Evaluation of Web Search Engines with Thai Queries

Virach Sornlertlamvanich, Shisanu Tongchim and Hitoshi Isahara
Thai Computational Linguistics Laboratory
112 Paholyothin Road, Klong Luang, Pathumthani, Thailand
{virach,shisanu}@tcllab.org, isahara@nict.go.jp

Abstract

This paper discusses some challenging issues that are found in the evaluation of web search engines by using Thai queries. The discussions are based on our experience in evaluating and comparing the search performance of 7 search engines on Thai queries. The issues addressed in this paper will help in improving further evaluations of search engines for Thai.
Keywords: *Web Search Engine, Evaluation, Thai.*

1 Introduction

The research on retrieval effectiveness of algorithms or search engines for web documents has been conducted extensively. The research may be done on prepared test collections like the Web Tracks of TREC [6], while several studies have been done on public web search engines [4, 2, 3]. Most studies in this area have been conducted for English. However, the results for non-English may differ from the findings on English. In 2006, we initiated an evaluation of retrieval effectiveness for Thai [5]. We focused on the retrieval performance of public search engines. Seven public search engines were evaluated by using 56 Thai queries. Some findings in previously published research on search engine evaluation and Web Tracks of TREC can be applied to our evaluation on Thai. However, there are some issues that are not found in those experiments on English. In this paper, we will discuss our experience of evaluating search engines on Thai by comparing with the previously published results on English. This paper centers on issues uniquely found in the search engine evaluation by using Thai queries, rather than the performance of search engines.

We will first discuss the current status of search engine usage in Thailand. The results are based on the referrer data recorded by the largest web statistics collector in Thailand, Truehits[1]. The results will show the popularity of search engines and their success in finding web documents for given queries. Moreover, Truehits also records input keywords used by users. Then, we will address the language aspect of these input queries. The discussion will center on how the language usage in these queries may affect the evaluation of search engines on Thai. Then, we will focus on the performance measures. We argue that some standard measures may not be appropriate due to the nature of our evaluation. Lastly, we present the use of metasearch models to improve the current search results.

[1] truehits.net

2 Current Status of Web Retrieval in Thailand

When conducting an evaluation of search engine performance, it is important to know and understand user behavior in using search engines. In our case, the study of user behavior in using search engines for Thai web documents can be done by using the data from Truehits. Truehits is the largest web statistics collector in Thailand. The usage statistics collection has been done by providing a small script to members. Every member places this script on their web pages. Each time these web pages are viewed, some information has been collected and sent to Truehits. Various kinds of information have been collected, for example, web browser vendor, screen resolution, operating system, referrer information, etc. By analyzing the referrer data, Truehits can identify what pages users were visiting or accessing immediately before coming to the current page. Thus, Truehits can keep track of what search engines and keywords users were using to find the websites of Truehits members.

We are particularly interested in statistics of search engine usage, especially in keyword usage. Before going to the analysis of keyword usage in the next section, we would like to mention about the current search engine market in Thailand. Currently, Truehits has recorded that there are more than 3 million search engine usage per day (as of March 11, 2007). According to this number, about 480,000 keywords have been used in each day. These numbers reflect the amount of usage recorded by Truehits 's members. Thus, the real numbers should be larger than these values. Despite

the extensive usage of search engines in the Thai community, the search engine market in Thailand seems to be a low-competition market. Before August 2004, there are two main players in the search engine market, namely Google and Yahoo with about 50% and 30% market share respectively. After this period, Truehits 's statistics show that Google has entirely dominated the Thai search engine market for more than two years. Since August 2004, Google has maintained its high market share (86.92%-97.23%). Therefore, it seems that there is no continual competition among local and global search engines or no search engine can compete with Google. The lack of competition would result in slow development in search techniques and the coverage of web collections.

3 Query Log Analysis

Truehits also records keywords used by users to find websites. This list of keywords is obtained from the analysis of the referrer data. Thus, they will be queries conforming to the following criteria:

- The recorded keywords are examples of real usage. They are extracted directly when users follow the links provided by search engines to Truehits 's members.
- The results of search engines that are not members of Truehits will not be recorded, even though users follow the result links.
- By using these keywords, at least one result will be found for each keywords. Keywords without any results found cannot be recorded.
- They are recorded when users follow the links from the result pages of search engines. In case that some keywords are queried to search engines and users do not follow the links in search results, these keywords cannot be recorded.

These criteria are also applied to other statistics about search engines. Some results have to be found by search engines first and they seems to be relevant in order to attract users to follow the result links. Therefore, we expect that the number of keywords in real usage should be larger than this. Improper keywords without any returned results from search engines will not be recorded. Likewise, search engines that return uninteresting results may not be recorded since users may not follow the provided links to websites. Although only successful keywords have been recorded, the recorded keywords still provide some useful information and insights about how users use search engines.

