Jansen, B.J. 2003. Operators Not Needed? The Impact of Query Structure on Web Searching Results?. Information Resource Management Association International Conference, p. 814 – 817. Philadelphia, PA, 18 - 21 May 2003.

OPERATORS NOT NEEDED? THE IMPACT OF QUERY STRUCTURE ON WEB SEARCHING RESULTS

Bernard J. Jansen
School of Information Sciences and Technology
The Pennsylvania State University
001 Thomas Bldg
University Park, PA, 16801, USA
Phone: 814-856-6459 Fax: 814-865-6426
Email: jjansen@acm.org

TRACK:
Human Computer Interaction

ABSTRACT

Most Web searchers use extremely simple queries. There is an assumption that the correct use of query operators will improve the quality of results. We test his assumption by examining the impact of query operators on the documents retrieved from Web searching. We compare the results from queries without operators to results from the same queries using a variety of operators on several major web search engines. There were 1900 queries submitted, which returned 18,332 documents. In general, there was an average 66% similarity between results from the queries with and without operators. Implications on the effectiveness of current searching techniques, for future search engine design and of future research are discussed.

INTRODUCTION

The vast majority of Web queries contain no query operators (Hoelscher, 1998; Jansen, Spink, & Saracevic, 2000; Spink, Jansen, Wolfram, & Saracevic, 2002). The use of Boolean operators is typically about 8% in these Web searching studies. It has been assumed that correct usage of query operators would increase the effectiveness of Web searches. However, it appears that the majority of Web searchers continue to use very simple queries, with little to no use of query operators, even though many of these techniques (e.g., phrase searching and must appear operators) are easy to employ (Korfhage, 1997) and well known (Sullivan, 2000a). Web searchers seem to be employing an ineffective and inefficient strategy for finding information.

Studies and data suggest that Web users may be finding the information they want using simple queries, however. A survey of users on the Excite Web search engine reports that nearly 70% of the searchers reported locating relevant information on the search engine (Spink, Bateman, & Jansen, 1999). Web searchers are not utilizing advanced searching operators, but they appear to be finding information using a technique that should be ineffective or at least inefficient.

The objective of this study is to determine the effect of query operators on the results retrieved by Web search engines. This knowledge is essential to understanding how users search the web, for the development of instructional material for web searching, and for design of search interfaces the support the information seeking process. In this paper, we present an overview of related literature, research methodology, and research results from various perspectives. We end with a discussion of results and directions for future research.

RELATED STUDIES

Search engines are the major information retrieval (IR) systems for users of the Web, with 71% of Web users accessing search engines to locate other Web sites (CommerceNet/NielsenMedia, 1997). There are approximately 3,200 search engines on the Web (Sullivan, 2000b). Those utilized in this research are Alta Vista, Excite, FAST Search, GoTo, and Northern Light.

There have been relatively few studies comparing the retrieval results of different search engines using different approaches to query formulation(Eastman, 2002; Jansen, 2000). Eastman (2002) explored the precision of search engines using a variety of topics and query formulations. The researcher notes that precision did not necessarily improve with the use of the advanced query operators. We could locate no study focusing the on the change in results from a large number of queries across multiple search engines. Jansen (2000) examines the changes in results using a small sample, fifteen queries, and five search engines utilizing different searching operators. The researcher reports a 70% similarity in results between queries with no operators and the queries with operators.

There has been some research examining Web searching in general. Research shows that Web queries generally have two terms (Jansen, Spink, Bateman, & Saracevic, 1998; Silverstein, Henzinger, Marais, & Moricz, 1999), cover a variety of topics (Wolfram, 1999), and are primarily noun phrases (Jansen & Pooch, 2001; Kirsch, 1998). Other studies show that most Web searchers, usually about 80%, never view more than ten results (Hoelscher, 1998; Jansen et al., 2000; Silverstein et al., 1999). Examining Web information systems, the ability of Web search engines to successfully retrieve relevant documents has been investigated several times (Leighton & Srivastava, 1999; Zumalt & Pasicznyuk, 1998).

