A Comparative Study on the Effectiveness of Semantic Search Engine over Keyword Search Engine using TSAP Measure

A. K. Mariappan
Information Technology Department,
Easwari Engineering College,
Chennai.

R. M.Suresh, PhD.
Computer Science Department,
Jerusalem College Engineering,
Chennai.

V.Subbiah Bharathi, PhD.
Computer Science department,
DMI College of Engineering,
Chennai.

ABSTRACT
The evaluation of search engine effectiveness has gained considerable momentum in the last few years. The *effectiveness*, measures the ability of the search engine to find the relevant information for the given user query. In recent years considerable research efforts have been devoted in developing semantic search engines aims to improve the traditional information search and retrieval process. We have seen number of semantic search engine projects and frameworks being implemented in various domains. In this paper, we have provided the results of retrieval effectiveness of Semantic Search engine against Keyword search engine using TREC Style Average Precision (TSAP) measure with little modification.

General Terms
Information Retrieval Technology, World Wide Web, Evaluation of search engine, Semantic search engine.

Keywords
Bing, Hakia, Information Retrieval, TSAP, Search Engine, Semantic Search

1. INTRODUCTION
Research in Information Retrieval (IR) technology has been evolved over years which helped the success of World Wide Web (WWW). IR research community has developed variety of techniques to find relevant information from large repositories. Baeza et al [1] described Vector Space model and probabilistic model for information retrieval. This model has been extended using Latent semantic Indexing [4], Machine Learning [5], and Probabilistic Latent Semantic Analysis [6] to improve the information retrieval process. Due to rapid increase of volume information on the web pose great difficulties in retrieving information efficiently. In order to present relevant information on top the result sets, some of the

commercial search engines used Page Ranking [2] and HITS [3] by exploiting link structure of the web. Despite huge success in ranking the relevance of information, those search engines face problems on two counts: (i) due to information overload and (ii) not understanding the user query semantically. This open problem has motivated a new thinking in search system which understands semantics (meaning) of the user query is called Semantic Search Engine.

Semantic search systems are next generation of search engines and its main goal is to provide better search results. The semantic search engine "understands" the meaning of the query input supplied by user and finds the relevant documents by "understanding" the contents. The general approach of semantic search is to match the query input against the internal fact database and search for corresponding facts and documents.

Evaluation is the key to understand and making progress in building better search engine for the today's World Wide Web. For a given query, we can define the effectiveness as a measure of how well the ranking produced by the search engine corresponds to a ranking based on user relevance judgments. A typical evaluation scenario involves the comparison of the result list for two or more systems for a given set of queries. The quality of the result list for each system is then summarized using an effectiveness measure that is based on relevance judgments. The relevance judgment should be done by the people who asked the questions, or by independent judge who have been instructed to how to determine relevance for the application being evaluated.

The remainder of the paper is organized as follows. The next section provides related work in information retrieval evaluation. In section 3 contains methodology adopted in our work. The 4th section presents the experimental results and in the last section we conclude the paper.

2. RELATED WORK
The Information Retrieval (IR) evaluation has its origins in the Cranfield II project [7]. It is the paradigm of the computer science oriented IR research, seeking to build better IR models and systems.

Leighton [9] evaluated four search engines namely Lycos, Infoseek, Webcrawler and WWW Worm based on precision measure using eight questions and rated Lycos and Infoseek higher.

Ding and Marchionini [10] investigated Infoseek, Lycos and Opentext for precision, duplication and degree of overlap using five queries.

Chu and Rosenthal [11] have used ten queries of varying complexity by evaluating the first ten results for relevance assessment and the findings revealed that AltaVista performed better than Excite and Lycos.

Clarke and Willet [12] evaluated AltaVista, Excite and Lycos and they found AltaVista outperformed other two search engines in terms of precision, recall and coverage.

Shafi and Rather [14] investigated five search engines for retrieving Scholarly information using biotechnology related search term and they used first ten results for estimation of precision and recall.

Voorhees and Harman [8] used the Text REtrieval Conference (TREC) evaluation model. In this model the test collection consisting of a document database, a set of fairly well defined

queries, and a set of relevance assessment identifying the documents that are topically relevant to each query. IR algorithms are evaluated for their ability to find relevant documents and the test results are expressed in term of precision and recall.

Hawking et al [13] investigated the effectiveness of twenty public search engines using TREC- inspired methods and a set of 54 queries taken from Web Search Logs. The World Wide Web is taken as the test collection and a combination of crawler and text retrieval system evaluated.

