

# An Approach For Evaluation Of Semantic Performance Of Search Engines: Google, Yahoo, Msn And Hakia

**Marwan S. M. Aldabbagh**

Department of Computer Sciences / College of Education  
University of Mosul

**Received**  
**11 / 04 / 2011**

**Accepted**  
**27 / 06 / 2011**

.Hakia MSN Google Yahoo

Hakia

Google Yahoo

Hakia

MSN

" "

" "

" "

" "

## ABSTRACT

Ten queries of different subjects were run over yahoo, Google, Msn and Hakia search engines. Second set of experiment comprise four phrases of the same meaning but different syntax were attached to each query and the combinations were run over Hakia search engine. Four thousand retrieved documents were classified as relevant, duplicate different source, non – relevant, duplicate same source and inaccessible

documents. Calculation of relative precision of the two sets of data, show maximum relative precision for Hakia search engine followed by exchange places search engines of Yahoo and Google, where the lowest relevant precision is shown by Msn. For the second set of data, phrase “information about“ with query shown maximum relevant precision followed by phrase “what is“ where as phrase “description of“ and “definition“ show lowest relevant precision. Comparison of web document retrieval performance were accounted for in the discussion.

## **1. Introduction**

The present study deals with four most visited search engines namely: Yahoo, Google, Msn and Hakia. The first three engines are keyword – based engine, where as the last, Hakia, is semantic meaning – based search engine [1]. It is well known that different search engines are in fact a computer program layed down for private or public search and retrieve web documents that meet the user needs [2]. Variation is noticed in the frequency of uses of the four– mentioned search engines. World wide uses show that Yahoo ratio is (8.34%), Google (79.77%), Msn (8.67%) and others (3.27%), where as for Asia: Yahoo (12.23%), Google (77.24%), Msn (3.43%) and others (7.1%). Hakia is one of few public semantic search engines developed by previous studies, where as, NAGA semantic search engines is not available for public uses. It is worth mention here that NAGA was developed by Kasneci et al.[3], and Lee et al.[4], proposed semantic system for better retrieval effectiveness. It seems that the subject of search engines currently gain importance and by now represent a well defined trend of research within the discipline of information technology [5]. Other efforts were directed to performance comparative studies of the different search engines intended to motivate researchers and provider of search engines for further improvement of current system [6]. The frequent finding of the precise document(s) by the user of search engine is considered as a case of perfect performance. The effectiveness of the search engine retrieval of relevant documents depends on the server of the country or countries where the computer set is logged. The present study is concerned with evaluation of Yahoo, Google, Msn and Hakia search engines within the context of perfect performance of search engines.

## **2. Methodology**

To achieve the goal of the present study, four world most used search engines (Yahoo, Google, Msn and Hakia) were selected for comparison of web document retrieval performance. A total of ten queries were selected and run over the four – mentioned search engines. The queries consist of one or two terms and cover various topics as shown in table (1).

**Table (1): List of Queries**

Query Number	Query
Q1	Acne
Q2	Air bag
Q3	Construction
Q4	Fire fighter
Q5	Immigration
Q6	Information Technology
Q7	Insomnia
Q8	Nature
Q9	Forest
Q10	Wireless Network

Four additional phrases of the same meaning but with different syntax were used with each query and run one by one on Hakia semantic search engine. The selected phrases include: what is; information about; definition and description of. All queries and queries with phrases were run on search engines within minimum interval of time.

After each run of query or query with phrases, the first 5, 10, 15 and 20 retrieved document were evaluated using the human relevance judgment and the retrieved documents were classified as “relevant“ or “non – relevant“ or “duplicate same source“ or “duplicate different source“ or “in accessible“ as shown in table (2) and table (3) for the result of search of queries and queries with phrases respectively. Such tables are not displayed in Tumer et al., study [7]. For the purpose of calculation of relative precision. A total number of relevant and duplicate of different source (mirror image) retrieved documents were calculated [7]. And, hence, called relevant (combined). Retrieved documents.

**Table (2): Number of relevant, duplicate different source, non relevant, duplicate same source and in accessible retrieved (queries) documents using four search engines**

Cut – off	Relevant Doc.	Duplicate different Source doc.	non relevant doc.	Duplicate same source doc.	In accessible	Search Engine type
First 5 doc.	36	-----	11	3	-----	Yahoo
First 10 doc.	64	2	27	6	1	
First 15 doc.	94	2	38	15	1	
First 20 doc.	118	2	56	23	1	
First 5 doc.	35	-----	14	1	-----	Google
First 10 doc.	68	-----	28	4	-----	
First 15 doc.	96	-----	47	7	-----	
First 20 doc.	124	1	64	10	1	
First 5 doc.	31	2	13	1	3	Msn
First 10 doc.	58	2	31	6	3	
First 15 doc.	87	2	48	10	3	
First 20 doc.	113	2	67	11	5	
First 5 doc.	34	2	13	1	-----	Hakia
First 10 doc.	65	4	30	1	-----	
First 15 doc.	93	10	44	3	-----	
First 20 doc.	127	10	57	6	-----	