RESEARCH DESIGN AND METHODOLGY

We investigate the effect of complex queries (i.e., those using advanced syntax, such as Boolean operators) on the results retrieved by Web search engines relative to the results retrieved by simple queries (i.e., those with no advanced syntax).

Selection of Queries and Results

We randomly selected a stratified sample of 100 queries from an Excite search service transaction log. Queries of the following lengths were selected for this study: 10 queries of 4 terms, 31 queries of 3 terms, and 59 queries of two terms. Along with selecting the queries, there is the issue of results. Based on typical Web searcher behavior, only the first ten results in the results list were selected for comparison. We examined these results only for changes in the first ten results. We did not evaluate the results for relevance.

Searching Rules

These five search engines offer a variety of advanced searching options. Some searching options are available from each search engine's main search page, others on a 'power' searching page. For this research, only those advanced searching options available from the search engine's main page were utilized. Of the five search engines, two offer four advanced search options (+, ", AND, and OR) from the main page, and three search engines offer two advanced searching options (+, and "). All of the search engines offer dropdown boxes (e.g., language of results, document collections to search) for refining the search. When dropdown boxes were present on the main search page, the default options were utilized.

Research Structure

Each of the 100 original queries was submitted to the five search engines for a total of 500 queries. The query was then modified with the advanced searching operators supported by the respective search engines. The entire process of submitting the simple and advance queries took 5 minutes or less. For example, the simple query *digital library* could be modified using the must appear operator *(+digital*

+ *library*), phrase searching operator (*"digital library"*), the AND operator (*digital AND library*), and the OR operator (*digital OR library*). These modified queries are the complex queries.

RESULTS

Of the 500 simple queries, 498 returned at least 10 results. One query returned 3 results, and one query returned no results. Therefore, there were 4983 results to use as the baseline (i.e., 498 x 10 + 3). As stated, results that appeared greater than position 10 in the results list were not utilized. Of the 1400 complex queries, 1325 returned ten or more results. There were 31 queries that returned fewer than ten results but more than zero results. There were 44 queries that returned no results. Altogether, there were 13,349 results returned by the complex queries. Combined with the 4,983 results from the simple queries, a total of 18,332 results were used in the analysis.

The match had to be exact when comparing the results between the simple and complex queries. The documents listed had to be the identical page at the same site. Different pages from the same site were not counted as matches. The identical pages at different sites were not counted as matches. Furthermore, if results appeared in both lists but in a different order, they were counted as matches as long as both were listed in the first ten results.

Simple Versus Complex Query Comparison

The aggregate results of the analysis of the 18,332 results are displayed in Table 1.

Table 1: Comparison of Simple versus Complex Queries on Major Web Search Engines.

Category	Average Number of Matching Results	Standard Deviation	Mode	Paired sample t
Simple Queries	9.99	0.03	10	-
Complex Queries	6.55	3.77	10	40.287
				$p < 0.01$

The baseline mean for the simple queries was 9.99, and the mean for the complex queries was 6.55. This means that, on average, 6.55 of the ten results retrieved by the complex queries also appeared in the baseline results for the corresponding simple query on that search engine. The results were analyzed using a paired sample t test, as reported in Table 1. The analysis revealed a significant difference between the two groups ($t=40.287$; $p<0.01$).

Results by Query

The number of matching results by queries length is graphically displayed in Figure 1.

