The Effects of Search Engines and Query Operators on Top Ranked Results

Bernard J. Jansen
School of Information Sciences and Technology
The Pennsylvania State University
University Park, PA, 16801, USA
Email: jjansen@acm.org

Caroline M. Eastman
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29208.
E-mail: eastman@cse.sc.edu

Abstract

We examine whether the use of query operators changes the documents retrieved by three popular Web search engines. One hundred queries containing query operators were selected from the transaction log of a major Web search service. The query operators were then removed from these one hundred advanced queries. Both the original and modified queries were submitted to three major Web search engines. A total of 600 queries were submitted, and 5,748 of the documents retrieved were examined. Changes in the ranking of the top documents retrieved were examined. The significant results of our research are that the effectiveness of query operators is dependent on the specific search engine utilized, and that generally there is approximately 60% similarity between retrieved results across all search engines. Implications on the effectiveness of current searching techniques, for future search engine design, and of future research are discussed.

1. Introduction

Web searchers seldom use advanced query structure, such as Boolean operators or phrase searching, when using information retrieval (IR) systems [1]. Numerous Web studies note the near absence of query operators such as AND, OR, NOT, MUST APPEAR (+), and PHRASE (“ ”) in Web queries [2-4]. The use of Boolean operators is typically about 10% in these Web searching studies

It is generally assumed that the proper use of query operators would increase the effectiveness of Web searches. The objective of this study is to determine the effect of query operators by measuring the change in the top ranked results retrieved by three major Web search engines, America On-line Search (AOL), Google, and Microsoft Search (MSN). This knowledge is key to understanding how users search the web and for improvement of IR systems. In this paper, we present related literature, our research methodology, our research results and discussion. We also present directions for future research.

2. Related Studies

The effect of query formulation is an established field of information retrieval, although most research has been on traditional IR systems [5]. There have been relatively few studies comparing the retrieval results of different web search engines based query reformulation [6, 7]. Eastman (2002) explores the precision of search engines using a variety of topics and query formulations, noting that precision did not necessarily appear to improve with the use of the advanced query operators. Jansen (2000) examines the changes in results when different searching operators are utilized, reporting a 70% similarity in results between queries with no operators and the same queries with operators added.

Others studies have examined the difference between average and more sophisticated web queries [3, 4, 8]. Jansen et al. [3] reports a Boolean operator usage of about 8%. Silverstein et al. [8] report an advanced operator usage of approximately 20% for Alta Vista users. Spink et al. [4] show a Boolean usage rate of about 10% for Excite users. None of these studies examined the effect of query operators on changes in web search results.

We could locate no large study focusing on Web search engines that investigated the impact of query operators on retrieved results compared to just utilizing the query terms. In this study, we utilize actual queries submitted by real users. All of these queries contained operators chosen by real searchers. We focus on three web search engines that dominate the market in terms of number of visitors.

3. Research Design and Methodology

There is an expectation that properly structured queries using query operators have a greater probability of locating relevant information than the identical queries without operators. In order to locate more relevant information, there should be changes in the results retrieved by queries using operators relative to queries not using operators. Of course, one must control other factors

such as term usage, IR system, and document collection. We investigate the effect of using advanced queries (i.e., those using query syntax, such as Boolean operators) on the results retrieved by Web search services relative to the results retrieved by basic queries (i.e., those with no query operators).

3.1 Research Question

Our research question is to investigate the effect of query operators on the top ranked retrieved results relative to using no operators and the identical query terms.

3.2 Methodology

The methodology of our research is outlined below.

Selection of Queries

The specific queries used in this research were selected from a transaction log of a subset of queries submitted to the Excite search service on 1 May 2001. The transaction log contained over 1.2 million queries. Excite supported several advanced query operators. For this study, we investigated the AND, OR, MUST APPEAR, and PHRASE searching operators.

All queries that did not contain one of these operators were eliminated from the transaction log. We also eliminated all queries that were obviously seeking pornography, as determined by the researchers. We generated four transaction logs, one for each of the query operators used in this study. We then qualitatively reviewed each of the queries in the four transaction logs, removing those queries that were improperly constructed. Since we were not investigating the effectiveness of improperly formed queries, we did not want these queries to skew our results. The queries using more than one distinct operator were also removed. Of the remaining queries in each transaction log, we randomly selected 25 from each transaction log for use in this study.

