An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia

Duygu Tümer[1], Mohammad Ahmed Shah[2], Yıltan Bitirim[1]

[1] Department of Computer Engineering, Eastern Mediterranean University,
Famagusta / T.R.N.C.
027860@students.emu.edu.tr, yiltan.bitirim@emu.edu.tr
[2] Department of Computer Engineering, Middle East Technical University
Northern Cyprus Campus, Güzelyurt / T.R.N.C.
shah@metu.edu.tr

Abstract

This paper investigates the semantic search performance of search engines. Initially, three keyword-based search engines (Google, Yahoo and Msn) and a semantic search engine (Hakia) were selected. Then, ten queries, from various topics, and four phrases, having different syntax but similar meanings, were determined. After each query was run on each search engine; and each phrase containing a query was run on the semantic search engine, the first twenty documents on each retrieval output was classified as being "relevant" or "non-relevant". Afterwards, precision and normalized recall ratios were calculated at various cut-off points to evaluate keyword-based search engines and the semantic search engine. Overall, Yahoo showed the best performance in terms of precision ratio, whereas Google turned-out to be the best search engine in terms of normalized recall ratio. However, it was found that semantic search performance of search engines was low for both keyword-based search engines and the semantic search engine.

Keywords: Information retrieval, Semantic search performance, Semantic search engine, Keyword-based search engine, Evaluation.

1. Introduction

Web search engine is a computer program that allows users to search and retrieve web documents with queries for their information needs. The most popular search engines are Google [1], Yahoo [2] and Msn [3] with 71.9%, 17.7% and 4.2% volume of search ratios (based on US Internet usage), respectively [4]. Google, Yahoo and Msn are keyword-based search engines. 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].

Researchers have conducted several studies to evaluate retrieval effectiveness of keyword-based search engines and some have proposed search engines, methods and technologies for semantic search. Atsaros et al. compared the performance of keyword-based search engines (site-specific and general purpose) in terms of precision and relative recall [7]. Kasneci et al. proposed NAGA semantic search engine which uses knowledge base, introducing a graph-based query language and a novel scoring mechanism [8]. However, NAGA is not a publicly available search engine. Lee et al. proposed a semantic web services search system to provide better retrieval effectiveness using clustering and ontology approaches [9].

Although the research trend on semantic search engines is increasing, publicly available search engines on the web are, generally, keyword-based search engines. Keyword-based search engines, such as Google and Yahoo, have a large user-base [4]. Therefore, evaluation of semantic search performance of popular keyword-based and semantic search engines is a valuable work intended to motivate researchers and search engine providers to improve current systems further.

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.

978-0-7695-3612-5/09 $25.00 © 2009 IEEE
DOI 10.1109/ICIMP.2009.16

Table 1. Query List

Query Number	Query	Query Number	Query
Q1	network switch	**Q6**	air conditioner
Q2	computer engineer	**Q7**	nightmare
Q3	fashion	**Q8**	atom
Q4	evolution	**Q9**	reincarnation
Q5	baby sitter	**Q10**	myopic

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.

2. Methodology

Initially, three keyword-based search engines, namely, Google, Yahoo and Msn, and a semantic search engine Hakia were selected. Afterwards, ten queries that contain various topics and consist of one or two terms[1] 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, both, the selected keyword-based search engines 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 [11]).
- 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.

After each run of the query or PwQ, the first twenty[2] documents retrieved were evaluated using binary human relevance judgment and with this, every document was classified as "relevant" or "non-relevant". Total 1600 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 [12]; (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 "non-relevant" [12]; and (4) if, for some reason, a retrieved document became inaccessible, it was classified to be "non-relevant" [12].

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 [14]. As such, using precision at different cut-off points is helpful in estimating the distribution of relevant documents over their ranks [12].

The score-rank curve is strongly associated with the normalized recall, say R_{norm} [15]. R_{norm} is based on optimized expected search length [16].

[1] Jansen and Spink examined a transaction log of Dogpile in their study [10] and they have revealed that 70% of users used queries that consist of two terms or less.

[2] In the study of Spink and Jansen [13], a data collection from Dogpile was analyzed and one of the findings was that the ratios of the users that viewed only the first page and those that viewed only the first two pages of document search results were about 71.9% and 15.8%, respectively.