----- no data

**Table (3): Number of relevant, duplicate different source, non relevant, duplicate same source and in accessible retrieved (queries with phrases) documents using Hakia search engine**

Cut – off	relevant doc.	duplicate different source doc.	non relevant doc.	duplicate same source doc.	In accessible	Phrase
First 5 doc.	34	2	14	1	-----	What is
First 10 doc.	65	4	30	1	-----	
First 15 doc.	93	10	44	3	-----	
First 20 doc.	127	10	57	6	-----	
First 5 doc.	26	-----	22	2		Definition
First 10 doc.	51	-----	40	6	3	
First 15 doc.	74	2	60	10	4	
First 20 doc.	103	2	77	13	5	
First 5 doc.	28	-----	20	-----	2	Description of
First 10 doc.	51	-----	37	10	2	
First 15 doc.	83	4	49	12	2	
First 20 doc.	113	5	66	14	2	
First 5 doc.	38	1	9	1	1	Information about
First 10 doc.	71	3	22	3	1	
First 15 doc.	103	5	35	3	3	
First 20 doc.	130	8	54	4	4	

----- no data

Relative precision is defined as the ratio of the number of relevant documents (combined) retrieved to the number of total document retrieved [8]. Calculation of relative precision at different cut – off points (5, 10, 15, 20) is helpful in estimating the distribution of relevant document over their rank [7].

Relative precision were calculated for each pair of query – search engine table (4) and for each pair of query with phrase – Hakia search engine. table (5). Such tables are absent in Tumer et al., study [7].

**Table (4): Relative precision for each pair of query- search engines**

Cut – off	Precision retrieved	Search engine type
First 5 doc.	72	Yahoo
First 10 doc.	66	
First 15 doc.	64	
First 20 doc.	60	
First 5 doc.	70	Google
First 10 doc.	68	
First 15 doc.	64	
First 20 doc.	62	
First 5 doc.	66	Msn
First 10 doc.	60	
First 15 doc.	59	
First 20 doc.	58	
First 5 doc.	72	Hakia
First 10 doc.	69	
First 15 doc.	69	
First 20 doc.	69	

**Table (5): Relative precision for each pair of query+ phrase- Hakia search engines**

Cut – off	Precision retrieved	Phrase
First 5 doc.	72	What is
First 10 doc.	69	
First 15 doc.	69	
First 20 doc.	69	
First 5 doc.	52	Definition
First 10 doc.	51	
First 15 doc.	51	
First 20 doc.	53	
First 5 doc.	56	Description of
First 10 doc.	51	
First 15 doc.	58	
First 20 doc.	59	
First 5 doc.	76	Information about
First 10 doc.	74	
First 15 doc.	72	
First 20 doc.	69	

### 3. Discussion

Few comments can be made on examination of data displayed in table (2) and table (3). In both tables, it can be noticed that the number of relevant and non – relevant documents retrieved are rather correlated with cut – off point values (5, 10, 15, 20) relative to duplicate (same and different source) and inaccessible documents. A part from the phrase “description of“ run on Hakia search engine at cut – off point value of (5) table (3), data of same source duplicate documents retrieved were displayed in both tables. Comparison of data displayed in table (2) show that Hakia is the only search engine consistently display zero number of inaccessible document retrieved at all cut – off point values, where as table (3) show that zero inaccessible document retrieved is in the case of the phrase “what is“ run at Hakia search engine at all cut – off point values. Such details of data is not shown by Tumer et al., study [7] and, therefore, comparison can not be made.

Relative precision together with cut – off point value are showed in table (4) and (5) and displayed in Fig.(1) and Fig.(2) respectively.