Twenty-five of the queries selected contained the AND operator; twenty-five contained the OR operator; twenty-five contained the MUST APPEAR operator; and twenty-five contained the PHRASE operator. Each query contained one or more uses of the same operator. Query lengths ranged from two to eight terms. The Boolean and other operators were not counted as terms. For identification, we refer to these queries as the advanced queries. We refer to the queries without operators as the basic queries.

3.2 Selection of Documents

Studies show that approximately 80% of Web searchers never view more than first ten results in a results list [2, 8, 9]. Based on this evidence of typical Web searcher behavior, only the first ten results in the results list were utilized for comparison in this study. If duplicates occurred within the first ten, only one of the duplicates was utilized in the analysis. The strength of this analysis is that we are utilizing real search engines with real document collections in a manner consistent to that utilized by real searchers.

3.3 Searching Environment

Search engines are the major portals for users of the Web, with 71% of Web users accessing search engines to locate other Web sites [10]. There are approximately 3,200 search engines on the Web [11], with a handful dominating in terms of usage. These include AOL Search (AOL), Google, and MSN Search (MSN), which are the search engines used for this research. Our selection criterion was that these are three of the most popular Web search engines in terms of number of unique visitors per month [12].

3.4 Searching Rules

All the search engines supported all the query operators in some form, but there are frequently minor changes to the searching rules. At the time of the study, AOL directly supported the use of the AND, OR, MUST APPEAR, and PHRASE operators from its main page, although it also provided an advanced search option that facilitated the use of operator functionality. Google directly supported the AND, OR, MUST APPEAR, and PHRASE operators, although it states that the use of AND is not necessary. MSN directly supported the AND, OR, and MUST APPEAR operators. There was a drop down box for PHRASE searching. All search engines provided an advanced search mode, which directly supported all of the operators considered here as well as other features.

3.5 Data Collection Method

Each of the 100 original advanced queries was submitted to one of the search engines. We then modified the query by removing the advanced searching operator(s) and submitting the basic query to the same search engine. For example, the query with the MUST APPEAR operator *+furniture +moving +equipment* would be modified to *furniture moving equipment*. The entire process of submitting the advanced and basic 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. The process was repeated for all queries and all search engines.

After each query was submitted, the uniform resource locators (URLs) for the top ten results were saved for comparison. The top results for each advanced query were compared to the top results for the corresponding basic query. The match had to be exact when comparing the results. The documents listed had to be the identical

page at the same site. Different pages from the same site were not counted as matches. Identical pages at different sites were not counted as matches. If documents appeared in both results lists but in a different order, they were counted as matches as long as both were listed in the first ten results.

4. Results

Of the 600 queries submitted, 570 retrieved 10 or more results. There were 13 queries that retrieved no results. There were a total of 5,748 documents retrieved by all queries on all three search engines that were ranked in the top ten results. Table 1 shows the retrieval results by number of documents retrieved by all, advanced and basic queries.

Table 1: Number of Results Retrieved by Queries

Number of Results (Max. of 10)	All Queries	Advanced Queries	Basic Queries
10	570	287	283
9	0	0	0
8	0	0	0
7	1	0	1
6	2	1	1
5	1	1	0
4	0	0	0
3	2	1	1
2	7	2	5
1	4	1	3
0	13	7	6
Total Number of Queries	600	300	300

4.1 Results by Operator and Search Engine

An analysis was conducted for query operators by search engine, with the results displayed in Table 2.

Table 2: Comparison of Results by Search Engine and Operator.

