PAPER *Special Issue on Knowledge, Information and Creativity Support System*

Improving Search Performance: a Lesson Learned from Evaluating Search Engines using Thai Queries

Shisanu TONGCHIM[†], ***Nonmember***, **Virach SORNLERTLAMVANICH**[†], ***and*** **Hitoshi ISAHARA**[††], ***Members***

SUMMARY This study initiates a systematic evaluation of web search engine performance using queries written in Thai. Statistical testing indicates that there are some significant differences in the performance of search engines. In addition to compare the search performance, an analysis of the returned results is carried out. The analysis of the returned results shows that the majority of returned results are unique to a particular search engine and each system provides quite different results. This encourages the use of metasearch techniques to combine the search results in order to improve the performance and reliability in finding relevant documents. We examine several metasearch models based on the Borda count and Condorcet voting schemes. We also propose the use of Evolutionary Programming (EP) to optimize weight vectors used by the voting algorithms. The results show that the use of metasearch approaches produces superior performance compared to any single search engine on Thai queries.

key words: *search engine evaluation, metasearch*

1. Introduction

Web search engines are essential tools for finding required information on the World Wide Web. The use of search engines for finding web documents is not limited to English. Many search engines support several languages, while some search services focus on languages other than English. Despite the number of languages supported by search engines, most studies on the performance of public search engines have been carried out for English. However, the performance of finding relevant documents may differ between English and other languages. The results based on English may not be generalized to other languages. Therefore, additional research based on other languages is necessary for providing information about the search engine performance on other languages.

In this study, we conduct some experiments for assessing the performance of search engines based on Thai queries. This study can be divided into three main parts. In the first part, we evaluate the performance of search engines using queries written in Thai. A challenge in developing information retrieval algorithms and other natural language processing techniques for Thai is that there are no explicit word boundaries. Therefore, efficient search engines should have the ability to deal with this ambiguity successfully.

The second part of this study examines the relation and degree of overlap of the returned results from search engines. The results reveal that the results from different engines appear to have a low degree of overlap. That is, the majority of web documents are retrieved by only one engine. The results also show that the correctness of search results tends to improve in results with a high degree of overlap.

The findings from the investigation of the degree of overlap encourage the use of metasearch approaches to combine search results. Consequently, the third part of this study examines some metasearch techniques based on the results of performance evaluation of search engines. The metasearch models are based on two different voting schemes: Borda count and Condorcet methods. Six different metasearch models are examined. The metasearch models based on the Borda count method, called Borda-fuse and Weighted Borda-fuse, were proposed by Aslam and Montague [1]. The metasearch models based on the Condorcet method, called Condorcet-fuse and Weighted Condorcet-fuse, were proposed by the same authors in their later study [2]. Their studies are based on English. We apply these methods to the problem of combining Thai results. We also propose the use of Evolutionary Programming (EP) to optimize weight vectors used by the Borda count and Condorcet algorithms. These algorithms are referred to as Evolutionary Borda-fuse and Evolutionary Condorcet-fuse.

The rest of this paper is organized as follows: Section 2 discusses related work which is roughly divided into three main topics, namely the evaluation of web search engines, the analysis of the degree of overlap and the metasearch approaches. Section 3 provides the description and the results of the blind evaluation of web search engines. Section 4 analyzes the relation and degree of overlap among the returned results. Section 5 evaluates the use of metasearch models to combine the results from different engines. Finally, Section 6 concludes our work.

2. Related Work

This paper is related to three research areas. This sec-

[†]The authors are with Thai Computational Linguistics Laboratory, NICT Asia Research Center, Pathumthani, 12120, Thailand.

[††]The author is with NICT, Kyoto, Japan.

tion provides a brief review of literature in each area.

2.1 Evaluation of web search engines

The comparisons among public search engines have regularly appeared in literature. The early studies conducted experiments by using the number of engines or the number of queries which sometimes can be considered to be significantly insufficient. 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.

