CHARACTERISTICS OF THE BOOLEAN WEB SEARCH QUERY

Estimating success from characteristics

Sunanda Patro and Vishv Malhotra
School of Computing, Private Box 100, University of Tasmania, Hobart 7001 Australia

Keywords: Boolean query, web search engines, precision, recall.

Abstract: Popular web search engines use Boolean queries as their main interface for users to search their information needs. The paper presents results a user survey employing volunteer web searchers to determine the effectiveness of the Boolean queries in meeting the information needs. A metric for measuring the quality of a web search query is presented. This enables us to relate attributes of the search session and the Boolean query with its success. Certain easily identified characteristics of a good web search query are identified.

1 INTRODUCTION

Web search engines are central to the Web browsing today. Popular web search engines index the information on the Web in a manner that allows the web searchers to quickly locate some information of interest to them most of the times. These days few users bookmark or memorise many sites to surf the Web. A user survey (Spink, 2002) in 1999 reported that some 70% of the information-seeking web-interactions begin through a search engine. Another survey in 2000 (Holscher and Strube, 2000) reported this number to be 81%. The increased sophistication of the search engines would reduce the need to remember site addresses and increase the proportion of searches that begin from a search engine over time.

There have been many surveys to characterise user interactions with the search engines. Some researchers have focused on the transaction logs to determine nature of the search sessions and the characteristics of queries written by the users. Others have used questionnaire-based survey to determine what the searchers seek to do.

We have not seen any survey that relates the characteristics of a search session or the query with the success they have in satisfying users' information needs. The issue is important for it alone can tell us how effective are the Boolean query interfaces of the popular web search engines in helping the searchers satisfy their information needs by locating the best documents. A user survey (Broder, 2002) reports that over two-thirds of the users, when asked about the topic of their search, state: I seek *a good site on this topic, but I don't have a specific site in mind.* More than three-quarters of the surveyed users desire to access *the best site regarding this topic*.

Analyses of the web search logs, for example, Jones *et al* (1998), Jansen (2000), Jones *et al* (2000) however, suggest that users do not make full use of the resources and facilities provided by the search engines. A typical query tends to be simple with only one to three words or phrases in it. An average query session lasts only one or two attempts at query refinement. The anecdotal evidence suggests that the users seeking uncommon and poorly understood information face a difficult task in searching the Web for their information needs.

This paper reports a survey that relates the properties of the search sessions with the success achieved by the searchers in meeting their information needs. Unlike the logs of transactions used by the other researchers, we used volunteers to search the Web for specified information. The success of the human effort was compared against the search queries synthesised by a program. The program is known to perform better than humans at very high level of confidence (alpha=0.0001). This allows us to assign objective measures to the successes of human queries. However, it makes the process very labour-intensive. The volunteers had to download a lot of pages for their queries as well as for the synthesised query and then classify each document as being relevant or non-relevant to the information needs. This limits the amount of data we collect to draw inferences.

The paper is organized as follows. Section 2

Patro S. and Malhotra V. (2005).
CHARACTERISTICS OF THE BOOLEAN WEB SEARCH QUERY - Estimating success from characteristics.
In *Proceedings of the First International Conference on Web Information Systems and Technologies*, pages 339-344
DOI: 10.5220/0001226403390344

explores results from related works to lay down context and background for the survey results. Section 3 gives an overview of the survey and defines metrics to measure the quality of the queries. Section 4 contains the analysis of the data collected from the survey. Section 5 is the concluding section of the paper.

2 THE RELATED WORKS

The information seeking interactions over the Web often start with a search using a search engine. With the exception of dynamically changing information sources such as news items, search engines are effective tools for locating the information. However, it is not uncommon for the web-savvy, domain experts (Holscher and Strube, 2000) to bookmark certain specialised and trusted websites to quickly access information that they believe meets their need for quality. Likewise, other users too may bookmark some trusted sites for reference; for example, a site providing information about the medical emergencies.

The continuing need to bookmark the sites may be viewed as an evidence of a limitation of the current web search paradigm. Experienced web users report *lack of domain knowledge regarding the individual search question as a significant obstacle in construction of a query* (Holscher and Strube, 2000). Query construction remains a challenging task in certain circumstances for an average searcher.