	All	AOL	G	MSN	AOL	G	MSN	AOL	G	MSN	AOL	G	MSN
	All	**+**	**+**	**+**	**"**	**"**	**"**	**AND**	**AND**	**AND**	**OR**	**OR**	**OR**
Average	6.01	8.20	8.52	5.57	5.60	7.24	1.80	8.88	6.12	7.04	6.44	0.72	6.00
SD	4.05	3.14	3.48	4.61	3.87	3.63	2.16	2.15	3.33	4.02	2.95	1.77	3.76
Matching Results	**No.**	**No.**	**No.**	**No.**	**No.**	**No.**	**No.**	**No.**	**No.**	**No.**	**No.**	**No.**	**No.**
10	102	13	21	11	8	11	0	14	6	13	2	0	3
9	30	6	0	0	1	4	0	5	2	2	3	0	7
8	20	1	0	1	0	2	0	3	2	1	8	0	2
7	15	1	0	0	2	0	2	1	3	1	3	0	2
6	13	0	0	0	2	1	0	1	3	1	2	1	2
5	11	0	0	0	2	0	2	0	0	0	2	2	3
4	7	1	0	0	2	1	0	0	2	0	1	0	0

Table 2: Comparison of Results by Search Engine and Operator.

	All	AOL	G	MSN	AOL	G	MSN	AOL	G	MSN	AOL	G	MSN
	All	**+**	**+**	**+**	**"**	**"**	**"**	**AND**	**AND**	**AND**	**OR**	**OR**	**OR**
3	7	0	0	1	0	2	2	0	1	1	0	0	0
2	16	0	1	0	2	1	6	0	4	1	1	0	0
1	23	2	1	7	3	1	3	0	1	2	1	2	0
0	54	1	2	3	3	2	10	1	1	3	2	20	6
NR	2	0	0	2	0	0	0	0	0	0	0	0	0
Total	300	25	25	25	25	25	25	25	25	25	25	25	25

Notes: (1) NR – no results for both advanced and basic query pair; (2) Average – average number of identical results compared to using no query operators. (3) SD – standard deviation of identical results compared to using no query operators. (4) Total – total number of results lists. (4) No. – number of occurrences. (5) G – Google.

The top row lists the search engine, and the second row displays one of the advanced query operator for that search engine. The left most column is the number of possible matching results. Starting from a row in column 1, moving right across, the table displays the occurrences for each in the number (*No.)* columns, which is the number of times that the results from the advanced queries contained that number of exact matches with the basic queries. For example, beginning at the first column in the sixth row, there were 102 advanced queries that returned ten results identical to the corresponding basic queries for all search engines and all operators. Moving further to the right, each column shows for each search engine operator the number of times a query retrieved 10 identical results. The average number of matching results and the standard deviation is also given in rows three and four.

5. Discussion of Results

Table 2 shows that there were 102 (34%) advanced queries that retrieved results identical to those retrieved using the basic queries. This occurrence is by far the most frequent; the next highest occurrences were 54 (18%) advanced queries that retrieve no matching results and 30 (10%) that retrieved nine identical results. On average, 6.01 of the results are the same whether or not one uses query operators. In other words, the use of query operators would result in a change in 4 of top 10 retrieved results or inversely, NOT using query operators would result in a change in 4 of the top 10 retrieved results.

More importantly, the results in Table 2 show that to effectively employ query operators the searcher must have an understanding of the underlying IR system. For example, the operators that generally increase precision (AND, MUST APPEAR, and PHRASE) have little effect on the results from Google (7.24 to 8.52 results were identical on average), a moderate effect on AOL (5.60 to 8.88 results were identical on average) but a significant effect on MSN. The difference in results for the PHRASE operator on MSN was dramatic; only 1.80 results were identical on average. The number of identical results using MUST APPEAR on MSN (5.57) was lower than that for AOL (8.20) or Google (8.52); the average number of identical results using AND on MSN was intermediate (7.04) between AOL (8.88) and Google (6.12). Conversely, the OR operator that one would expect to increase recall had moderate effects on AOL (6.44 identical results on average) and MSN (6.00 identical results on average), but a significant effect on Google (only 0.72 results were identical on average). This certainly points out that the application of searching techniques using query operators can not be applied wholesale but must be utilized in conjunction with an understanding of the underlying IR system.

6. Conclusions and Future Research

This research studied whether the use of query operators changed the documents retrieved by three popular Web search engines. One hundred queries containing query operators were selected from the transaction log of a major Web search service. We removed the query operators from these one hundred advanced queries and submitted both the queries with operators and queries without operators to AOL, Google, and MSN. A total of 600 queries were submitted, and 5,748 documents retrieved were utilized in this study. Changes in the ranking of the top ten documents retrieved were examined by comparing the results from the queries with operators to those without operators.