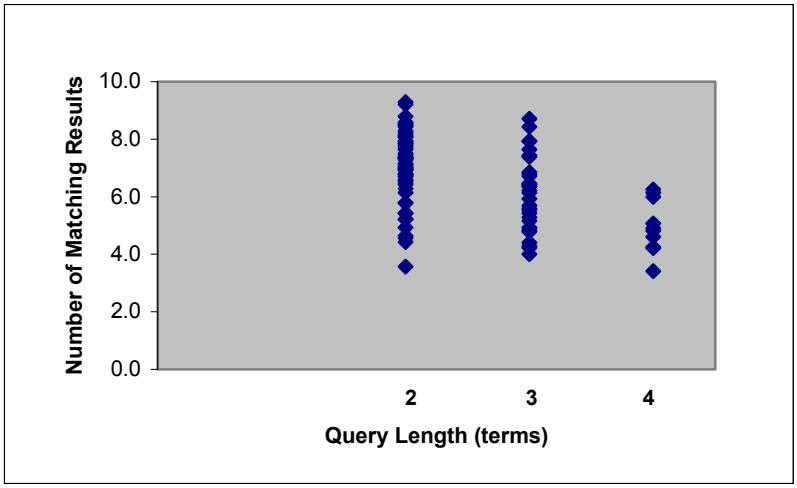

Figure 1: Number of Matching Results between Simple and Complex Queries

In terms of the number of matching results, the range for three and two term queries was similar (see Figure 1). The range for four-term queries was lower. However, there was a great deal of disparity among all three query lengths. One of the lowest, a two- term query, *internet capitalist,* had an average of only 3.6 matching results. One of the queries with the greatest number of matching results was a three-term query, *grape seed extract*. As a synopsis, the top 15 queries with the greatest overlap of results are displayed in Table 2.

Table 2: Comparison of Results by Query for Top 15 Queries.

Query	Average Number of Results that Appear in Baseline	Standard Deviation	Mode
bonsai trees	9.3	1.3	10
fuzzy logic	9.2	1.4	10
ear cleaning	8.8	2.0	10
grape seed extract	8.7	2.1	10
bread machines	8.6	1.9	10
self esteem	8.6	2.9	10
car insurance	8.5	2.8	10
bull riding	8.5	2.2	10
adult friend finders	8.4	2.5	10
morgan horse	8.4	2.2	10
nudist colonies	8.3	2.4	10
scream queens	8.2	2.5	10
neck pain	8.1	2.4	10
talk shows	8.1	3.2	10
poetry contest	8.1	2.7	10

Results by Search Engine

The analysis was conducted for query operators by search engine, with the results displayed in Table 3.

Table 3: Comparison of Results by Search Engine and Operator.

Matching Results	No.	AV	EX	FS	GT	NL	AV	Ex	FS	GT	NL	EX	NL	Ex	NL
		+	+	+	+	+	"	"	"	"	"	AND	AND	OR	OR
Average		3.0	7.9	10	9.4	10	6.0	5.0	3.8	7.2	6.0	7.9	3.1	10	2.57
SD		3.2	3.5	0.0	1.5	0.0	4.5	3.8	2.9	3.4	2.8	3.5	2.9	0.0	2.9
Paired sample t*		21.11	6.00	-	3.74	-	8.54	12.63	19.69	7.46	23.36	6.04	23.36	-	25.53
10	651	9	67	100	77	100	48	21	1	44	7	67	5	100	5
9	45	4	2	0	12	0	0	3	2	3	14	1	3	0	1
8	49	1	4	0	3	0	1	6	8	4	13	5	3	0	1
7	54	0	3	0	0	0	1	7	8	4	15	3	5	0	8
6	45	4	1	0	3	0	1	3	11	7	8	1	4	0	2
5	54	5	3	0	1	0	1	8	11	2	10	3	5	0	5
4	55	10	2	0	1	0	3	4	6	1	7	2	13	0	6
3	59	4	3	0	1	0	1	7	5	6	7	3	11	0	11
2	91	18	4	0	0	0	6	9	8	6	4	4	17	0	15
1	79	15	4	0	1	0	9	5	8	8	7	4	8	0	10
0	163	25	7	0	0	0	20	18	20	1	3	7	26	0	36
NR	55	5	0	0	1	0	9	9	12	14	5	0	0	0	0
Total	1400	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Note: (1) AV – Alta Vista, EX – Excite, FS – FAST Search, GT – GoTo, NL – Northern Light, NR – No Results returned by query. (2) Missing pair-t values could not be calculated due to a zero standard deviation. * $p<0.01$

