ISSN:0975-9646
Rajeev Kumar et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (2) , 2012,3655-3658

A Detailed Study on Semantic Search Performance of Keyword and Meta Search Engines

Rajeev Kumar[1], Rajat Goel[2.]
[1]*M.Tech Student, Department of Computer Science & Engineering, IIMT College of Engineering, Meerut*
[2] *Department of Computer Science & Engineering, Translam Institute of Technology & Management, Meerut*

***Abstract*– With the development of the Web, an information "Big Bang" has taken place on the Internet. The continued rapid growth in information volume makes it increasingly difficult to find, organize, access and maintain the information required by users. Today, when we use a web search engine, the search engine can't tell if the web page is actually relevant for our search. This research paper critically analyzes the performance of popular search engines based on distinct technologies as, Semantic Search, Keyword Search & Meta Search. We have used three keyword based search engines (*Google, Bing, & Yahoo*), Meta Search Engine (*Dogpile*), and Semantic Search Engine (*Hakia*). The queries are used on every search engine to judge the performance of the search engine on the basis of the relevancy of the results returned by the search engine. The first twenty documents on each retrieval output were used to judge the performance of search engines on different criteria such as precision ration & normalized recall ratio.**

***Keywords*: Semantic Web, search engine, web page summarization, web content, Semantic search engine, Keyword based search engine, Evaluation, Meta Search Engine.**

I. INTRODUCTION

Web search engines work by storing information about many web pages, which they retrieve from the html itself. These pages are retrieved by a Web crawler (sometimes also known as a spider) — an automated Web browser which follows every link on the site. Exclusions can be made by the use of robots.txt. The contents of each page are then analyzed to determine how it should be indexed (for example, words are extracted from the titles, headings, or special fields called meta tags). Data about web pages are stored in an index database for use in later queries. A query can be a single word. The purpose of an index is to allow information to be found as quickly as possible.

When a user enters a query into a search engine (typically by using key words), the engine examines its index and provides a listing of best-matching web pages according to its criteria, usually with a short summary containing the document's title and sometimes parts of the text. The index is built from the information stored with the data and the method by which the information is indexed. Unfortunately, there are currently no known public search engines that allow documents to be searched by date. Most search engines support the use of the boolean operators AND, OR and NOT to further specify the search query. Boolean operators are for literal searches that allow the user to refine and extend the terms of the search.

The most popular search engines are Google[1], Yahoo[2] and Bing[3] with 65.45%, 14.79% and 13.27% volume of search ratios (based on US Internet usage), respectively[4]. Google, Yahoo and Bing are keyword-based search engines. Dogpile is the popular Meta Search engine in the series of distinct technology based search engines on the web. However, semantic search engines are an alternative to these search engines. The difference of semantic search engines from conventional search engines is that the semantic search engines are meaning-based. Hakia[5] is one of the emerging publicly available semantic search engines on the web[6].

Perfect search engine model might be the one that always finds the precise document(s) on the web for the user. The result of a perfect search engine would, ideally, satisfy the expectations of its users, whenever a query is searched. The inspiration, for this study, is to motivate researchers and search engine providers towards reaching this perfect search engine model.

Ten queries, from various topics, and four phrases, having different syntax but similar meanings, were determined. The queries are listed in the following table (*Table 1: Qurery List*).

Table 1: Query List

Query Number	Query	Query Number	Query
Q1	zodiac signs	Q6	Movie
Q2	criminal trials	Q7	Facebook
Q3	faithful dogs	Q8	ebay
Q4	earthquake	Q9	reincarnation
Q5	iPad	Q10	nightmare

This paper is organized as follows: Section 2 describes the methodology employed to evaluate search engines in terms of precision and normalized recall, Section 3 reports and discusses the experimental findings and the last section concludes the paper.

II. METHODOLOGY

Initially, three keyword-based search engines, namely, Google, Yahoo and Bing, a Meta search engine Dogpile, and a semantic search engine Hakia were selected. Afterwards, ten queries that contain various topics and consist of one or two terms[7] were randomly determined as shown in Table 1. (Note that for some queries, Hakia displays categorized documents in its retrieval output before displaying web results. In order to have compatible retrieval outputs for search engine evaluations, during query selection process, particularly those queries were used that provide web results in the retrieval output, without the categorized documents, when run on Hakia). Ten queries were run on, all the three, the selected keyword-based search engines, Meta Search engine as well as the semantic search engine. Additionally, four phrases, with different syntax but similar meanings, were used with each query and run, one by one, on the semantic search engine Hakia. The phrase-with-query (PwQ) forms were as follows: (1) "*what is* <query>"; (2) "*information about* <query>"; (3) "<query> *definition*"; and (4) "*description of* <query>".