As research in this area progresses, more systematic and well designed experiments have been carried out. Leighton and Srivastava [6] compared five commercial search engines by using 15 queries in early 1997. They measured the precision on the first 20 returned results. They found that three search engines were superior to the other two. Gordon and Pathak [7] evaluated eight search engines by using 33 topics from faculty members. The top 20 returned results from each search engine were judged by the faculty members. The findings showed that there were statistical differences among search engines for precision, but not the retrieval effectiveness. Later, Hawking *et al.* [8] applied an extended TREC-8 Large Web task methodology [9] to compare 20 search engines. The experiment was based on 54 topics originated by anonymous searchers. The top 20 results of each engine were judged. They found that there was a significant difference in the performance of the search engines. They also compared 11 search engines using two different types of query, i.e. online service queries and topic relevance queries [10]. They found a strong correlation between the performance results on both types of query.

2.2 Measuring the degree of overlap

The research on the returned results from web search engines is not limited to only the relevance judgement. Some studies aimed to examine the properties of the ranked results, or to compare the results among search engines.

In 2005, Dogpile [11] which is a metasearch provider conducted a study about the degree of overlap in the first page results from search engines. They used their findings to support their claims about the importance of using metasearch. Some of their findings are as follows:

- By submitting 12,570 queries to major four search engines, the majority of results (84.9%) were unique to one of these engines. Only 1.1% of returned results were found by all four search engines.
- Since the majority of the first page results are unique to only one engine, using only one web search engine may miss desired results. They used the number of unique search results missed by using only one search engine to support this claim. The results showed that 68%-72% of the first page results will be missed when using only one search engine.

Spoerri [12] developed a visualization technique for showing the degree of overlap in search engine results. The article itself does not measure the degree of overlap. The proposed technique not only provides an overview of overlap in search engine results, but it can be used to perform filtering operations visually, e.g. assigning different weights to search engines in order to create a new ranking function.

2.3 Metasearch approaches

Some studies proposed metasearch techniques to combine the final results from several search engines or information retrieval algorithms. Most studies in this area have been conducted on English queries. Kamps and Rijke [13] compared various algorithms on several European languages. They found that the results on English differ from those of other European languages.

Aslam and Montague [1] categorized metasearch techniques by the data they require. Some techniques require training data, while some do not use any training data. Some techniques require relevance scores, while some use only ranks.

Our work is carried out on results from public web search engines. The relevance scores of these results are not available. Therefore, we will consider only the metasearch techniques that utilize ranks, rather than relevance scores.

Aslam and Montague [1] proposed the use of a voting system, called the *Borda count*, as a fusion algorithm for metasearch. In the Borda count, voters rank choices or candidates in order of preference. Each candidate gets a number of points, depending on the position ranked by each voter. In a simple implementation, where there are n candidates, the top ranked candidate receives n points, the second ranked candidate gets $n-1$ points, and so on. Finally, the candidates are ranked according to the total points. In a single-winner election, the candidate with the most points wins. However, it is possible to use the Borda count in a multiple-winner election by selecting the candidates with the most points.

The fusion of ranked results from different search engines can be analogous to a multiple-winner election. Each search engine acts like a voter, while the returned results from each search engine are the ranked candidates. Thus, the Borda count can be applied to the problem of metasearch.

Aslam and Montague [1] developed two algorithms based on the Borda Count, namely *Borda-fuse* and *Weighted Borda-fuse*. Borda-fuse assigns the same weights to all engines, while Weighted Borda-fuse allows the use of different weights.

Later, Aslam and Montague [2] proposed a new algorithm based on another voting system, called the *Condorcet method*. Generally speaking, the Condorcet method finds the winners by comparing each candidate against every other candidate. In each pairwise comparison, the winner is the candidate that is ranked in the higher positions by the majority of voters. The winners are determined from the results of every possible pairing. They reported that the Condorcet models perform better than the Borda count models in combining the ranked results.

3. Performance Evaluation of Search Engines

3.1 Blind evaluation

The first part of this study is to evaluate search engines based on user preference of returned documents. Seven public search engines are included in this study: SiamGURU[†], Sansarn[††], Google, Yahoo, MSN, AltaVista and AlltheWeb. SiamGURU and Sansarn are Thai-focused search sites since their services center on Thai web documents. Unlike the first two engines, the rest of engines have wider collections of web data and support other languages as well. These engines are referred to as global search engines in the rest of this paper.