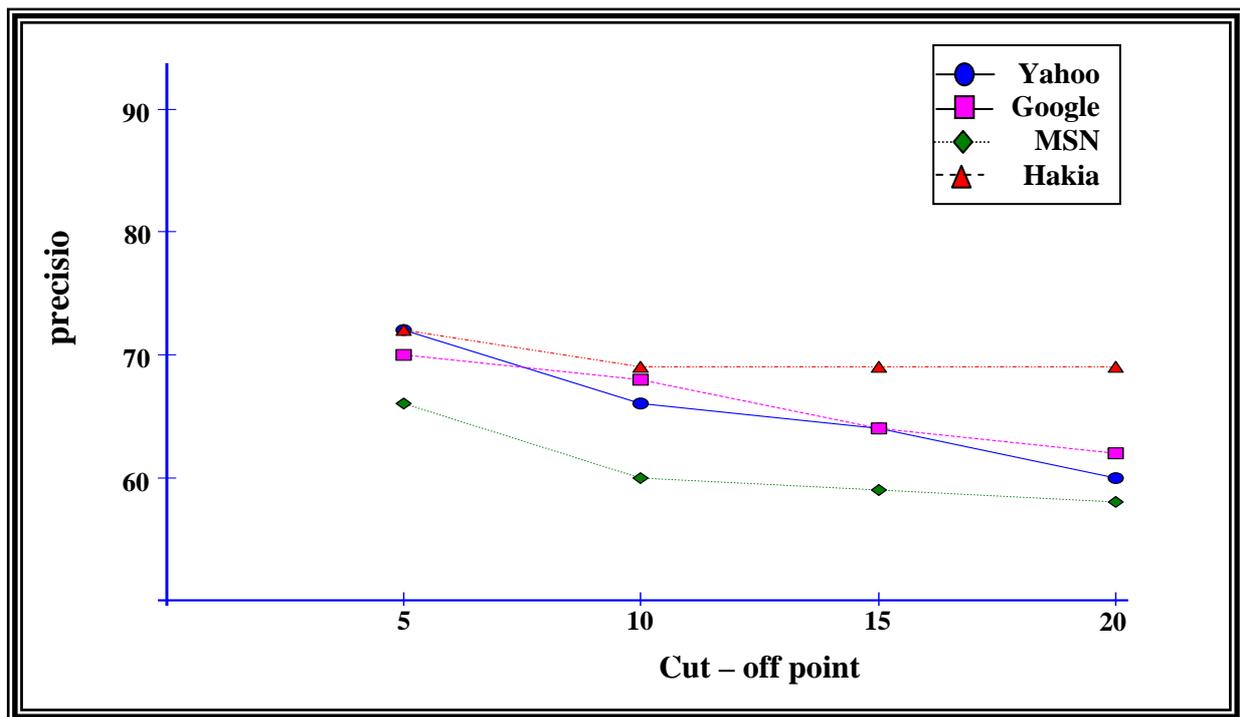
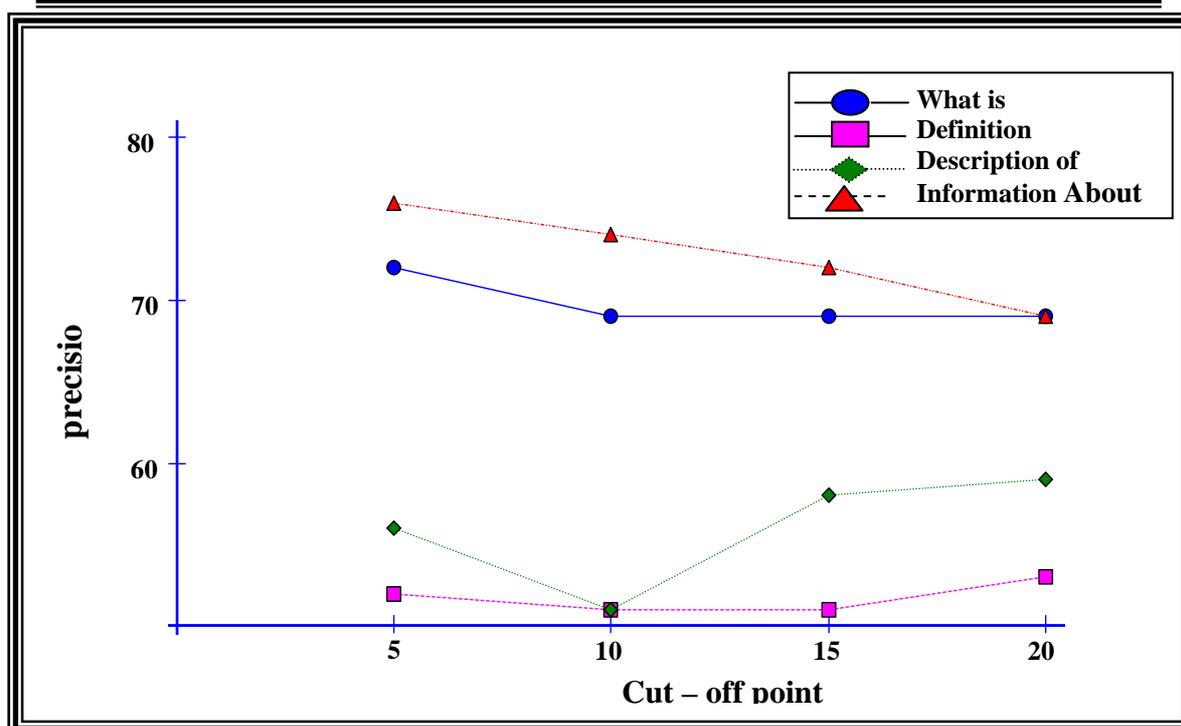


