Improvement on the effectiveness of Web Searching by using advance query operator among THAILAND's search engine sites

Thoedsak Sukdoa, Chonawat Srisa-an

Faculty of Information Technology, Rangsit University

ABSTRACT

Search Engine is one of the reasons for e-commerce success in today's world. They are playing an important role in supplying both international and local communities with various dimensions of business information for the customers and other interested parties.

Almost all E-Business transaction use simple classical keyword-based methods for searching commercial goods. One of the main problems that plague modern search engines is poor relevance. In many situations, search engines retrieve thousands of pages at least partially satisfying input query. Of course human attention remains more or less constant, and most humans are able to cope only with about 100-200 retrieved documents.

Advance query operators technique is one of proven method to improve the effectiveness of Web searching. However, only 10% of Web searchers utilize advance query operators, with the other 90% using simple query.

In this paper, we will examine the effects of query operators on the top popular THIALAND search engine in Thai keywords. We selected one best thousand queries from our web-logs. Each of the original queries contained query operator such as AND,OR,MUST APPEAR(+), or PHASE(""). We then modify the operators from these one thousand advance queries. This objective of this paper to determine the effect of advance query operators among THAILAND search engines.

Index Terms— web mining, text mining, information retrieval, search engines, Expert Systems and AI in e-Business

1. Introduction

One of the main problems that plague modern search engines is poor relevance. In many situations, search engines retrieve thousands of pages at least partially satisfying input query. Of course human attention remains more or less constant, and most humans are able to cope only with about 100-200 retrieved documents.

The correct use of query operators would increase the effectiveness of web searches by increasing the total number of retrieved document, increasing the number of relevant documents retrieved. Google advice that advance search statement can increase the accuracy of query search. However, 10% of web searchers use advance operators such as Boolean operators and phase searching while the others use an extremely simple queries. Most searchers are lazy to compose a complex query. In a survey, about 70% of the searches can locate relevant information from search engine[Spink et al. 1999]. Additionally, Web search engines continue to attract large numbers of web searchers. Many of the most popular web sites in terms of visitors are web search engines, implying that most view search engines are the best method available for finding information on the web.

2. The History of THAILAND E-business

In 1998, Mweb Company Limited (Thailand) was established as part of the MIH Group. The MIH Group has been involved with communication businesses covering more than fifty countries world wide. The MIH Group headquarter is located in The Netherlands. Mweb changed to the new name, 'Sanook Online Limited' on the 4th January 2007. Based on the number of users, and its popularity among advertising agencies and commodity vendors, NECTEC [3] reported in 2007 that Sanook holds the leading position of Thai's most popular online media. The data has indicated that the number of unique IP addresses connected to Sanook is more than 220,000 users a day. Moreover, one of the popular web information service sites, Alexa.com [4], points out that Sanook is ranked first among Thai portal websites since 2005 and its traffic is ranked 321 among 500 most popular sites in the world. Sanook also aims to provide its users with access to a large entertainment and information repository [5]. This will be achieved through its own content development and a variety of media.

In 2007, the number of Internet users accessing Thai sites increased to almost 70 million [3]. While there are new opportunities with the increasing number of users, there are potential issues related to deterioration in the system performance. In addition, the increased popularity of the sites also attract competitors. It is therefore necessary that Sanook needs to generate new business ideas to enable it to stays ahead of the competitors. In 2005, Torboon Phuangmaha, the new CEO of Sanook, announced new policies to promote the flagship service of Sanook -the Thai Web Directory. The sole objective is to maintain its top rank

position among all the Thai websites and to continually increase its revenue. He considers Sanook's competitors belong to two categories: a) internal competitors which are the local competitive companies, and b) external competitors which Sanook may join them as business alliances [7]. Based on its past performance and strategies, it could be estimated that the forecast sale of Sanook in fiscal year of 2007 to 2008 is likely to achieve the goal of the management.