Note that the number of engines used in this work is less than those used in some studies (e.g. 20 engines in [8]). The first reason is that only a small number of search engines have been found to support Thai when we conducted a survey. Moreover, several metasearch engines cannot handle Thai queries correctly, although these engines receive the results from Thai-supported search engines. The second reason is that the current search engine market seems to be shared by just few companies [14]. Many search providers now utilize services from other companies, rather than using their own engines. Moreover, some companies were acquired by others, while some companies (e.g. Northern Light) already closed their public search services. Therefore, our experiment is unable to cover all engines used in previously published articles. We also would like to include other Thai search engines in our evaluation. From our survey, however, only SiamGURU and Sansarn have actively operated. One of the biggest web portals in Thailand like Sanook[†††] has just opened the search feature. However, the search function is based on the results provided by Google. Therefore, we decide not to include Sanook in our evaluation.

We developed a web-based user interface for the evaluation. This interface accepts keywords from judges. Then, it performs search operations by submitting the input keywords simultaneously to several search engines. The results from all search engines are merged into a single pool, and then presented to judges in random order. Therefore, judges do not know which each result originates.

We use 56 Thai queries in our evaluation. Each query is composed of selected keywords and a query description. We do not use natural language queries in this study since none of public search engines has been found to support natural language queries written in Thai. In our study, the length of queries ranges between 1 and 4 words.

All 56 queries are assigned to a team of 7 judges (each is responsible for 8 queries). The experiments were conducted in June 2006. The relevance judgments are binary. In particular, each result is judged whether its content is relevant to the assigned keywords and the query description or not. The inaccessible results are treated as irrelevant answers. The first 20 results from each engine which are typical results in the first two pages are used in this study.

3.2 Performance evaluation

In the literature on information retrieval, many evaluation measures are based on *Precision* and *Recall*. Precision is the ratio of relevant documents returned to the amount of all returned documents. Recall is the ratio of relevant documents retrieved to the total number of relevant documents in the collection. Typically, precision is plotted as function of recall.

In the evaluation of public web search engines, the number of relevant documents to a particular topic is usually unknown in practice. Thus, it is impractical to calculate recall. For this reason, other measures have been used to measure the performance of web search engines. Among these measures, Precision at n documents (P@n) is one of common evaluation measures used in the annual Text REtrieval Conference (TREC) web track and other literature. P@n means the proportion of relevant documents returned, calculated from the first n results returned from each engine. In our case, it is questionable to adopt this measure since the numbers of retrieved results on some queries is less than the document cutoff value (20). We therefore use Mean average precision (MAP) and Mean reciprocal rank of the first correct answer (MRR) which are standard TREC measures [15]. MAP is the average of the precision value obtained when each relevant document is retrieved. It rewards systems that rank relevant documents high. Unlike MAP, MRR is calculated only from the first relevant document retrieved. Both measures are equivalent when there is just one relevant

[†]http://www.siamguru.com/
[††]http://www.sansarn.com/
[†††]http://www.sanook.com/

document. Between two measures, MAP is the most meaningful measure. Thus, the comparison is mainly based on MAP.

3.3 Results of performance evaluation

The results are shown in Table 1. The results are sorted according to MAP. The average precision in terms of MAP differs dramatically between the highest ranked engine and the lowest ranked engine, ranging from 0.212 (for Google) down to 0.022 (for Sansarn). Google is the top performer for both measures, while Sansarn achieves the lowest performance. SiamGURU is second only to Google in terms of MAP, but not for MRR.

To ensure accurate comparison, statistical testing becomes necessary. Note that a recent study [16] showed that the t-test is highly reliable for comparing IR systems. In a situation in which several t-tests are carried out to compare all combinations of systems, however, the probability of making at least one Type I error increases as the number of systems in the comparison. That is, the probability of falsely rejecting the null hypothesis increases when several systems are compared by using multiple pairwise comparisons. The increased error rate is known as the familywise error rate. Thus, we use *repeated-measures ANOVA* for comparing the performance of search engines. Since MAP is more meaningful than MRR, the comparison will be based on MAP.

