MEASURING THE EFFECTIVENESS OF PUBLIC SEARCH ENGINES ON THAI QUERIES

Shisanu Tongchim and Virach Sornlertlamvanich and Hitoshi Isahara
Thai Computational Linguistics Laboratory
112 Phahon Yothin Rd.
Pathumthani 12120, Thailand
email: {shisanu, virach}@tcllab.org, isahara@nict.go.jp

ABSTRACT
This paper compares the retrieval performance of seven public search engines on Thai queries by using a blind evaluation. Two of the engines used in this study are Thai-focused search services, while the remaining engines are large, commercial search engines that have wider collections of web data and support several languages. The results are compared by Mean average precision (MAP) and Mean reciprocal rank of the first correct answer (MRR). These results are calculated from binary relevance judgments of the first 20 returned results, using 56 topics. Statistical testing shows that there are significant differences among engines.

KEY WORDS
Web search engine, Blind evaluation, Thai queries

1 Introduction

A number of studies examining the performance of web search engines regularly appear in literature. Most of the studies in search engine evaluation have been done by using English queries. A study based on languages other than English (i.e. Chinese) was shown in [1]. To our knowledge, however, no systematic evaluation of public search engines on Thai queries has been reported. In this paper, we compare the performance of seven search engines by using Thai queries. The evaluation is completely blind. That is, the results from different engines are pooled and presented to the judges in random order. No information about the search engines is shown. Besides comparing the performance of search engines through a blind evaluation, we also discuss about search engine popularity based on referrer data from Thai web sites. The results based on an analysis of referrer data have been provided by the largest web statistics collector in Thailand, Truehits[1]. The results from referrer data would suggest the popularity and success of search engines in real usage. These results will be compared with the findings from the blind evaluation. This will suggest whether there are any correlations between search engine popularity and the results of our blind evaluation or not.

By using Thai test queries, a challenging issue encountered by every search engine is how to deal with a language without explicit word boundary. In a typical indexing method, finding reliable information about word boundaries before creating indexes is essential. However, an inherent error of word segmentation is typically inevitable. This is one of the major problems in Thai text processing. Therefore, good search engines would cope well with this problem. Despite efficient retrieval algorithms, powerful web crawlers with comprehensive collected web data are also key factors for successful search engines. A search engine may fail to find any relevant documents either because the retrieval algorithm fails to find them or the crawler has not fetched them. Thus, the evaluation conducted in this study compares not only the retrieval algorithms but also the crawlers used by different search engines.

It is interesting to note that only a small number of public search engines have been found to support Thai queries. For example, we have examined eight search engines that were used in the study by Gordon and Pathak [2]. To date, only two of them have been found to support Thai queries. Many search engines fail to retrieve Thai documents possibly due to the problem in one of two main search engine components, or both. For the crawler part, Thai documents may be ignored by crawlers due to their policies or their own interest. These crawlers may keep only supported languages to their collections. Another possible reason is due to the retrieval algorithms. That is, their algorithms may not support Thai. Unlike some languages (e.g. Chinese), few dedicated Thai search engines have been developed and opened to the public. To our knowledge, there are two dedicated Thai search engines, namely SiamGURU[2] and Sansarn[3]. Some search sites claim to be Thai search engines, but their results are adopted from other larger engines. In this study, SiamGURU and Sansarn are included in our evaluation.

The rest of this paper is organized as follows: Section 2 discusses related work. Section 3 discusses about search engine popularity in Thailand estimated from referrer data. Section 4 provides the description of our experiment. Section 5 presents the experimental results and discussion. Finally, Section 6 concludes our work.

[1] http://truehits.net/

[2] http://www.siamguru.com/

[3] http://www.sansarn.com/

2 Related Work

Several studies in web search engine evaluation have been proposed in the past decade. These studies differ from each other in several aspects, especially their methodologies and findings. Early studies in this area were conducted on few search engines using a small number of test queries. Ding and Marchionini [3] compared 3 search engines by using 5 topics. Chu and Rosenthal [4] also compared 3 engines on 10 topics. Nicholson [5] replicated the experiment by Ding and Marchionini [3] 10 times over the ten-week period. The results showed that the rankings of engines change from time to time.

In early 1997, Leighton and Srivastava [6] conducted a comparison among five commercial search engines. They submitted 15 queries to search engines, and measured the precision on the first 20 returned results. However, they divided the first 20 links into three groups (namely, the first three links, the next seven links and the last 10 links), and assigned different weights to these groups. The findings showed that three search engines were superior to the other two.