3. METHODOLOGY

The purpose of this work is to evaluate the effectiveness of the semantic search engine against keyword search engine. So we have selected the popular keyword based general purpose search engine Bing [15] with Hakia [16] – an upcoming Semantic search engine which uses computational linguistics, fuzzy logic, semantics and mathematics for its SemanticRank algorithm to enhance the results.

3.1 Search Queries

To evaluate the effectiveness of search engines we have selected fifteen search queries of Computer Science related topics. The queries are categorized in to three types based on their search complexity as a Simple query, compound query and complex query. These fifteen queries are listed below.

Simple queries:

I. Algorithm
II. Firmware
III. Debugging
IV. Softloading
V. Registers

Compound queries:

I. Software Testing
II. Device Drivers
III. Open Source
IV. Grid Computing
V. Virtual Memory

Complex queries:

I. Sliding Window Protocol
II. Software quality Management
III. Code Optimization Techniques
IV. Information Technology in Defense Services
V. Research Challenges in Semantic Web

3.2 Test set up

Each query (listed above) was submitted to each of the search engine (Bing and Hakia) which retrieves a quite large number of results but only the top ten results were evaluated to limit the study since most of the users will look up top results only. The listed queries are submitted on the same day on each the search engines in order to avoid any changes that may be caused due to system updating [12]. The relevancy of each result retrieved is checked manually and relevancy is classified into three levels: relevant, not relevant and less relevant.

3.3 Evaluation criteria

Since it is difficult to evaluate all the relevant results to a given query for each of the search engine, the traditional recall and precision is not suitable for evaluating in such situation. A popular measure for evaluating the effectiveness of search engines is the TREC- Style Average Precision (TSAP) [13].

In this work, TSAP at cutoff N, denoted as TSAP@N, will be used to evaluate the effectiveness of the search engines.

$$TSAP@N = \left(\sum_{i=0}^{N} r_i\right) / N$$

Where ri = 1/i if the ith ranked result is relevant, ri= 1/2i if the ith ranked result is less relevant and ri= 0 if the ith ranked result is not relevant/dead link/ duplicate result. The cutoff value N = 10. TSAP@N is an average precision and unlike the true TREC measure does not include a recall component. It is observed that TSAP@N tends to yield a larger value when more relevant results appear in the top N results and when the relevant results are ranked higher.

4. EXPERIMENTAL RESULTS

We evaluated the value of TSAP@10 for each of the search engine selected using the list of queries presented in the previous section. The results are reported in Tabular form for each query category.

Table-1: TSAP@10 of Bing for Simple Query

Search Query	Relevant	Less Relevant	Not relevant	Dead link/ Duplication	TSAP @10
Q.I	9	1	0	0	0.288
Q.II	3	6	0	1	0.217
Q.III	5	3	2	0	0.234
Q.IV	2	0	8	0	0.133
Q.V	8	1	1	0	0.251
Mean TSAP@10					0.225

Table-2: TSAP@10 of Bing for Compound Query

Search Query	Relevant	Less Relevant	Not relevant	Dead link/ Duplication	TSAP @10
Q.I	8	2	0	0	0.274
Q.II	9	1	0	0	0.268
Q.III	6	2	1	1	0.234
Q.IV	9	0	1	0	0.278
Q.V	8	0	1	1	0.258
Mean TSAP@10					0.262

Table-3: TSAP@10 of Bing for Complex Query

Search Query	Relevant	Less Relevant	Not relevant	Dead link/ Duplication	TSAP @10
Q.I	7	1	2	0	0.248
Q.II	6	2	1	1	0.242
Q.III	3	6	1	0	0.213
Q.IV	0	1	9	0	0.050
Q.V	4	1	3	2	0.117
Mean TSAP@10					0.174

Table-4: TSAP@10 of Hakia for Simple Query

Search Query	Relevant	Less Relevant	Not relevant	Dead link/ Duplication	TSAP @10
Q.I	8	0	2	0	0.258
Q.II	7	0	3	0	0.199
Q.III	6	3	1	0	0.296
Q.IV	4	0	6	0	0.177
Q.V	2	0	8	0	0.150
Mean TSAP@10					0.216

Table-5: TSAP@10 of Hakia for Compound Query

Search Query	Relevant	Less Relevant	Not relevant	Dead link/ Duplication	TSAP @10
Q.I	8	2	0	0	0.274
Q.II	8	2	0	0	0.277
Q.III	3	3	1	3	0.092
Q.IV	7	0	1	2	0.212
Q.V	9	0	1	0	0.282
Mean TSAP@10					0.227