The use of repeated-measures ANOVA assumes that the assumption of sphericity is not violated. Mauchly's test is used to test whether the sphericity assumption is violated or not. The results of Mauchly's test indicate that this assumption is violated ($\chi^2(20) = 172.78, p < .001$). Thus, degrees of freedom are corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.48$). The results of repeated-measures ANOVA show that there are significant differences among the performance of search engines, $F(2.90, 159.41) = 14.66, p < .001$.

Difference between pairs of search engines is assessed by means of Bonferroni test ($p < .05$). The results reveal that Google statistically outperforms three search engines (MSN, Yahoo, Sansarn), but not for SiamGURU, AltaVista and AlltheWeb. When comparing SiamGURU with other search engines, there is a significant difference between SiamGURU and Sansarn, but not for the rest of engines. Moreover, Sansarn has statistically lower performance than all engines. Overall, there are some significant differences among these search engines.

The findings about the search engines performance on Thai queries have something conforming with statistics of real usage. Statistics about search engine usage in Thailand have been recorded by Truehits [17]. Truehits.net is the largest web statistics collector in Thailand operated by Government Information Technology Services (GITS)[†]. The service has been opened to public. Truehits not only collects web access statistics in terms of the number of visitors to a particular page, it also keeps 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.

Table 1 The results for the seven search engines

	MAP	MRR
Google	0.212	0.713
SiamGURU	0.194	0.585
AlltheWeb	0.171	0.634
AltaVista	0.150	0.603
Yahoo	0.128	0.540
MSN	0.111	0.617
Sansarn	0.022	0.151

Table 2 Coverage of relevant documents

	(%)
Google	20.18
Yahoo	14.42
MSN	11.30
SiamGURU	17.34
AlltheWeb	15.53
AltaVista	17.13
Sansarn	4.09

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 (about 90%). The results in this study confirm that Google performs well in both performance measures and this may be one reason why Google is popular in real usage.

4. Overlap and Relation Among Results from Different Engines

From relevant documents obtained from all search engines, the coverage of each search engine for relevant documents can be calculated as shown in Table 2. Even the best search engine like Google covers only 20.18% of relevant documents from the seven search engines. Despite the fact that most searching for Thai web documents relies on Google (about 90%), users would miss the majority of relevant documents. The results also suggest that it is not safe to rely on the results from a single search engine.

The results from all engines are compared to examine the degree of overlap and their relation to the correctness. The results are presented in Table 3. The first column, Degree of overlap (n), means the results

[†]http://www.gits.net.th

in the second and third columns are based on returned results shared by n search engines. The second column shows the percentage of returned results found in n engines. The third column is the percentage of relevant results in the returned results found in n engines.

From the table, 75.72% of returned documents are unique to one of 7 engines, while among these unique results, 19.44% are relevant. Only 0.05% of returned results are found in all engines, while 50% of these results are judged to be relevant. Overall, the majority of returned results are unique to one of 7 engines. As the degree of overlap increases, the documents have a higher chance of being relevant. In order to test the correlation between the degree of overlap and the percentage of relevant results, the Kendall's tau, τ, is calculated. The results indicate that there is a positive relationship between the degree of overlap and the percentage of relevant results, $\tau = .905, p < .01$. The findings from this analysis encourage the use of metasearch approaches to improve the performance by combining prospective results. In the next section, some experiments on metasearch approaches are carried out.

5. Metasearch Approaches

The previous analysis suggests the possibility of using some metasearch algorithms to improve the performance. In this work, we explore six algorithms based on the Borda count and Condorcet voting scheme. Aslam and Montague [1] proposed two algorithms based on the Borda count, i.e. *Borda-fuse* and *Weighted Borda-fuse*. Later, they proposed two algorithms based on the Condorcet method, i.e. *Condorcet-fuse* and *Weighted Condorcet-fuse* [2].