The number of unique IP addresses connected to Sanook is more than 220,000 users a day. Moreover, one of the popular web information service sites, Alexa.com [4], points out that Sanook is ranked first among Thai portal websites since 2005 and its traffic is ranked 321 among 500 most popular sites in the world

3. Methodology

The effect of using queries with operators are investigated. This section describes the specified research questions and the methodology used to investigate them. The original queries we used represent real needs of real users that they submitted to a real web search engine. These users employed query operators in the manner they did in order to improve results. These queries are a sample of how users actually employ operators. Second, we utilized three of the most popular web search engines as measured by number of unique visitors.

3.1. Set Hypothesis from Selected Query Source

First Step is to select queries. We selected the specific queries submitted to our web logs. The operators used in this research are the AND, OR, MUST APPRAR and PHASE searching operators. We then eliminated all queries that did not contain one of one of these operators from the transaction log. We then generated four transaction logs, one of each of the queries operators used in this study.

As the results, Two hundred of the queries selected contained the AND operators; Two hundred contained the OR operators; Two hundred contained the MUST APPREAR operators; Two hundred contained the PHASE operators. Each query contained one or more uses of the same operator.

Hypothesis:

H1: The use of the AND query operator will result in a decrease in coverage.

H2: The use of the OR query operator will result in an increase in coverage.

H3: The use of the MUST APPREAR query operator will result in a decrease in coverage.

H4: The use of the PHASE query operator will result in a decrease in coverage.

H5: The use of the AND query operator will result in an increase in relative precision.

H6: The use of the OR query operator will result in an increase in relative precision.

H7: The use of the MUST APPREAR query operator will result in an increase in relative precision.

H8: The use of the PHASE query operator will result in an increase in relative precision.

H9: The use of the AND query operator will result in higher ranking document.

H10: The use of the OR query operator will result in higher ranking document.

H11: The use of the MUST APPREAR query operator will result in higher ranking document.

H12: The use of the PHASE query operator will result in higher ranking document.

No	Web Site	Agv visitors/day	Search Engine Feature
1	www.sanook.com	418,902	Its own engine
2	www.kapook.com	309,741	N/A
3	www.mthai.com	216,587	Its own engine
4	www.dek-d.com	207,416	Power By Google
5	www.teenee.com	170,662	N/A
6	www.exteen.com	166,970	Its own engine
7	www.manager.com	164,689	Its own engine
8	www.bloggang.com	144,159	Power By Google
9	www.playpark.com	127,320	N/A
10	www.narak.com	111,763	Its own engine
11	www.hunsa.com	90,551	Power By Google

Table 1: Popular web site in Thailand [1-19 MAY 2008]

3.2. Selection of Search Engine Services

Search engines are the major portals for users of the web, 71% of web users accessing search engine to locate other web sites [Sullivan 2000]. From statistic [Table 1], we choose three web sites as follows: sanook.com, manager.com and teenee.com. The reason to choose those web sites is because those sites had its own search services.

Statistic shows that about 80% of web searchers never view more than the first ten results in the result list [Silverstein et al. 1999]. Based on this statistic, we extract only the first ten results for comparison of relative precision and ranking in this study. If duplicates occurred within first ten results then we distinct only one results. For the analysis of coverage, we utilize the reported number of document by the respective search engines.

3.3. Reviewer process

We submitted each of the 100 original queries to one of three search engines. We then modified the query by removing the advance searching operator(s) and submitted it to the same search engine again. As example, queries with the MUST APPREAR operator(+E-Business+Conference+THAILAND), the PHASE operator("E-Business Conference THAILAND"), the AND operator (E-Business AND Conference AND THAILAND)

or the OR operator (E-Business OR Conference OR THAILAND) would be modified to the query. The process of submitting of submitting the original and modified query pair took approximately five minutes or less on each search engine. Therefore, the opportunity for the document collection to change between query submissions was minimal. Data collection occurred between January 1, 2007 and January 1, 2008.

After we submitted each query, we recorded the number of reported retrieved documents. Additionally, we saved the URLs for the top ten results. We did not evaluate identifiable Sponsored link or Sponsored Sites. Typically, search engines present these sponsored links within a separate area of the result page, usually on the right hand side of the browser window.