Table-6: TSAP@10 of Hakia for Complex Query

Search Query	Relevant	Less Relevant	Not relevant	Dead link/ Duplication	TSAP @10
Q.I	5	2	3	0	0.211
Q.II	5	3	2	0	0.198
Q.III	2	4	4	0	0.168
Q.IV	0	5	4	1	0.114
Q.V	4	1	1	4	0.159
Mean TSAP@10					0.170

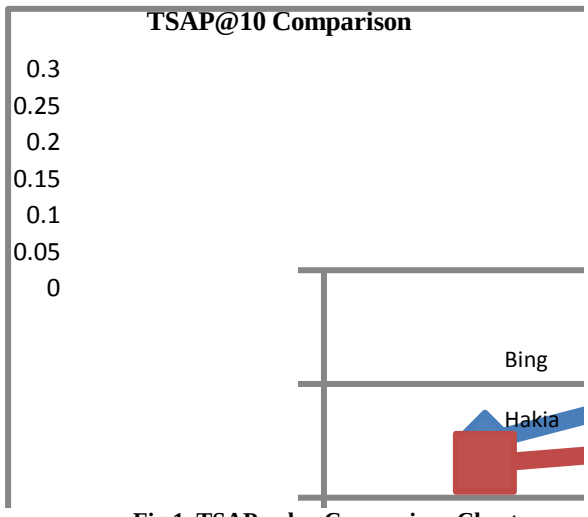

Fig.1. TSAP value Comparison Chart

5. CONCLUSION

The results of the experiment showed that Bing search engine is superior in compound query category whose mean TSAP@10 was 0.262. Both of the search engines are producing very closer results for the other two categories of queries. From the above results we observed that, some of the result pages are unreachable and duplication. In order to discourage dead links / duplications, we have awarded zero score for that page during evaluation. The study also revealed that, for the simple query category, both of the search engines producing many result pages from Wikipedia and free online dictionary. Though, the goal of the semantic search system is not achieved in this study may come up soon to produce better search result. In future, we plan to include other parameters such as coverage, update, user effort, and system response in addition to retrieval performance for evaluating the search engines.

6. REFERENCES

[1] R A Baeza-Yates and B A Ribeiro-Neto, 1999: Modern Information Retrieval, ACM Press.

[2] S Brin and L Page, 1998: The Anatomy of a Large Scale hypertextual web search engine. Computer Networks and ISDN Systems,Vol.30, No.1-7,pp107-117.

[3] J M Kleinberg,1998: Authoritative sources in a hyperlinked environment. SODA, pp.668-677.

[4] S. C. Deerwester, S.T. Dumais, T.K..Landauer, G.W. Furnas and R.A.Harshman, 1990: Indexing by latent semantic analysis. JASIS, Vol. 41,No. 6, pp.391-407.

[5] H Chen,1995:Machine learning for information retrieval: Neural Networks, Symbolic learning, and genetic algorithms. JASIS, Vol. 46, No. 3, pp.194-216.

[6] T Hofmann, 1999:Probabilistic latent semantic analysis UAI, pp.289-286.

[7] Cleverdon, C.W.,1967: The Cranfield tests on index language devices. ASLIB Proceedings, 19, pp.177-194.

[8] Voorhees, E. & Harman, D. Overview of TREC 2001. In The Tenth Text Retrieval Conference (TREC 2001).

[9] Leighton, H. (1996): Performance of four WWW index services, Lycos, Infoseek, Webcrawler and WWW Worm. http://www.winona.edu/library/webind.htm.

[10] Ding, W., & Marchionini, G.1996: A comparative study of The Web search service performance. In proceeding of the ASIS 1996 Annual Conference, October, 33, pp136-142, 1996.

[11] Chu, H., & Rosenthal, M. (1996): Search engines for the World Wide Web: A Compartive study and evaluation methodology. Proceedings of the ASIS 1996, Annual Conference, 33, pp.127-135.

[12] Clarke.S.,and Willett,P. 1997: Estimating the recall performance of search of search engines. ASLIB Proceedings, 49(7), pp. 184-189, 1997.

[13] D. Hawking, N. Craswell, P. Bailey, K. Griffiths., 2001:Measuring Search Engine Quality. Information Retrieval Journal, 4(1):33-59, 2001.

[14] Shafi, S. M., & Rather, R. A. (2005). Precision and recall of five search engines for retrieval of scholarly information in the field of biotechnology. Webology, 2(2). http://www.webology.ir/2005/v2n2/a12.html.