search Engines Using User Effort Measures," Library and Information Science Research Electronic Journal 13(2), 2003.

[4] Albertoni R., Bertone A., & De Martino M., (2004), "Semantic Web and Information Visualization, Proceedings of the1st Italian Workshop on Semantic Web Application and Perspective, DEIT, pp.108-114,Ancona, Italy,December10, 2004.

[5] D. Tumer, M. A. Shah and Y. Bitirim "An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia," Fourth International Conference on Internet Monitoring and Protection, 2009.

[6] M. Andago, P.L Phoebe and A. M Thanoun, "Evaluation of a Semantic Search Engine against a Keyword Search Engine Using First 20 Precision," In Proceedings of the 29th Annual International Conference on Research and Development in Information Retrieval, ACM Press, pp. 735–746, 2010.

[7] Lou Junwei, Xiao, "Resaerch on Information Retrieval System based on Semantic Web and multi-agent",ICICCI,IEEE 2010.

[8] Qing Chen, " Towords Web-based Information Retrieval in Grid Envirnoment", IEEE 2010

[9] Guha,R.McCool and Miler, "Semantic Search," 2011.

[10] A.Gulli and A. Signorini, "the indux able Web is more than 11.5 billion pages," 2013.

[11] K.Varsha,P.Sandhya and K.Supreet, "Information Retrieval: Today and Tomorrow" IJCA, Vol.116, 2015.

[12] Hakia search engine availabe at :http://www.hakia.com/

[13] DuckDuckGo available at:htt://duckduckgo.com/

[14] http://www.w3.org/TR/2004/REC-rdf-primer-20040210, 2004.

[15] W. Roush, "Search beyond Google," Technology Review, 2004.