Three independent reviewers who had not performed the original searches evaluated each of the retrieved web sites. We first reviewed the summary presented by the search engine; they then retrieved the full text of web sites based on the reviewers made independent relevance judgments on each of the sites based on the reviewer's interpretation of the original query terms using topical relevance[Hawking 2000]. Reviewers evaluated the results of each search separately. We provided each of the reviewers a written explanation of the reviewing and the relevance judgment tasks. The reviewers rated each document using a four-point relevance scale. A score of 4 indicated a totally relevant document. A score of 3 indicated a partially relevant document. A score of 2 indicated a somewhat relevant document. A score of 1 indicated a non-relevant document. The average score must be at least 3.0 to be deemed as relevant for the study. The calculated agreement across the three raters using the individual reviewer's original rating for each document was quite reasonable.

TABLE II: Number and Percentage of Relevant Results by Rank Position by Operator

Rank Position	AND		OR		MUST APPEAR		PHASE	
1	99	11%	78	10%	94	12%	100	11%
2	105	12%	77	10%	93	12%	101	11%
3	100	11%	84	11%	83	11%	91	10%
4	94	11%	84	11%	81	10%	95	10%
5	93	11%	79	10%	63	8%	92	10%
6	85	10%	75	10%	78	10%	88	10%
7	81	9%	68	9%	72	9%	88	10%
8	77	9%	73	10%	76	10%	88	10%
9	82	9%	70	9%	73	9%	93	9%
10	69	8%	66	9%	64	8%	83	9%
Total	885	100%	754	100%	777	100%	912	100%

4. Conclusion

The results are discussed in the areas of coverage, relative precision, and ranking. The PHASE operator was the only operator that results in a decrease in coverage across all three search engine. The AND and MUST APPEAR operators also use to narrow a query, although they did not do so in this study. The OR operator had no significant effect on the coverage.

www.exteen.com has a search engine feature but advance query does not help much in term of relevant pages.

www.manager.com also has a search engine feature and use yahoo's search engine in separate areas. Its search engine retrieves some non-relevant page.

www.sanook.com returns the best coverage and relevant pages by using advance query feature.

Choice of search engine use of operators did not have an impact on relative precision. It appears that there is It appear that is little advantage to using OR in query, but there may be an advantage, at least in some case, in using the PHASE operator. A difference in ranking might be expected to make some difference to the user since it is more convenient to have relevant items at the top of the list. However, this study found only spotty improvements to ranking with no general improvement use any operator.

References

[1] D. F. Ferguson and R. Kerth, "WebSphere as an E-business Server," in 0018-8670/2001, vol. 2007: IBM, 2001.

[2] Steve Elliot and N. Bjorn-Andersen, "Part 1: Evaluating Commercial Web Sites: Development and Application of a Framework," in Electronic Commerce: B2C Strategies and Models, S. Elliot, Ed. Sussex, Eng.: John Wiley & Sons, Ltd, 2002, pp. 256-275.

[3] NECTEC (National Electronics and Computer Technology Center), "The Internet Index of Thailand," vol. 2007, 2007.

[4] "Site Stats for Sanook.com," 2007.

[5] "About Sanook," vol. 2007: Sanook Online Ltd., 2007.

[6] T. Phuangpakha (CEO of Sanook Online Ltd.), "Mweb proact to Online Marketing Business: Build Sanook.com to the Biggest Online Marketplace in Thailand," Meet the Press, 7 September 2005, 2005.

[7] T. Phuangpakha, "CEO of Sanook Online Ltd.," 8 May 2007.

[8] J. Colborn, Search Marketing Strategies: A Marketer's Guide to Objective-Driven Success from Search Engine. Amsterdam: Elsevier Butterworth Heinemann, 2006.

[9] "About Hunsa.com," vol. 2007, 2007.

[10] I. Thiraniti, "Product and Sale Director-Voice Content," Teleinfo Media Public Co. Ltd.," 9 May 2007.

[11] SULLIVAN, D. 2000 Search watch. Access On 1 June 2000.

[12] N. Slack, S. Chambers, and R. Johnston, Operations Management. London: Pearson Education, 2001..