Approximately 34% of the results lists were identical, with all 10 results being the same. On average,

approximately 64% of the time, 6 or more of the results will be identical. On deeper evaluation, it is apparent that effective employment of searching techniques utilizing query operators must be accompanied by an understanding of the algorithmic workings of the underlying IR system. Our results show that some operators may have a significant effect on the results returned by one search engine but only a slight effect on another.

It would be interesting to see the effect that these operators have on the changes in precision within the top ten documents (P@10, which is a measure of the number of relevant documents within the top ten ranked documents retrieved by a search engine). This is certainly an area for future research, which we plan to address. However, based on previous studies, one can project what the results of such a study might be.

Several researchers have examined the effect of query changes on the precision of web search engines [6, 13-15]. Although each of these studies utilized various numbers of queries, types of queries and number of search engines, some general trends emerge, namely that precision is approximately 0.5 (50%) to 0.7 (70%) on average for web search engines. Using these reported precision rates, we can estimate what the effect of the change in results are on the P@10 measures for our research.

Using an estimated precision rate of 0.5 and four new results on average using query operators, one can estimate that of the four new results two would be relevant and two would not be relevant. However, the precision rates from earlier studies are from a variety of queries, both with and without operators. So, one should also assume that two of the four results that were retrieved by the queries without operators would also be relevant and two would not be relevant. Therefore, one can estimate that the use of query operators will result in no increase in precision on average. This estimate is consistent with prior research, such as Eastman (2002), who reports that precision actually decreased in many cases with the use of query operators and that, on the average, there appeared to no significant effect on precision.

7. References

[1] C. Borgman, "Why are Online Catalogs Still Hard to Use?," *Journal of the American Society for Information Science*, vol. 47, pp. 493-503, 1996.

[2] C. Hölscher and G. Strube, "Web Search Behavior of Internet Experts and Newbies," *International Journal of Computer and Telecommunications Networking*, vol. 33, pp. 337-346, 2000.

[3] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic, "Real Life Information Retrieval: A Study of User Queries on the Web," *SIGIR Forum*, vol. 32, pp. 5-17, 1998.

[4] A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic, "From E-sex to E-commerce: Web Search Changes," *IEEE Computer*, vol. 35, pp. 107-111, 2002.

[5] A. Chowdhury, S. Beitzel, and E. Jensen, "Analysis of Combining Multiple Query Representations with Varying Lengths in a Single Engine," presented at The IEEE 3rd International Conference on Information Technology Coding and Computing, Las Vegas, Nevada, 2002.

[6] C. M. Eastman, "30,000 Hits May be Better than 300: Precision Anomalies in Internet Searches," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 879-882, 2002.

[7] B. J. Jansen, "An Investigation into the Use of Simple Queries on Web IR Systems," *Information Research: An Electronic Journal*, vol. 6, pp. 1-10, 2000.

[8] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz, "Analysis of a Very Large Web Search Engine Query Log," *SIGIR Forum*, vol. 33, pp. 6-12, 1999.

[9] B. J. Jansen, A. Spink, and T. Saracevic, "Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web," *Information Processing and Management*, vol. 36, pp. 207-227, 2000.

[10] CommerceNet/NielsenMedia, "Search Engines Most Popular Method of Surfing the Web," vol. 2000: Commerce Net/Nielsen Media, 1997.

[11] D. Sullivan, "Search Watch," vol. 2000: Search Engine Watch, 2000.

[12] Nielsen/Netrating, "Top Web Properties," vol. 2002: Nielsen/Netrating, 2002.

[13] E. Selberg and O. Etzioni, "On the Instability of Web Search Services," in *Presented at RIAO 2000: Computer-assisted information retrieval*, vol. 2002. Paris, France, 2000.

[14] S. Nicholson, "Raising Reliability of Web Search Tool Research Through Replication and Chaos Theory," *Journal of the American Society for Information Science*, vol. 51, pp. 724-729, 2000.

[15] W. Ding and G. Marchionini, "A Comparative Study of Web Search Service Performance," presented at The 59th Annual Meeting of the American Society for Information Science, Medford, NJ, 1996.