Table 2. Number of relevant documents retrieved

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

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, R_{norm} 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. [17], 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 $R^{+}_{\max}$ gives the maximum number of R^{+} [12].

3. Experimental 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 and Msn 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 fashion"). Yahoo retrieved more relevant documents and Msn 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 Msn. Google, Yahoo and Msn 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.

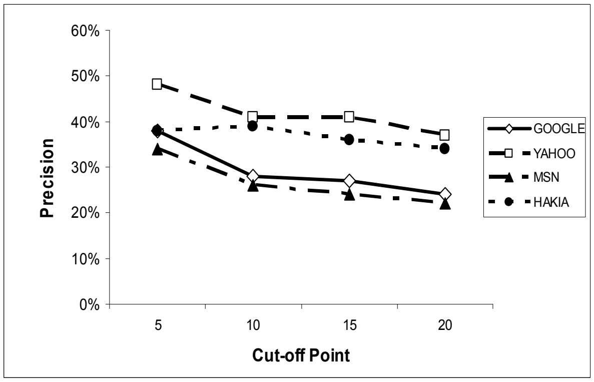

Figure 1. Mean precision ratios of search engines

Mean precision ratios of keyword-based search engines and the semantic search engine at various cut-off 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 Msn 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 cut-off points and its best precision ratio was 48% at cut-off point 5. However, Msn 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.

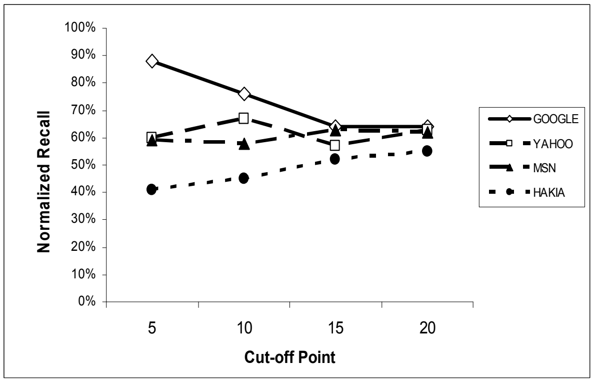

Figure 2. Mean normalized recall ratios of search engines

Figure 2 shows mean normalized recall ratios of keyword-based search engines and those of the semantic search engine at various cut-off points. Yahoo's normalized recall ratio was approximately the same as that of Msn 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 Msn were approximately the same at cut-off point 15 and the difference between Google, Yahoo and Msn 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.

4. Conclusion and Future Work

In this paper, an investigative evaluation on search performance of keyword-based and semantic search engines is detailed. It was found that Google, Yahoo and Msn 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 Msn at all cut-off points. However, Yahoo retrieved the highest number of relevant documents at all cut-off 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 Msn 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.

5. References

[1] Google, http://www.google.com (Accessed on 09 February 2009)

[2] Yahoo, http://www.yahoo.com (Accessed on 09 February 2009)

[3] Msn, http://www.msn.com (Accessed on 09 February 2009)

[4] Hitwise, "Hitwise US – Leading Search Engines", October 2008. http://www.hitwise.com/datacenter (Accessed on 12 November 2008)

[5] Hakia, http://www.hakia.com (Accessed on 09 February 2009)

[6] 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.

[7] G. Atsaros, D. Spinellis and P. Louridas, “Site-Specific versus General Purpose Web Search Engines: A Comparative Evaluation”, 12th Pan-Hellenic Conference on Informatics, pp. 44-48, IEEE Computer Society, Samos Island, Greece, August 2008.

[8] G. Kasneci, F. M. Suchanek, G. Ifrim, M. Ramanath and G. Weikum, “NAGA: Searching and Ranking Knowledge”, 24th IEEE International Conference on Data Engineering, pp. 953-962, Cancun, Mexico, April 2008.

[9] D. Lee, J. Kwon, S. Yang and S. Lee, “Improvement of the Recall and the Precision for Semantic Web Services Search”, 6th IEEE International Conference on Computer and Information Science, pp. 763-768, Melbourne, Australia, July 2007.

[10] 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.

[11] Wordtracker, “The Top 200 Long-Term Keyword Report”, The Wordtracker Report, 05 February 2008.

[12] 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.

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

[15] 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.

[16] 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.

[17] 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.