Borda-fuse and Condorcet-fuse assign the same weights to all engines, while Weighted Borda-fuse and Weighted Condorcet-fuse use the precision based on training data as the weights of engines. That is, Borda-fuse and Condorcet-fuse do not require training data. It can combine the ranked results directly. In contrast, Weighted Borda-fuse and Weighted Condorcet-fuse need training data to determine the precision value of each engine. The precision values calculated from training data are used as the weights for engines in the Borda count and Condorcet methods. Aslam and Montague [1] pointed out that the use of precision values as weights may not always be optimal. It would be ideal if some techniques are used to fine-tune the weight vector used by the Borda count and Condorcet. The results will reveal whether the use of precision values as weights is still promising or it is not optimal.

In this paper, we explore the use of Evolutionary Programming (EP), which is a class of Evolutionary Algorithms (EAs), to optimize the weight vector used by the Borda count and Condorcet. Classical EP (CEP) uses Gaussian mutation as the primary search operator. However, CEP may converge slowly on certain classes of problems. The Improved Fast Evolutionary Programming (IFEP) [18] was proposed by Yao *et al.* to overcome this problem. IFEP utilizes two types of search operators: Gaussian mutation and Cauchy mutation. Algorithm 1 outlines the procedure of IFEP. The use of two different mutation operators is to balance between exploration and exploitation. This work uses IFEP to fine-tune the weight vector of the Borda count and Condorcet. These algorithms are denoted as *Evolutionary Borda-fuse* and *Evolutionary Condorcet-fuse*.

Table 3 Degree of overlap (n), the percentage of results found in n engines and the percentage of relevant results in the returned results found in n engines

Degree of Overlap (n)	% of documents	% of relevant documents
1	75.72	19.44
2	7.97	25.16
3	10.23	27.30
4	4.39	31.79
5	1.32	40.38
6	0.33	53.85
7	0.05	50.00

Algorithm 1: Improved Fast Evolutionary Programming

initialize the population of μ individuals, $(x_i, \eta_i), \forall i \in \{1, \ldots, \mu\}$
evaluate the fitness of each individual, $(x_i, \eta_i), \forall i \in \{1, \ldots, \mu\}$

while *the halting criterion is not satisfied* **do**
  for *each individual* $(x_i, \eta_i), \forall i \in \{1, \ldots, \mu\}$ **do**
    create a single offspring $(x'_i, \eta'_i)_1$ from (x_i, η_i) by Gaussian mutation

    create a single offspring $(x'_i, \eta'_i)_2$ from (x_i, η_i) by Cauchy mutation

    evaluate the fitness of $(x'_i, \eta'_i)_1$ and $(x'_i, \eta'_i)_2$

    select the best offspring (x'_i, η'_i) out of $(x'_i, \eta'_i)_1$ and $(x'_i, \eta'_i)_2$

  end
  select the μ individuals out of (x_i, η_i) and $(x'_i, \eta'_i), \forall i \in \{1, \ldots, \mu\}$

end

Algorithm 2: Pairwise Comparison

set $C = 0$
for *each search engine* S_i **do**
  If S_i ranks d_1 above d_2, $C = C + w(S_i)$
  If S_i ranks d_2 above d_1, $C = C - w(S_i)$

end
If $C > 0$, rank d_1 better than d_2
Else rank d_2 better than d_1

Table 4 An example of voting profile

	2 engines	1 engine	2 engines
1st	a	b	d
2nd	d	a	a
3rd	b	c	b
4th	c	d	c

Table 5 Parameter setting of IFEP

Population size	20
Number of Generations	100
Tournament size	3
Number of Objective Variables (n)	7
Range of Objective Variables	$[0,1]^n$

5.1 Borda count

In the Borda count, voters rank choices in order of preference, rather than just electing the most favorite choice. When applying the Borda count to the metasearch problem, voters are search engines. Each engine returns the ranked results in order of relevance scores. Each result gets a number of points, depending on the position ranked by each search engine. Then, the results are ranked according to the total points.

For example, assuming that there are 5 engines. The returned results of these 5 engines are given in Table 4. Each engine returns a list of four results ranked in order of relevance scores. The top ranked result receives 4 points, the second ranked result gets 3 points, and so on. From the table, the Borda score of 'a' is calculated as: (4×2)+(3×1)+(3×2)=17. By using the same calculation, the Borda scores of 'a', 'b', 'c', 'd' are 17, 12, 6, 15 respectively. The ranked results based on the total scores are 'a', 'd', 'b' and 'c'.