In 1998, Gordon and Pathak [2] compared eight search engines by using 33 topics from faculty members. All searches were performed by highly trained searchers. The assessment was done by the faculty members on the top 20 returned results from each search engine. The findings showed that absolute retrieval effectiveness was quite low. Moreover, there were statistical differences among search engines for precision, but not the retrieval effectiveness.

Later, Hawking *et al.* [7] compared 20 search engines by using 54 topics originated by anonymous searchers. The top 20 results were judged. The findings showed that there was a significant difference in the performance of the search engines. They also did a comparison among 11 search engines using two different types of query (namely, online service queries and topic relevance queries) [8]. They found a strong correlation between the performance results on both types of query.

3 Search Engine Popularity and Success

The results of search engine usage discussed in this section have been provided by Truehits.net [9]. Truehits is the largest web statistics collector in Thailand operated by Government Information Technology Services (GITS)[4]. The service has been opened to public. Every member places a small script provided by Truehits on their web pages. Each time these web pages are viewed, this script is executed and then some statistics are sent to Truehits. Truehits not only collects web access statistics in terms of the number of visitors to a particular page, it also collects other statistics, for example, web browser vendor, screen resolution, operating system, referrer information, etc. By tracking the referrer data, it is possible to identify what pages users were visiting or accessing immediately before coming to the current page. From this technique, Truehits can keep track of what search engines users were using to find the websites of Truehits members.

Statistics based on analyzing the referrer data provide two indications about search engines. Firstly, the results provide an indication of how the popularity of each search engine changes over time. Secondly, they also suggest about the success of each search engine in finding websites since the data originate from every visit by using search engines.

The results of search engine usage in Thailand based on referrer analysis over the past 3 years is illustrated in Figure 1. Before July 2004, the search engine market in Thailand was shared by two main players, Google and Yahoo. After that period, the search engine market is entirely dominated by Google. Recently, there is a new search provider that holds the second position, Sanook[5]. Sanook is one of the largest web portals in Thailand. The search service of Sanook was just introduced in February 2006. However, the search function of Sanook is provided by Google. From this reason, we could say that Google currently dominates the search engine market in Thailand. Note that Sanook is not included in our evaluation by the same reason.

4 Experiment

4.1 Blind Evaluation

The objective of our experiment is to evaluate search engines based on user preference of returned documents. A web-based user interface is developed for the evaluation. This interface accepts keywords from users and works like meta-search engines. It transmits the input keywords simultaneously to several individual search engines. The results from all search engines are then presented to users. Since the evaluation is completely blind, no information about the search engines queried is shown. Our interface parses the returned pages from different engines, and extracts the results. The results are merged into a single pool, and then presented in random order. Thus, users do not know which each result originates. A screenshot of our user interface is shown in Figure 2.

4.2 Search Engines

Seven public search engines are considered in this study: Google, Yahoo, MSN, AltaVista, AlltheWeb, SiamGURU and Sansarn. Note that SiamGURU and Sansarn are Thai-focused search sites, while other engines have wider collections of web data and support other languages as well. In the rest of this paper, the engines, except SiamGURU and Sansarn, are referred to as global search engines.

[4] http://www.gits.net.th

[5] www.sanook.com

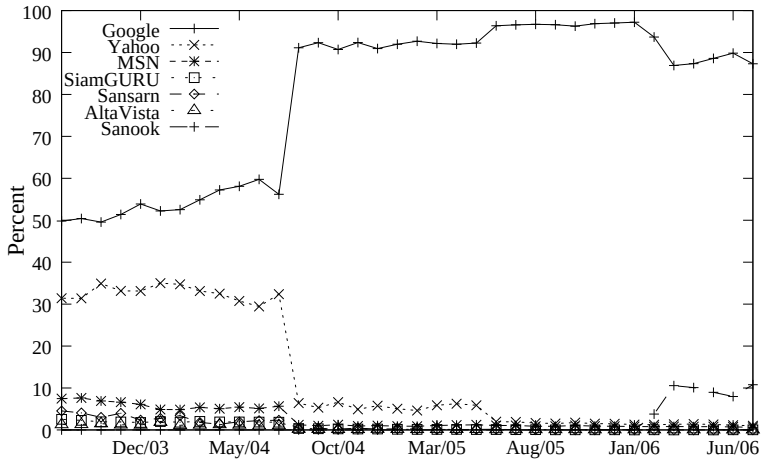

Figure 1. The results of search engine usage in Thailand based on referrer analysis over the past 3 years [Source: truehits.net]