The first column in Table 3 is the heading for the number of matching results. The top row lists the searching engine; the second row displays the corresponding advanced query operator. From each row in column 1, one can move right across the table to the occurrences for each in the *No.* column, which is the number of times that the results from the complex queries contained that number of exact matches. For example, there were 651 complex queries that return ten results identical to the corresponding simple queries. Moving further to the right, each column shows the number of occurrences for each search engine and operator for a given number of matching results. The average number of matching results and the standard deviation is also given.

As Table 3 illustrates, the effect of the specific operators varied depending on the search engine involved. With Alta Vista, the average for the must appear operator was half of what it was for phrase searching. With Excite, the average for phrase searching was about half of the other three operators. With FAST Search, there was a marked drop using phrase searching. The matching results of the GoTo operators were both greater than seven matches. The default algorithms for Excite, FAST Search, and Northern Light are illustrated with 100% matches between the simple and complex queries.

Table 3 shows that there were 651 (47%) complex queries that retrieved identical results as the simple queries. All ten results from these 651 complex queries were identical to the results from the simple queries. This occurrence is by far the most frequent; the next highest occurrence was 163 (12%) complex queries that retrieve no matching results.

The results were analyzed using a paired sample t test, as reported in Table 3, fifth row. The analysis revealed a significant difference between the results of each search engine operators relative to the results retrieved by the simple queries on the respective search engines, with the except of when there was no difference (i.e., noted as -).

We conducted a regression analysis to determine any significant relationship among the variables, query length, search engine, and query operator on the results retrieved. The overall model was significant ($F= 21.99$, $p<0.01$) with an R-squared of 0.05. Query length was a significant predictor of results ($t=-6.156$, $p<=0.01$), with a beta weight of -0.164. As query length increased, the number of matching results decreased. Query operator was also a significant predictor ($t=6.156$, $p<0.01$), with a beta weight of 0.145. Although significant, as the beta weights show, neither query length nor query operator had a substantial impact on the number of matching results. Search engine was an insignificant predictor of matching results.

DISCUSSION OF RESULTS

Approximately 66% of the results were identical regardless of how the searcher entered the query. Referring to the data displayed in Table 1, a paired sample t-test ($t= 40.287$, $p<0.01$) shows that the results from the simple queries are significantly different from the results for complex queries. However, the betas show that the impact of operators is relative low indicting that there are other factors that influence results. For example, terms have been show to impact query results (Spink & Saracevic, 1997). Additionally, as with all tests of statistical significance, one must ask "what different does this make in the 'real world'?".

Is it practical to learn and utilize the query operators if on average they are only going to present about three or four results that are different from those retrieved by just entering the query terms? Are the three or four different results worth the increased probability of entering a complex query incorrectly? As the complexity of queries increases, so does the probability of error.

The findings of this research suggest that the use of queries operators is generally not worth the trouble for the typical Web searcher (i.e., one who uses two terms and is interested only in the first ten results). Based on their conduct, it appears that most Web searchers do not think it is worth the trouble either. The relative precision of simple Web queries meets the information needs of most Web searchers.

In reviewing the analysis by search engine, outlined in Table 3, there was a great deal of overlap between query results for most search engines. The mode for all five search engines, regardless of search operator was ten. With Excite, 78% of the results are identical, regardless of the present or absence of advanced searching operators. Based on the rather random results retrieved, Alta Vista appears to adhere to the theoretical model of no ranking feature when Boolean-like operators are used in a query.

CONCLUSIONS AND FUTURE RESEARCH

This research indicates that use of complex queries appears to have a moderate impact on the results retrieved. Approximately 66% of the top ten results on average will be the same regardless of how the query is entered. Based on the actions of most web searchers, the approximately three or four different results may not be worth the increased effort required to learn the advanced searching rules or the increased risk of making a mistake.