5.2 Condorcet

Like the Borda count, each voter in a Condorcet election ranks the list of choices in order of preference. Finding Condorcet winners can be done by performing a series of pairwise comparisons. Each candidate is compared against other candidates. The winner of each pairing is the candidate that a majority of voters prefer more than the other one. For example, a pairwise comparison is conducted between 'a' and 'b' of the example voting profile. The candidate 'a' is the winner of this comparison since it is ranked ahead of 'b' by four of the five voters. By comparing every candidate against every other candidate, the ranked candidates can be found based on the results of pairwise comparisons.

Montague and Aslam [2] implemented the metasearch models by using the QuickSort algorithm. Firstly, a list of all returned results is created. This list is sorted by the QuickSort algorithm. The comparison function used by QuickSort is performed as the pairwise comparison mentioned earlier. Let S_i be the i_{th} search engine and $w(S_i)$ be the weight assigned to this engine. The pairwise comparison is outlined in the Algorithm 2. In the simplest form, the weight of every search engine is assigned to one.

5.3 Borda-fuse, Weighted Borda-fuse and Evolutionary Borda-fuse

We use the topics and the assessment of returned results in Section 3 to examine the use of metasearch approaches. For each topic, each search engine returns the first 20 ranked results or fewer candidates. In our implementation, the top ranked result receives 20 points, the second ranked candidate gets 19 points, and so on. If the number of returned results is less than 20, only ranked results are assigned scores. Once the total scores of results have been counted, each metasearch approach selects the top 20 results ranked by the total scores as the final answer.

In the Borda-fuse algorithm, all engines are assigned their weights to one. In reality, there are some differences in the performance of search engines. Thus, the use of different weights for calculating the total scores may improve the performance. Weighted Borda-fuse uses the precision values in terms of MAP as the weights of search engines. Note that the weights based on MAP range between 0 and 1. In order to calculate the weights in Weighted Borda-fuse, a training set is required. In our experiment, the 56 topics are randomly divided into two sets: (1) 40 topics as training data, (2) 16 topics as test data. In order to achieve statistically significant results, the experiments are repeated 100 times. That is, 100 sets of training data and test data are used in the evaluation.

Evolutionary Borda-fuse uses IFEP to find the weights for the seven engines. Therefore, the number of objective variables is 7. The parameters of IFEP are shown in Table 5. The fitness calculation is based on the 40 topics of training data. The best individual after 100 generations of each run is evaluated on the 16 topics of test data.

5.4 Condorcet-fuse, Weighted Condorcet-fuse and Evolutionary Condorcet-fuse

In addition to three variants of Borda-fuse, three metasearch models based on the Condorcet method are examined. In Condorcet-fuse, the weight of every engine is assigned to one. Weighted Condorcet-fuse uses MAP as the weights of search engines. The experiment is carried out in the same way as the Borda-fuse algorithms. That is, 40 topics are used as training data and 16 topics are used as test data. The experiments are repeated 100 times. We also use IFEP to optimize the

weight vector used by the Condorcet. This algorithm is referred to as Evolutionary Condorcet-fuse. The parameter setting is the same as Evolutionary Borda-fuse.

5.5 Experimental results

The results averaged over 100 runs are shown in Table 6. In addition to the results of six metasearch approaches, the results of the top two search engines are presented as a baseline. Clearly, the use of metasearch approaches significantly improves the performance. All metasearch approaches outperform the top two engines from Section 3. Moreover, the use of different weight assignments can further improve the performance of metasearch models. The performance of using IFEP to optimize the weight vector is not better than that of Weighted Borda-fuse and Weighted Condorcet-fuse. This suggests that the use of the MAP values as weights is quite optimal. No significant performance can be improved by IFEP.