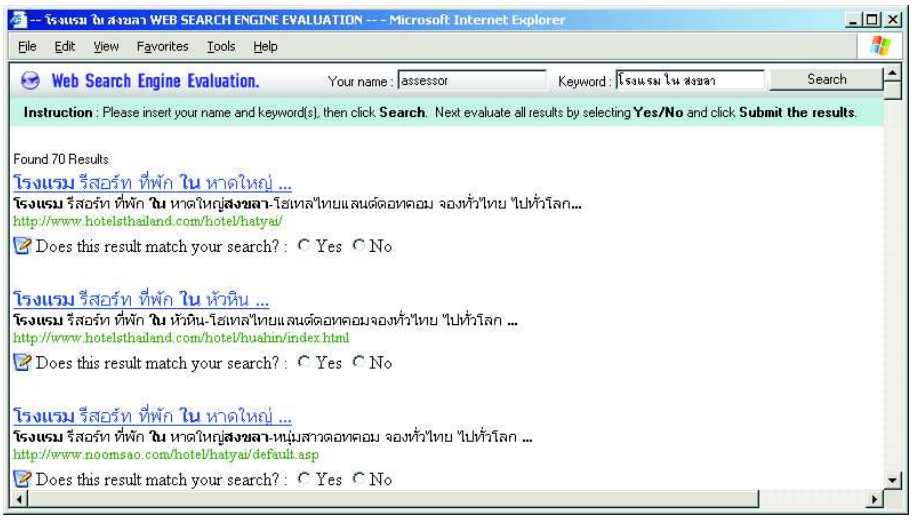

Figure 2. A screenshot of the web-based user interface used in the blind evaluation

We acknowledge that the number of engines used in this work is less than those used by some studies (e.g. 20 engines in [7]). There are two main reasons behind this decision:

1. The search engines used in this study must support Thai queries. While we conducted a survey of search engines, only a small number of search engines have been found to meet this requirement. As mentioned earlier, only two from the list of eight engines used in the study by Gordon and Pathak [2] are found to support Thai. Moreover, only five from the list of twenty engines in [7] can handle Thai queries. All of these five engines are included in our study. It is interesting to note that several meta-search engines cannot handle Thai queries correctly. Although these engines receive the results from Thai-supported search engines, several results are simply misleading to non-relevant articles.

2. Like the dynamic nature of WWW, search services change from time to time. Currently, the search engine market seems to be shared by just few companies [10]. Many search sites now use services provided by other companies rather than using their own engines. Many companies were acquired by other companies. Some companies (e.g. Northern Light) already closed their public search services. With the constant change in search service, the list of engines will differ from those used in previously published articles.

4.3 Test Queries

We use 56 Thai queries for evaluating the search performance. Some studies (e.g. [7]) use natural language queries in their studies. However, it is impossible for Thai language. To our knowledge, none of public search engines has been found to support natural language questions writ-

ten in Thai. Thus, keywords are selected for each query by hand. An example of a query and keyword selection is illustrated in Figure 3. The first line shows the query in English, while the second line shows the query in Thai. The third line is the query in Thai with word segmentation. For the sake of clarity, each word boundary is shown with a slash symbol ('/'). However, it does not exist in actual Thai text. The last line shows an example set of selected keywords from the Thai query. In our study, the length of queries ranges between 1 and 4 words.

4.4 Performance Measures

We assign 56 queries to a team of 7 judges (each is responsible for 8 queries). The relevance judgments are binary. That is, each result is judged whether its textual content is relevant to the keywords or not. In the experiment, the first 20 results from each search engine for each query are used. The inaccessible results are treated as irrelevant answers.

For each query, the results from seven search engines are pooled together, giving a maximum pool size of 140. The average actual pool size (when duplicates are possible) is 107.04 (76.46% of the maximum), while the average number of unique documents returned from all search engines is 70.32 (50.23% of the maximum). The average number of relevant documents for one query is 15.21, 10.87% of the maximum.

The average number of results returned by each engine for a query is $107.04/7 = 15.29$. This is somewhat different from other experiments based on English queries (e.g. the average number of results returned by each engine is 35.18 in the TREC-8 Web track experiment [11]). Our experiment based on Thai queries obtains fewer returned results. This reason influences the decision on performance measures used in our study.

Many evaluation measures used in Information Retrieval studies are based on *Precision* and *Recall*. Precision is the proportion of returned documents which are relevant, while Recall is the proportional of relevant documents that are retrieved. In typical, a calculation of recall is necessary to know exactly how many relevant documents there are. Since it is hard to answer such a question in the evaluation of public web search engines, some studies (e.g. [2]) used relative recall instead. Relative recall is calculated in relative to the number of returned documents that are judged to be relevant. However, some studies (e.g. [7]) objected to the use of this measure.