An analysis of usage logs from a digital library (Jones et al 1998) and a direct user survey (Holscher and Strube, 2000) report that the user queries are short and generally consist of 1, 2 or 3 words (terms) only. Domain savvy users construct queries with well chosen but fewer words than those who do not know the search domain well. Average query lengths have been reported to be between 2 to 3 words. A typical searcher tends to rely on the default operator for the search engine to define their queries (Jones et al 1998). An average query is 2.21 words long with a standard deviation of 1.05. The reported length of a typical user session, as measured by the number of queries in a session, to satisfy the user's information need is also short. Few sessions extend beyond 4 queries. Average session length of 2.04 queries is reported by Jones (2000). Significantly, they report that 64.4% of the queries do not lead to the searcher viewing any document. The above reported facts combine to suggest that a session is concluded once the searcher has viewed a document. The average number of the documents viewed by the users has been reported to be about 2.5. No viewing together with viewing of one or two of the top ranked links account for over 90% of the post search viewing actions.

To summarise, the web search queries used for information searching are not very sophisticated. It is unclear how effectively the web searchers are able to locate the most relevant document for their information needs.

3 THE SURVEY AND METRICS

The survey reported in this paper explores the following question: How well does the common Boolean query paradigm supports human searchers in devising web search query that simultaneously support good coverage (recall) and good precision when the search domain is relatively unfamiliar to the searcher. Are there clues to spot a good query?

Before presenting the results from our survey, we define the metrics that we will use to measure the quality of the query.

3.1 Measuring Quality of a Query

Information retrieval literature (Chakrabarti 2003, Baeza-Yates and Ribeiro-Neto, 1999, Witten and Frank 2002) is the main influence on the web search practices and measurements. Recall and precision are often quoted as the common metrics. Given a corpus of documents containing r relevant documents and n non-relevant documents, let q be the number of documents a query selects. Suppose the precision of the query be p. That is, query selects q*p relevant documents and the remaining q*(1-p) documents are non-relevant to the information needs of the search. Recall for the query is computed as q*p/r. As a single metric to measure the quality of a query, F-factor or harmonic mean of the recall and precision is often used in information retrieval literature (Manning and Schtze 1999; Powers 2002). The measure is defined as $2/(1/p+r/(q*p))$.

However, the Web is a huge and ever-expanding collection of documents and resources. Nor is it fully indexed – only a fraction of the documents on the Web have been indexed by the search engines. Thus it is not possible to use traditional information retrieval based metrics to quantify the quality of a web search query. It is not possible to provide values for r for every conceivable information need.

Precision is commonly used measure of the web query quality. We will use P@20 – the number of relevant documents among the first 20 links returned by a query – as the measure of precision in this paper.

Recall is more difficult to determine. Instead we define *coverage* as a measure of the utility of the documents accessed by the query. Let e be the estimated number of documents a query returns. Google prints this estimate for each query. We prefer the coverage to have values in the same range as precision, 0 to 20. For a query with precision p that locates e documents, we estimate the number of relevant documents returned by the query to be e*p.

The marginal utility of the relevant documents to the searcher decreases with size. All reported surveys have suggested that users are more likely to view the top ranked links from the output of a web search query than those lower down (see for example, Jones, 1998 and Jones, 2000). We choose to use logarithm to base 2 of e*p as measure of the utility of the returned links. This utility will be used as the metrics to express query coverage.

The choice of the base is somewhat arbitrary but is motivated by the following remark in Jones (2000): *12.7% of all viewed documents were located at the first position in the result list. The next most common location was the second position (6.8% of viewed documents).* The interest has diminished to about one-half for the second ranked document.

3.2 How the Data was Obtained

To collect data for this exercise, we first identified a number of topics to search. Each topic area was identified by a single word keyword. For each of these topics we prepared a short description in the form a checklist to provide a consistent basis to describe the nature of information that is to be searched.

For each topic, the title keyword was used to search and download about 25 or more pages. These pages were then classed as *relevant* or *irrelevant* based on the checklist. The classified pages were then used to synthesise a query that selects relevant documents while rejecting irrelevant documents using the algorithm detailed in (Malhotra et al 2005).

In each survey session, a volunteer was given a topic to search along with the checklist and some sample relevant documents. All volunteers used Google search engine for their search starting with the single topic word as their first query. We shall refer to this query as a *naïve query*. The volunteer then refined the query to select the best collection of documents. No constraint was placed on the volunteer regarding the time, number of tries or quality of their query.

The volunteer specifies the final query when they have the query formed. In each session, twenty documents were downloaded using the volunteer's query and another 20 were downloaded using the synthesised query. The volunteer then classified the two sets of documents based on their understanding of the information needs consistent with the checklist provided to them.

A total of 39 sessions were surveyed. Some topics were searched by more than one volunteer. Likewise some volunteers helped us with more than one topic. Statistical summary of the data is provided in Table 1.

Table 1: Summarising statistics for the survey and other data reported in Table 1.