Thai has no explicit word nor sentence boundary. This feature has an important influence on the keyword formation and the results obtained. When analyzing the log of keywords used by users, there are two main types of queries. The first one is like English queries. Each query is composed of one or several individual words separated with spaces. That is, users manually select important words as queries. The second query type is composed of one or few short phrases, or sometime (in a few cases) a short sentence. In general, they look like short phrases without manual word segmentation.

We randomly choose 1210 queries from the query log provided by Truehits. The number of words or phrases based on spaces is shown in Table 1. From the table, the majority of queries are a single word/phrase, while the second and third positions are the queries with two and three words/phrases respectively.

As mentioned earlier, users may form queries as sets of individual words and short phrases without manual word segmentation. In order to test this assumption, we apply a word segmentation algorithm to the 686 queries written as a single phrase. The word segmentation is done by using the maximum matching algorithm. The results of word count based on the output of the word segmentation algorithm are shown in Table 2. We acknowledge that an inherent error of word segmentation is typically inevitable. However, the results would provide an overview of queries written as a single word/phrase. From the table, the majority of queries written as a single phrase are composed of 2-3 words. Some queries are composed of several words, e.g. 76 queries out of 1210 queries are composed of five or more words.

Table 1. Query distribution based on the number of words/phrases

Number of Words/Phrases	
1	686
2	359
3	119
4	38
5	4
> 5	4

Table 2. Number of separated words for a single phrase

Number of Words	
1	65
2	234
3	227
4	84
5	45
> 5	31

Since most search engines try to find an exact match for each Thai query, no operation (e.g. word segmentation, query reformulation) has been applied to the original query string. Our assumption is that the longer the query is, the fewer the results are found. We test this assumption by submitting the 686 queries written as a single phrase to 6 public search engines. The numbers of returned results (in percentage) for each query length are shown in Table 3. When queries are short like a single word, most queries (70.15%) have the number of returned results more than 30. In the longer queries, fewer results are found. In many cases, no result can be found with queries composing of 5 words or more (about 27%). This is quite surprise, since all keywords can be recorded only when some results are found. This would reflect a lack of consistency in finding relevant results for Thai queries.

In this section, some statistics about the keyword usage have been measured and discussed. The results provide an overview of keyword usage and its impact. In the next section, we will look into some examples of keyword usage from the query log. Some challenging issues in searching for these keywords will be discussed.

4 Some Issues of Keyword Usage

Like the ordinary list of keywords submitted to search engines, the list of Thai keywords from Truehits has diverse characteristics. Many of them are short and ambiguous. Many queries contain typos or ill-written strings. Many queries are clean and specific, like names of organization or persons. In this section, we will discuss some possible problems from these real world queries.

4.1 Incorrect or inconsistency transliteration

There are many transliterated words in the list of queries. They are English words written by using Thai characters. One problem arises when each user has a different way for transliteration. For example, the word "Internet" has been transliterated into several forms: "อินเตอร์เน็ต", "อินเทอร์เน็ต" and "อินเตอร์เน็ท". There are also several queries that contain transliteration of the word "Internet", e.g. "อินเทอร์เนตไร้สาย" (Wireless Internet). Since most search engines try to find a word in its given form, some relevant documents may be missing. This problem also affects the performance of word segmentation algorithms.

4.2 Word boundary issue

Some queries in the query log are short, but specific. For example, one example is the word "ข่า" which is a plant name. We submit this word to 7 public search engines. Surprisingly, several search engines do not precisely recognize the word boundary of Thai texts. When we take a closer look at the first page results from these seven search engines, five of them return some results that do not contain the word "ข่า". These results contain the words like "**ข่าว**" (News) or "เครือ**ข่าย**" (Network). These words contain the string "ข่า" without any relation to the word "ข่า". This would suggest that the word boundary in Thai texts is still a challenging issue in the current search technology.

4.3 Indivisible unit issue

Some queries are indivisible units although each query can be considered as a set of words. For example, a query found in the query log is "กรมอุตุนิยมวิทยา" which is the "Thai Meteorological Department". This query can be considered as two words: "กรม" (Department) and "อุตุนิยมวิทยา" (Meteorology). Since this word represents a unique entity, it may be recognized as an indivisible unit. We have found that there are some queries that resemble to this word, but they are ill-written. For example, we found at least three queries that can be considered to refer to this word: "กรมอุตุ", "กรมอุตุนิยม" and "กรมอุตุวิทยา". The use of these keywords usually leads to the websites that have improper forms of the word "กรมอุตุนิยมวิทยา", rather than the website of the Thai Meteorological Department.