Precision at n documents (P@n) is one of common evaluation measures used in TREC web track and other literature. P@n means the proportion of returned documents which are relevant, calculated from the first n results returned from each engine. P@1 shows the precision of the first results that users see, while P@10 indicates the precision of typical results in the first page. Hawking *et al.* [7] presented the performance of search engines by plotting P@n against the document cutoff values ($1 \leq n \leq 20$). Although P@n is easy to understand, we decide to use other measures due to two reasons:

1. The numbers of returned results on several queries are less than the cutoff value (20). It is questionable about how to calculate the average P@n as a function of cutoff values ($1 \leq n \leq 20$). For example, system $\mathcal{A}$ finds only 2 results for one topic. P@1 and P@2 can be directly determined, but not the case of P@3-20. This situation happens to every search engine in our study. Thus, it is questionable whether the use of P@n as an evaluation measure is really justified.

2. The findings from some studies [12, 13] showed that P@n is less reliable than some measures.

We use Mean average precision (MAP) and Mean reciprocal rank of the first correct answer (MRR) as our evaluation measures. Both measures are standard TREC measures [14]. MAP is based on the performance over all relevant documents. This measure is the average of the precision value obtained when each relevant document is retrieved. Assuming that the precision of each relevant document is zero when it cannot be found. Thus, this measure rewards systems that show relevant documents at the early positions. While MAP is calculated from all relevant documents, MRR is calculated from the first relevant document found. Both measures are equivalent for topics with just one relevant document. Since it is hard to know all relevant documents in Web environment, MAP is calculated from known relevant documents in the judging pool instead, like the calculation of relative recall.

We also provide the results in terms of the success at n documents (S@n). This measure is the proportion of queries that at least one relevant document is found in the top n documents. However, we mainly use MAP for comparing the search engines since it is the most meaningful measure among other measures used in this study. Moreover, MAP is also more reliable than some measures (e.g. P@n when n is small) [12, 13].

5 Results and Discussion

The results are presented in Table 1, sorted according to MAP. From the table, the average precision in terms of MAP ranges from 0.212 (for Google) down to 0.022 (for Sansarn). Google is the top performer for all measures, while Sansarn achieves the lowest performance for all measures. SiamGURU is second only to Google in terms of MAP, but not for other measures. When comparing with the third rank in MAP (i.e. AlltheWeb), SiamGURU has better performance in terms of MAP, but not for other measures (i.e. MRR, S@1, S@5, S@10 and S@20). MAP evaluates the completeness of relevant documents found as well as the ranking system, while other measures consider only the rank of the first relevant document found. Therefore, the results imply that SiamGURU performs better in terms of the number of relevant documents found, but not for the ranking algorithm.

English query	Where can I find hotels in Songkhla province?
Thai query	ที่ไหนสามารถหาโรงแรมในจังหวัดสงขลา?
Thai query with manual word segmentation	ที่ / ไหน / สามารถ / หา / **โรงแรม** / **ใน** / จังหวัด / **สงขลา** / ?
Keywords	โรงแรม (hotel), ใน (in), สงขลา (Songkhla)

Figure 3. An example of a Thai query and keyword selection

The success at n documents provides some insights about the ranking algorithms used in the search engines. S@10 and S@20 of Google are equivalent. That is, Google always finds at least one relevant document in the first 10 documents if any relevant document can be retrieved. In addition, S@5 of Google is more than S@20 of other engines. From the table, the results show that the probability for finding a relevant document in the top 5 documents (S@5) by Google is 0.893, while the probabilities for finding a relevant document in the first 20 results (S@20) of other system range from 0.875 (for AltaVista) down to 0.482 (for Sansarn). This means that the probability in finding a relevant document in the first 5 documents of Google is higher than the probabilities in finding a relevant document in the first 20 documents of other systems.

The comparison by using differences in effectiveness measures alone cannot provide accurate information that one system is better than another. Statistical tests are often used to compare the systems. A recent study by Sanderson and Zobel [13] showed that the t-test is highly reliable for comparing IR systems. Thus, we use paired t-test, with $p \leq 0.05$, for comparing the search engines based on MAP. That is, confidence in the null hypothesis (no significant difference between two systems) is 5%. The results are shown in Table 2. The '$>$' symbol denotes the systems in the first column have better performance in terms of MAP than the systems in the first row, while the '$-$' symbol means there is no significant difference. From the table, there is no statistical difference between the top two performers, namely Google and SiamGURU. Statistically, Google outperforms all engines, except for SiamGURU. SiamGURU has statistically better performance than Yahoo, MSN and Sansarn, while its performance is comparable to AlltheWeb and AltaVista. Sansarn has statistically lower performance than all engines.