Repeated-measures ANOVA is used to compare the performance. Again, Mauchly's test is used to test the sphericity assumption first. Mauchly's test indicates that the sphericity assumption is violated ($\chi^2(27) = 604.24, p < .001$). Therefore, degrees of freedom are corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.39$). Repeated-measures ANOVA indicates that there are significant differences among the performance of search engines, $F(2.74, 271.46) = 167.48, p < .001$.

Pairwise comparisons are carried out using Bonferroni adjustment, $p < .05$. The results indicate that all metasearch models statistically outperform the top two search engines (Google and SiamGURU). The best algorithm (Weighted Borda-fuse) significantly outperforms all algorithms, except for Evolutionary Borda-fuse. Evolutionary Borda-fuse also statistically performs better than other algorithms, except for Weighted Borda-fuse. Although the use of different weights slightly improves the performance in Condorcet-fuse, there are no significant difference among three variants of Condorcet methods.

Overall, all metasearch models can improve the performance provided by public search engines. The metasearch models based on the Borda count perform better than the Condorcet models. This somewhat contradicts the findings reported in earlier studies [2], at least in the case of Thai queries. The findings in the earlier studies show that the Condorcet models perform better than the Borda count. However, this result cannot be generalized to the case of Thai queries.

Table 6 The results of the metasearch approaches and the top two search engines

	MAP
Google	0.208
SiamGURU	0.191
Borda-fuse	0.256
Weighted Borda-fuse	0.287
Evolutionary Borda-fuse	0.285
Condorcet-fuse	0.247
Weighted Condorcet-fuse	0.251
Evolutionary Condorcet-fuse	0.247

6. Conclusions

This research conducts an evaluation of public web search engines by using Thai queries. The results show that there are statistically differences among seven search engines. We also compare the returned results to measure the degree of overlap. The comparisons among the returned results show that the majority of results are unique to just one of the search engines. Further, the results shared by several search engines are likely to be relevant. These findings encourage the use of metasearch to improve the performance.

This study explores three metasearch approaches based on the Borda count voting schemes and other three approaches based on the Condorcet method. We introduce the use of evolutionary programming to optimize the weight vector of the Borda count and Condorcet models. The experiments of these metasearch techniques are conducted on the results of relevance judgments from the earlier search engine evaluation. The results show that all metasearch approaches statistically outperform the top search engines. Moreover, no improvement in the performance can be achieved by using evolutionary programming. This suggests that the use of average precision as weights is quite robust.

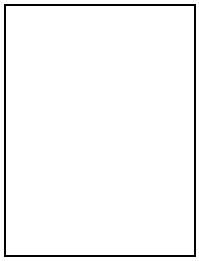
The main contributions of this work are as follows:

- A systematic evaluation of public search engine based on Thai queries is carried out.
- We show that the metasearch models can help in improving the performance of search engines on the returned results for Thai queries.
- We show that the use of average precision as the weight vectors of the metasearch models is quite optimal.
- We show that the metasearch models based on the Condorcet method may not always outperform the metasearch models based on the Borda Count, at least in the case of Thai queries.

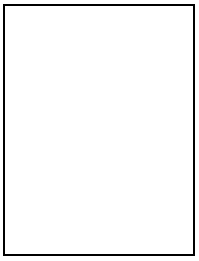
References

[1] J.A. Aslam and M.H. Montague, "Models for metasearch.," SIGIR, ed. W.B. Croft, D.J. Harper, D.H. Kraft, and J. Zobel, pp.275–284, ACM, 2001.
[2] M.H. Montague and J.A. Aslam, "Condorcet fusion for improved retrieval.," CIKM, pp.538–548, ACM, 2002.
[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, pp.136–142, 1996.
[4] H. Chu and M. Rosenthal, "Search engines for the world