5 Performance Measures for Web Retrieval Evaluation

In this section, we will review some performance measures that are normally used in the studies of search engine and web retrieval evaluation. We will discuss our experience in applying these measures to our work in search engine evaluation on Thai queries.

Many evaluation measures used in the field of Information Retrieval 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. Typically, precision is plotted as function of recall. However, a calculation of recall is necessary to know exactly how many relevant documents there are. It is impractical or almost impossible to find the number of relevant documents 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. [3]) objected to the use of this measure.

Some performance measures that are used in this research area are as follows:

- *Precision at n documents (P@n)* : It is one of common evaluation measures used in TREC web

Table 3. Percentages of returned results for a single phrase

Number of returned results	Number of separated words					
	1	2	3	4	5	> 5
0	2.46%	4.79%	9.25%	14.80%	26.67%	26.62%
1-5	2.15%	14.87%	19.38%	23.39%	19.56%	22.08%
6-10	0.31%	8.29%	7.31%	11.46%	8.44%	6.49%
11-15	1.23%	4.27%	4.58%	3.82%	4.89%	3.90%
16-20	3.38%	6.67%	8.37%	10.50%	13.33%	9.09%
21-25	0.92%	6.75%	8.46%	7.88%	4.89%	7.79%
26-30	19.38%	14.96%	13.30%	8.83%	8.89%	10.39%
> 30	70.15%	39.40%	29.34%	19.33%	13.33%	13.64%

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. Several studies plot P@n against n.

- *Mean Average Precision (MAP)* : MAP is the average of the precision value obtained when each relevant document is retrieved. It rewards systems that rank relevant documents high.

- *Mean Reciprocal Rank of the first correct answer (MRR)* : Unlike MAP, MRR is calculated only from the first relevant document retrieved.

- *Relative recall* : Relative recall is calculated based on the number of relevant documents known to be in the set of returned documents.

The TREC web track also incorporates other measures (e.g. speed of indexing, size of indexes). However, those measures can be applied in evaluating retrieval algorithms, rather than public search engines. Among all measures, P@n is a popular measure used in this research area. To achieve reliable results, the comparison based on precision at earlier cutoff (1..5) should be avoided [1]. That is, P@n where $n = 20, 30$ or more will improve the reliability of the results. From our experience, the numbers of returned results on several queries are less than the cutoff value (i.e. 20 in our previous work [5]). This may not affect P@n at earlier cutoffs. After a certain cutoff, however, no more results can be found. It is questionable whether P@n after this cutoff is meaningful. In this situation, MAP may be a more promising measure. However, the MAP calculation is based on the number of relevant documents known like relative recall. This is estimated from the relevance judgement of the document pool from all search engines. Thus, the accuracy in estimating the number of relevant documents depends on the size of document pool. This affects the document cutoff value used in the experiment.

Table 4. The results of the metasearch approaches compared with the top two search engines on Thai queries

		MAP
Search Engines	Google	0.214
	SiamGURU	0.193
Borda Count	Borda-fuse	0.262
	Weighted Borda-fuse	0.293
	Evolutionary Borda-fuse	0.292
Condorcet	Condorcet-fuse	0.250
	Weighted Condorcet-fuse	0.254
	Evolutionary Condorcet-fuse	0.250

6 Search Improvement Based on Existing Results

Based on our search engine evaluation conducted in June 2006, the best engine of 7 search engines in the test covered only 20.18% of relevant documents found from all engines. This means that a large number of relevant documents may be missed by users. To improve the search results, we have explored the use of metasearch models. Metasearch takes the returned results from a number of search engines or algorithms, and then merges the results into one ranked list. We consider only models that use only ranked results as their input since the relevance scores are not provided in general.

Table 4 show the results of metasearch models compared with the top two search engines from the previous evaluation. The metasearch models are based on two voting systems: Borda Count and Condorcet. The weighted versions of algorithms assign different weights to search engines, while the evolutionary versions of algorithms use Evolutionary Programming (EP) to optimize the weight vector. Overall, all metasearch models outperform the top search engine like Google. The use of different weight assignment also improves the performance of metasearch models.

7 Conclusions

In this article, we have discussed some issues based on our experience in evaluating public search engines on Thai queries. The query log analysis shows a mix of query formulations. The information from the query log analysis helps in designing later experiments that mimic user behavior. Some challenging issues based on the examples of keywords are discussed. We also point out that the number of returned documents from search engines affects the choice of performance measures.

References

[1] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *SIGIR*, pages 33–40, 2000.

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

[4] H. 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.

[5] S. Tongchim, V. Sornlertlamvanich, and H. Isahara. Measuring the effectiveness of public search engines on thai queries. In *Proceedings of The Fifth IASTED International Conference on communications, internet, and information technology*, 2006.

[6] E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. The MIT Press, 2005.