In order to have realistic results:

• Only keywords were used on keyword-based search engines since, in general, the users do not tend to use phrases (as observed in the most frequently used queries list of Wordtracker[8]).

• Beside the keywords, phrases were used for Hakia since one of the main features of Hakia, being a semantic search engine, is the use of phrases.information.

After each run of the query or PwQ, the first twenty[9] documents retrieved were evaluated using binary human relevance judgment and with this, every document was classified as "relevant" or "nonrelevant". Total 1800 documents were evaluated by the same author and in order to have stable performance measurement of search engines, all the searches and evaluations were performed in minimal non-distant time slots. While evaluating the retrieved documents following criteria were used:

(1) Documents that contain any explanation about the searched query were considered "relevant";

(2) documents having same content but originating from different web addresses (i.e., mirror pages) were classified to be different[10];

(3) in case of duplicated documents, the first document that was retrieved was considered in the evaluation process, whereas its duplicates were classified to be "nonrelevant"[10]; and

(4) if, for some reason, a retrieved document became inaccessible, it was classified to be "non-relevant" [10].

Precision and normalized recall ratios of keyword-based search engines were calculated at various cut-off points (first 5, 10, 15 and 20 documents retrieved) for each pair of query and search engine. Furthermore, these calculations were performed for each pair of query and Hakia, and for each pair of PwQ and Hakia. However, for each query topic, the average precision and normalized recall ratios of the query and four PwQs was considered in evaluating the semantic search engine Hakia.

Precision is defined as the ratio of the number of relevant documents retrieved to the number of total documents retrieved [11]. As such, using precision at different cut-off points is helpful in estimating the distribution of relevant documents over their ranks [10].

The score-rank curve is strongly associated with the normalized recall, say R_{norm} [12]. R_{norm} is based on optimized expected search length [13].

Hence, normalized recall considers Δ_1 to be better than Δ_2 if Δ_1 provides fewer non-relevant documents; here Δ_1 and Δ_2 are two different retrieval outputs. In this study, Rnorm was calculated at four cut-off points (cut-off 5, cut-off 10, cut-off 15 and cut-off 20) in order to get values parallel to those for precision. Formula 1, proposed by Bollmann et al.[14], was used to calculate normalized recall values at various cut-off points.

$$R_{norm}(\Delta) = \frac{1}{2}\left(1 + \frac{R^+ - R^-}{R^+_{\max}}\right) \qquad (1)$$

Where R^+ is the number of document pairs that have relevant documents ranked higher than non-relevant documents, R^- is the number of document pairs that have non-relevant documents ranked higher than relevant ones, and $\mathbf{R}^+_{\max}$ gives the maximum number of R^+ [10].

III. EXPERIMENT RESULTS

Retrieval performance of search engines can be evaluated using the number of zero retrievals (i.e., no documents retrieved) or retrievals containing no relevant documents (i.e., the precision ratio is zero) [12]. The number of relevant documents retrieved by each search engine for the first twenty documents retrieved is shown in Table 2.

While the queries and PwQs ran on the search engines, the expectation was to retrieve documents that contain an explanation regarding the queries. Google, Yahoo, Bing and Dogpile retrieved at least one relevant document for all queries; however, Hakia[3] failed to retrieve any relevant document for one of the queries ("information about faithful dogs"). Yahoo retrieved more relevant documents and Bing retrieved least number of relevant documents than other keyword-based search engines. Yahoo and Hakia[1] retrieved approximately the same number of relevant documents in total. However, Hakia[2] and Hakia[4] retrieved more relevant documents than Yahoo, and Hakia[5] retrieved more relevant documents than Google and Bing. Google, Yahoo and Bing retrieved approximately 75.5%, 63% and 78% non-relevant documents, respectively, and, for the original query and all PwQs, Hakia retrieved 62.5%, 59.5%, 84%, 51.5% and 74% non-relevant documents,respectively. Although Hakia[3] retrieved least number of relevant documents in total, Hakia[4] retrieved more relevant documents than all others.