Figure 1: Relative precision for each pair of query- search engines



**Figure 2:Relative precision for each pair of query+ phrase- Hakia search engines**

It can be noticed in Fig.(1) a trend of decreasing in relative precision with an increase in cut–off point values. This represent similarity to Tumer et al.[5] findings. Other similar features is the minimum value of relative precision of Msn search engine at all cut – off point values. On the other hand, maximum relative precision of Hakia search engine Fig. (1) is considered as crucial differences from the finding of Tumer et al.[5] study, where as, the relative precision of Yahoo and Google search engines exchanging places at high and low side of the relation by crossing or matching each other within range of all cut – off point values. The only deviation is in matching of relative precision of Yahoo search engine with that of Hakia search engine at cut – off point value of (5).

In general, comparison show lower relative precision of semantic search engine performance of Hakia search engine Fig. (2) relative to that of keyword – based search engines Fig. (1). Details of Fig. (2) show that relative precision is at maximum for retrieved documents in case of phrase “information about“ with query was run on Hakia search engine where as phrase “what is“ with query comes next, yet the two combinations match in relative precision value at cut – off point value of (5), while lower value of relative precision is shown in case of phrases “description of“ with query and “definition“ with query were used. The last phrase is the least relative precision. In this respect, no such data is displayed (tables and Figures) in Tumer et al., [7] study and, hence, no comparison can be made.

#### 4. Conclusion

1. Number of retrieved relevant and non – relevant documents is correlated with cut – off point value for keyword – based and meaning – based search engines.
2. Wide variations in the number of remaining retrieved document with cut – off point values were noticed.
3. For keyword – based search engines, the Hakia and Msn search engines show maximum and minimum relative precision respectively, while the relative precision of Yahoo and Google search engines are exchanging places by crossing and matching within range of all cut – off point values.
4. For semantic performance of Hakia search engine, the maximum and minimum relative precision belong to the phrases “information about“ and “definition“, while phrase “what is“ comes second and “description of“ is third

#### 5. References

- 1) C. Guggilla, S. Mohamed and S. Vardarajan., "Extracting semantic sets from the World Wide Web using unsupervised learning", Journées internationales d'Analyse statistique des Données Textuelles, India 2009.
- 2) C. Muller, T. Zesch. M. Muller, D Bernhard, K. Ignatova, I. Gurevych and M. Muhlhauser. "Flexible UIMA Components for Information Retrieval Research", 6<sup>th</sup> International Conference on Language Resources and Evaluation, pp. 24-27, Marrakech, Morocco, May, 2008.
- 3) G. Kasneci, F. M. Suchanek, G. Ifrim, M. Ramanath and G. Weikum, “NAGA: Searching and Ranking Knowledge”, 24<sup>th</sup> IEEE International Conference on Data Engineering, pp. 953-962, Cancun, Mexico, April 2008.
- 4) D. Lee, J. Kwon, S. Yang and S. Lee, “Improvement of the Recall and the Precision for Semantic Web Services Search”, 6<sup>th</sup> IEEE International Conference on Computer and Information Science, pp. 763-768, Melbourne, Australia, July 2007.
- 5) D. Tumer, Y. Bitirim and M. A. Shah, “An Empirical Evaluation on Semantic Search Performance of Keyword-Base and Semantic Search Engines: Google, Yahoo, Msn and Hakia” 4<sup>th</sup> IEEE International Conference on Internet Monitoring and Protection, pp. 51-55, Venice, Italy, May 2009
- 6) G. Atsaros, D. Spinellis and P. Louridas, "Site – Specific versus General Purpose Web Search Engines: A Comparative Evaluation", 12<sup>th</sup> Pan-Hellenic Conference on informatics, pp. 44-48, IEEE Computer Society, Samos Island, Greece, August 2008.
- 7) Y. Bitirim, Y. Tonata and H. Sever, “Information Retrieval Effectiveness of Turkish Search Engine”, Advances in Information Systems, Lecture Notes in Computer Science, T. Yakhno (Ed.) vol. 2457, pp. 93-103, Springer-Verlag, Heidelberg, October 2002.
- 8) R. Baeza-Yates and B. Ribeiro-neto, “Modern information Retrieval”, ACM Press/Addison Wesley, 1999.