Given that the typical web searcher seldom uses advanced operators, web search engines appear to be compensating for the searching characteristics of their users. Based on the results of this research, it appears that the ranking algorithms of these search engines generally adhere to the following rule: *Place those documents that contain all the query terms and that have all the query terms near each other at the top of the results list*. Although an over simplification of what can be complex algorithms, a ranking rule like this would negate the impact of most query operators for the topmost ranked documents.

This study measured the change in the results list of complex versus simple queries. The natural next step is to measure the change in precision. One might expect that the complex queries would improve precision (i.e., the ratio of retrieved relevant documents to the total retrieved documents); however, this assumption would have to be tested. Given the observed changes in ranking by some search engines, the introduction of Boolean operators may result in a precision decrease.

REFERENCES

CommerceNet/NielsenMedia. (1997). *Search Engines Most Popular Method of Surfing the Web* [web site]. Commerce Net/Nielsen Media. Retrieved 30 August, 2000, from the World Wide Web: http://www.commerce.net/news/press/0416.html

Eastman, C. M. (2002). 30,000 Hits May be Better than 300: Precision Anomalies in Internet Searches. *Journal of the American Society for Information Science and Technology, 53*(11), 879-882.

Hoelscher, C. (1998, July 1998). How Internet Experts search for Information on the Web. In *Proceedings of the World Conference of the World Wide Web, Internet, and Intranet*, Orlando, FL.

Jansen, B. J. (2000). An Investigation into the Use of Simple Queries on Web IR Systems. *Information Research: An Electronic Journal, 6*(1), 1-10.

Jansen, B. J., & Pooch, U. (2001). Web User Studies: A Review and Framework for Future Work. *Journal of the American Society of Information Science and Technology, 52*(3), 235-246.

Jansen, B. J., Spink, A., Bateman, J., & Saracevic, T. (1998). Real Life Information Retrieval: A Study of User Queries on the Web. *SIGIR Forum, 32*(1), 5-17.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *Information Processing and Management, 36*(2), 207-227.

Kirsch, S. (1998). *The future of Internet search (keynote address)* [website]. Keynote address presented at the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia. Retrieved 16 August, 1999, from the World Wide Web: http://www.skirsch.com/stk.html/presentations/sigir.ppt

Korfhage, R. (1997). *Information Storage and Retrieval*. New York, NY: Wiley.

Leighton, H., & Srivastava, J. (1999). First 20 Precision among World Wide Web Search Services (Search Engines). *Journal of the American Society for Information Science, 50*(1), 870-881.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum, 33*(1), 6-12.

Spink, A., Bateman, J., & Jansen, B. J. (1999). Searching the Web: A Survey of Excite Users. *Journal of Internet Research: Electronic Networking Applications and Policy, 9*(2), 117-128.

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From E-sex to E-commerce: Web Search Changes. *IEEE Computer, 35*(3), 107-111.

Spink, A., & Saracevic, T. (1997). Interaction in Information Retrieval: Selection and Effectiveness of Search Terms. *Journal of the American Society for Information Science, 48*(5), 382-394.

Sullivan, D. (2000a). *Search Engine Sizes*. Retrieved 30 August, 2000, from the World Wide Web: http://searchenginewatch.com/reports/sizes.html

Sullivan, D. (2000b). *Search Watch*. Search Engine Watch. Retrieved 1 June, 2000, from the World Wide Web: http://searchenginewatch.com/

Wolfram, D. (1999). Term Co-occurrence in Internet Search Engine Queries: An Analysis of the Excite Data Set. *Canadian Journal of Information and Library Science, 24*(2/3), 12-33.

Zumalt, J., & Pasicznyuk, R. (1998). The Internet and Reference Services: A Real-world Test of Internet Utility. *Reference and User Services Quarterly, 38*(2), 165-172.