Query Number	Google	Yahoo	Bing	Dogpile	Hakia[1]	Hakia[2]	Hakia[3]	Hakia[4]	Hakia[5]
Q1	5	9	6	8	7	9	1	6	3
Q2	3	12	4	10	4	4	7	7	9
Q3	2	2	1	2	2	6	0	10	9
Q4	5	6	3	5	7	12	3	15	4
Q5	9	10	2	8	7	6	4	6	1
Q6	5	5	5	5	7	6	3	9	2
Q7	2	1	2	2	7	8	1	11	4
Q8	2	9	3	8	11	9	2	9	2
Q9	10	13	9	11	16	15	7	12	13
Q10	6	7	9	8	7	6	4	12	5
Total	**49**	**74**	**44**	**67**	**75**	**81**	**32**	**97**	**52**
Avg (%)	**24.5**	**37**	**22**	**33.5**	**37.5**	**40.5**	**16**	**48.5**	**26**

[1]: The original query is used; [2]: "what is <query>" is used; [3]: "information about <query> " is used;
[4]: "<query> definition" is used; [5]: "description of <query>" is used.

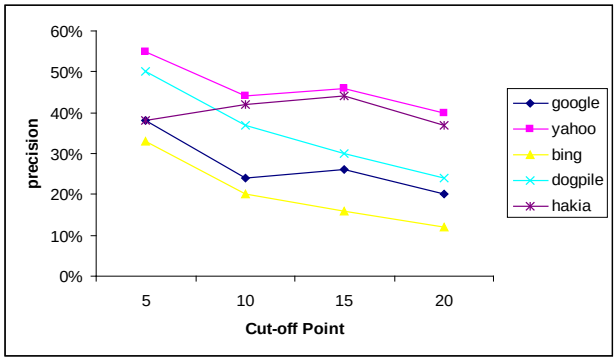

Figure1. Mean precision ratios of search engines

Mean precision ratios of keyword-based search engines, meat search engine and the semantic search engine at various cutoff points (for first 5, 10, 15, and 20 documents retrieved) are shown in Figure 1. Google's precision ratio is the same with Hakia at cut-off point 5 (38%).

With increase in cut-off point value, Google's precision ratios decreased for all cut-off points. Hakia's precision ratio slightly increased at cut-off point 10, then decreased slightly at cut-off points 15 and 20. Furthermore, Google retrieved more relevant documents than Bing for all cut-off points with approximately 3% difference. Although Yahoo retrieved approximately 4% more relevant documents than Hakia at cut-off points 10, 15 and 20, it retrieved the highest number of relevant documents at all cutoff points and its best precision ratio was 48% at cutoff point 5. However, Bing retrieved the least number of relevant documents at all cut-off points. Generally, precision ratios of search engines decreased when the cut-off point values were increased.

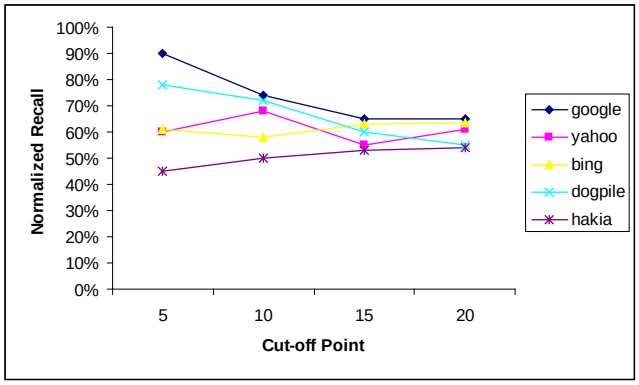

Figure2. Mean normalized recall ratios of search engines

Figure2 shows mean normalized recall ratios of keyword-based search engines, meta search engine and those of the semantic search engine at various cut-off points. Yahoo's normalized recall ratio was approximately the same as that of Bing at cut-off point 5. However, for cut-off points 10, 15 and 20, while the normalized recall ratio for either of these search engines increased, the other search engine's normalized recall ratio decreased. Normalized recall ratios of Google and Bing were approximately the same at cut-off point 15 and the difference between Google, Yahoo and Bing was about 1% at cut-off point 20. Google had the highest performance at cut-off point 5 (88%) but when the cut-off point increased to 10 and 15, Google's normalized recall ratio decreased. However, Google had the same normalized recall ratio at cut-off points 15 and 20. At all cut-off points, Google had the highest performance for displaying relevant documents retrieved in the top ranks of the retrieval output. Although Hakia's normalized recall ratio increased gradually at all cut-off points, Hakia had the least performance for displaying relevant documents retrieved in the top ranks of the retrieval output.