6 Conclusions

This article compares the performance of seven search engines by using Thai queries. Two of them (i.e. SiamGURU and Sansarn) are Thai-focused search sites, while the other engines (i.e. Google, Yahoo, MSN, AltaVista and AlltheWeb) are global search sites. The results show that there are statistically differences among the search engines. Overall, Google is the top performer for all measures. However, the difference in the mean average precision (MAP) between Google and the second performer (i.e. SiamGURU) is not statistically significant.

Despite the fact that Google currently dominates the search engine market in Thailand, the results of performance evaluation suggest that there are some competitive search providers, at least for Google and SiamGURU. However, the precision of returned results is just one factor. There are other factors (e.g. search time, reliability) that are also important to the success of search engines. The findings of this article would suggest that Google does better than other search engines for the precision of returned results and the popularity in real usage.

Acknowledgements

The authors would like to thank Mr. Norapat Karawawattana and Mr. Kergrit Robkop who implemented the blind evaluation system.

References

[1] F. Xin, Evaluating the quality of search engines: User-centered discussions on evaluation schema and comparative studies on Chinese and English search engines, Master's thesis, Peking University, 2003.

[2] M. Gordon and P. Pathak, Finding information on the world wide web: The retrieval effectiveness of search engines, *Information Processing and Management*, 35(2):141–180, 1999.

[3] W. Ding and G. Marchionini, A comparative study of web search service performance, *Proceedings of the 59th annual meeting of the American Society for Information Science*, 1996, 136–142

[4] H. Chu and M. Rosenthal, Search engines for the world wide web: a comparative study and evaluation methodology, *Proceedings of the 59th annual meeting of the American Society for Information Science*, 1996, 127–135

Table 1. The results for the seven search engines

	MAP	MRR	S1	S5	S10	S20
Google	0.212	0.713	0.607	0.893	0.946	0.946
SiamGURU	0.194	0.585	0.5	0.714	0.821	0.839
AlltheWeb	0.171	0.634	0.536	0.732	0.821	0.857
AltaVista	0.150	0.603	0.5	0.679	0.804	0.875
Yahoo	0.128	0.540	0.375	0.75	0.839	0.857
MSN	0.111	0.617	0.554	0.696	0.75	0.821
Sansarn	0.022	0.151	0.071	0.268	0.375	0.482

Table 2. Paired t-test

	Google	SiamGURU	AlltheWeb	AltaVista	Yahoo	MSN	Sansarn
Google		–	>	>	>	>	>
SiamGURU			–	–	>	>	>
AlltheWeb				>	>	>	>
AltaVista					>	>	>
Yahoo						–	>
MSN							>

[5] S. Nicholson, Raising reliability of web search tool research through replication and chaos theory, *Journal of the American Society for Information Science*, 51(8):724–729, 2000.

[6] H.V. Leighton and J. Srivastava, First 20 precision among world web search services (search engines), *Journal of the American Society for Information Science*, 50(10):870–881, 1999.

[7] D. Hawking, N. Craswell, P. Bailey, and K. Griffiths, Measuring search engine quality, *Information Retrieval*, 4(1):33–59, 2001.

[8] D. Hawking, N. Craswell, and K. Griffiths, Which search engine is best at finding online services?, In *Poster Proceedings of the 10th International World Wide Web Conference*, 2001

[9] Truehits, Truehits statistics, http://truehits.net/monthly/, 2006. (accessed August 1, 2006).

[10] D. Lewandowski, Web searching, search engines and information retrieval, *Information Services and Use*, 25(3-4):137–147, 2005.

[11] D. Hawking, E. Voorhees, N. Craswell, and P. Bailey, Overview of TREC-8 web track, *Proceedings of TREC-8*, Gaithersburg, Maryland USA, November 1999, 131–150

[12] C. Buckley and E.M. Voorhees, Evaluating evaluation measure stability, *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, 2000, 33–40

[13] M. Sanderson and J. Zobel, Information retrieval system evaluation: effort, sensitivity, and reliability, *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, 2005, 162–169

[14] N. Craswell and D. Hawking, Overview of the TREC-2004 Web Track, *Proceedings of TREC-2004*, Gaithersburg, Maryland USA, November 2004.