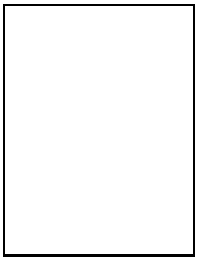
wide web: a comparative study and evaluation methodology," Proceedings of the 59th annual meeting of the American Society for Information Science, pp.127–135, 1996.
[5] S. Nicholson, "Raising reliability of web search tool research through replication and chaos theory," Journal of the American Society for Information Science, vol.51, no.8, pp.724–729, 2000.
[6] H. Leighton and J. Srivastava, "First 20 precision among world web search services (search engines).," Journal of the American Society for Information Science, vol.50, no.10, pp.870–881, 1999.
[7] M. Gordon and P. Pathak, "Finding information on the world wide web: The retrieval effectiveness of search engines.," Information Processing and Management, vol.35, no.2, pp.141–180, 1999.
[8] D. Hawking, N. Craswell, P. Bailey, and K. Griffiths, "Measuring search engine quality.," Information Retrieval, vol.4, no.1, pp.33–59, 2001.
[9] D. Hawking, E.M. Voorhees, N. Craswell, and P. Bailey, "Overview of the trec-8 web track.," TREC, 1999.
[10] D. Hawking, N. Craswell, and K. Griffiths, "Which search engine is best at finding online services?," Poster Proceedings of the Tenth International World Wide Web Conference, 2001.
[11] Dogpile, "Different engines, Different results." http://comparesearchengines.dogpile.com/OverlapAnalysis.pdf (accessed August 1, 2006), 2005.
[12] A. Spoerri, "Metacrystal: visualizing the degree of overlap between different search engines.," WWW (Alternate Track Papers & Posters), ed. S.I. Feldman, M. Uretsky, M. Najork, and C.E. Wills, pp.378–379, ACM, 2004.
[13] J. Kamps and M. de Rijke, "The effectiveness of combining information retrieval strategies for european languages.," SAC, ed. H. Haddad, A. Omicini, R.L. Wainwright, and L.M. Liebrock, pp.1073–1077, ACM, 2004.
[14] D. Lewandowski, "Web searching, search engines and information retrieval," Information Services and Use, vol.25, no.3-4, pp.137–147, 2005.
[15] N. Craswell and D. Hawking, "Overview of the TREC-2004 Web Track," Proceedings of TREC-2004, Gaithersburg, Maryland USA, November 2004.
[16] M. Sanderson and J. Zobel, "Information retrieval system evaluation: effort, sensitivity, and reliability.," SIGIR, ed. R.A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait, pp.162–169, ACM, 2005.
[17] Truehits, "Truehits statistics." http://truehits.net/monthly/ (accessed August 1, 2006), 2006.
[18] X. Yao, Y. Liu, and G. Lin, "Evolutionary programming made faster.," IEEE Trans. Evolutionary Computation, vol.3, no.2, pp.82–102, 1999.



Shisanu Tongchim received the B.Eng., M.Eng. and Ph.D. degrees in computer engineering from Chulalongkorn University, Bangkok, Thailand, in 1998, 1999 and 2004 respectively. In August 2004, he joined Thai Computational Linguistics Laboratory, where he is currently a research associate. His research interests include natural language processing, web mining, web information retrieval, etc.



Virach Sornlertlamvanich received his bachelor and master degrees in Engineering from Kyoto University in 1984 and 1986 respectively, and doctoral degree in Computer Science from Tokyo Institute of Technology in 1998. Currently, he is a co-director of Thai Computational Linguistics Laboratory, NICT Asia Research Center. He is a director of AAMT, and a member of AFNLP. His current research interests include natural language processing and information technology.



Hitoshi Isahara received the B.E., M.E., and Ph.D. degrees in electrical engineering from Kyoto University, Kyoto, Japan, in 1978, 1980, and 1995, respectively. His research interests include natural language processing and lexical semantics. He is a Leader of the Computational Linguistics Group and a Director of the Thai Computational Linguistics Laboratory (TCL) at the National Institute of Information and Communications Technology (NICT), Japan. He is a Professor at Kobe University Graduate School of Engineering, Japan. He is a Vice-President of the International Association for Machine Translation (IAMT), a President of the Asia-Pacific Association for Machine Translation (AAMT), a board member of GSK (Gengo Shigen Kyokai, linguisic resource association) and a board member of the Association for Natural Language Processing, Japan. He is a member of the Institute of Electronics, Information and Communication Engineers, Japan, the Association for Natural Language Processing, Japan, the Japanese Society for Artificial Intelligence, and the Information Processing Society of Japan.