IV. CONCLUSION AND FUTURE WORK

In this paper, an investigative evaluation on search performance of keyword-based, meta search engine and semantic search engines is detailed. It was found that Google, Yahoo and Bing retrieved at least one relevant document for all queries, whereas Hakia[3] failed to retrieve any relevant document for one of the queries. Although Hakia[4] retrieved more relevant documents compared to other search engines, Hakia[3] retrieved least relevant documents. In terms of overall performance, Hakia retrieved more relevant documents compared to Google and Bing at all cut-off points. However, Yahoo retrieved the highest number of relevant documents at all cutoff points with its best precision ratio being 48% at cut-off point 5. Google showed the highest performance for displaying relevant documents in the top ranks of the retrieval output at all cut-off points. Yahoo and Bing come next, while Hakia showed the least performance for displaying relevant documents retrieved in the top ranks of the retrieval output. Generally, precision ratios of search engines decreased with increased cut-off point values. However, it was seen from the results that the performance of search engines, when displaying relevant documents in the top ranks, is better than their relevant document retrieval. Finally, it was seen that semantic search performance of search engines was low regardless of the type of the search engine used. Therefore, search engines need to improve their systems, taking into consideration the importance of the role semantic search can play in helping users getting precise information from the web with minimal effort. As a future work, the most frequently used queries and phrases could be run on search engines. Furthermore, the number of search engines, queries and phrases could be increased. In addition, elaborate statistical analysis could be provided.

V. REFERENCES

[1]. J.Pollock: Semantic Web for Dummies, 2009
[2]. G.Antoniu and F.van Harmelen: Semantic Web Primer, 2nd edition in 2008.
[3]. D.Allemang and J.Hendler: Semantic Web for the Working Ontologist,2008.
[4]. P.Hitzler,R.Sebastian, M.Krotzsch: Foundation of Semantic Web Technologies, 2009.
[5]. Google, http://www.google.com (Accessed on 09 March 2012)
[6]. Yahoo, http://www.yahoo.com (Accessed on 09 March 2012)
[7]. Bing, http://www.bing.com (Accessed on 09 March 2012)
[8]. Hitwise,"Hitwise United States >>- the power of competitive intelligence", March 2012. http://www.hitwise.com/us/datacenter/main/dashboard-23984.html (Accessed on 09 March 2012)
[9]. Hakia, http://www.hakia.com (Accessed on 09 March 2012)
[10]. C. Müller, T. Zesch, M. Müller, D. Bernhard, K.Ignatova, I. Gurevych and M. Mühlhäuser, "Flexible UIMA Components for Information

Retrieval Research", 6th International Conference on Language Resources and Evaluation, pp. 24-27, Marrakech, Morocco, May 2008.

[11]. B. J. Jansen and A. Spink, "Sponsored Search: Is Money a Motivator for Providing Relevant Results?", IEEE Computer, vol. 40, no. 8, pp. 52-57, August 2007.

[12]. Wordtracker, "The Top 200 Long-Term Keyword Report", The Wordtracker Report, 05 February 2008.

[13]. A. Spink and B. J. Jansen, "Searching Multimedia Federated Content Web Collections", Online Information Review, vol. 30, no. 5, pp. 485-495, 2006.

[14]. Y. Bitirim, Y. Tonta and H. Sever, "Information Retrieval Effectiveness of Turkish Search Engines",Advances in Information Systems, Lecture Notes in Computer Science, T. Yakhno (Ed.), vol. 2457, pp. 93-103, Springer-Verlag, Heidelberg, October 2002.

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

[16]. Y. Y. Yao, "Measuring Retrieval Effectiveness Based on User Preference of Documents", Journal of the American Society for Information Science, vol. 46, no. 2, pp. 133-145,1995.

[17]. W. S. Cooper, "Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems", American Documentation, vol. 19, no. 1, pp. 30-41, 1968.

[18]. P. Bollmann, R. Jochum, U. Reiner, V. Weissmann and H. Zuse, "Planung und Durchführung der Retrievaltests",Leistungsbewertung von Information Retrieval Verfahren,H. Scheider (Ed.), pp. 183-212, Fachbereich Informatik, Technische Universitat Berlin, Germany, 1986.