Statistics	Naïve 1-word query		User devised query		Synthesised query	
	P@ 20	cover age	P@ 20	cover age	P@ 20	cover age
Minimum	0.5	14.6	10	8.4	12	12.6
Average	6.3	18.6	14.9	14.4	18.2	15.8
Maximum	17	20	20	20	20	20
Median	6	19.4	15	13.9	19	14.9
Std. Dev.	3.9	1.6	2.8	2.8	1.8	2.5

4 ANALYSIS AND INFERENCES

4.1 Session Length and Termination Condition

A number of influences determine the perseverance of the volunteers to devise a query that they believe effectively satisfy the information need. The session lengths observed in the survey consists of one to three refinements (after the initial query) giving an average session length of 2.64. The session length in our survey matches with those observed by the other researchers using different sources of data. Average session lengths of 2.02 and 2.8 queries are quoted in (Silverstein et al, 1999) from different researchers.

Precision of the query emerges as one of the main criteria used by the volunteers to access the query quality. No session returning less than 10 relevant documents among the top ranked 20 retrieved links was observed in the survey. Also, we note that all human queries have precision equal to or above the original naïve query – no volunteer has returned the original naïve query as their final choice. Again we believe that the precision of the naïve query sets a lower-bound on the precision for the volunteers. Every volunteer tried to exceed this target value.

Human users (that is, volunteers) do not seem to regard coverage of the query as a vital factor. Some queries given by the volunteers had lower harmonic mean of coverage-precision combine then the

original naïve query. Yet our observations support the following hypothesis at 99.99% confidence level (alpha = 0.0001): F-factor value of the users' Boolean web search query is more than the corresponding value for a naïve single word query for the topic.

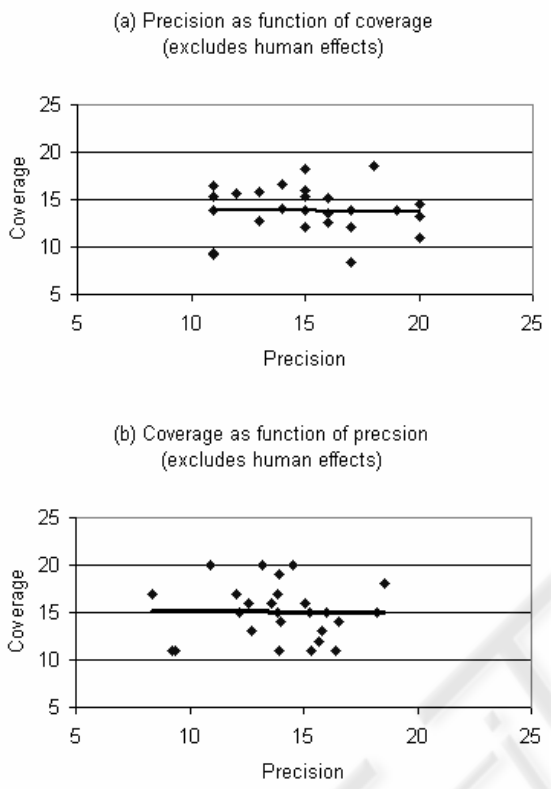

Figure 1: Relationships between precision and coverage of volunteer queries with the effects of query averaged out. Average value of the range variable emerges as the best prediction in these graphs. (Note: the trend line in (b) overlaps the grid line at C=15).

4.2 Trading Coverage for Improved Precision

The basic premise in devising a Boolean web search query to select relevant documents is that one can increase precision of the query by sacrificing some coverage. Ranking algorithms used by the web search engines to order the links play their part in this process.

This trade-off needs to be analysed to separate the effects of the query from those inherent in the topic due to the volume and nature of its presence on the Web. If we consider only those samples from survey which had the topics repeated, we may expect the effects of good queries cancelled against the poor queries for the same topic; thus the plotted relationships between precision (P) and coverage (C) will only be influenced by the Web specific properties of the topics.

Figure 1 depicts the relationship between precision and coverage for these samples in the survey. As is evident from these graphs, topic of the search does not contribute to any (positive or negative) trend between metrics precision and coverage. The correlation coefficient between these metrics is only 0.03 over these cases.

To determine the relationships between precision and coverage in the presence of query effects we consider all samples of the survey. In this case precision and coverage values for topics surveyed multiple times have been replaced by their average values. We note a significant correlation coefficient of 0.3 between the precision and coverage. In turn, as evident in Figure 2, the linear regression relationships between the metrics were determined to be: $P = 0.28C+10.8$ and $C = 0.32P+9.6$.

The positive correlation between precision and coverage and also between coverage and precision highlights a fundamental property of a good query. A better query is one that returns higher values for precision together with good coverage. Thus, F-factor is an appropriate measure of the query quality.

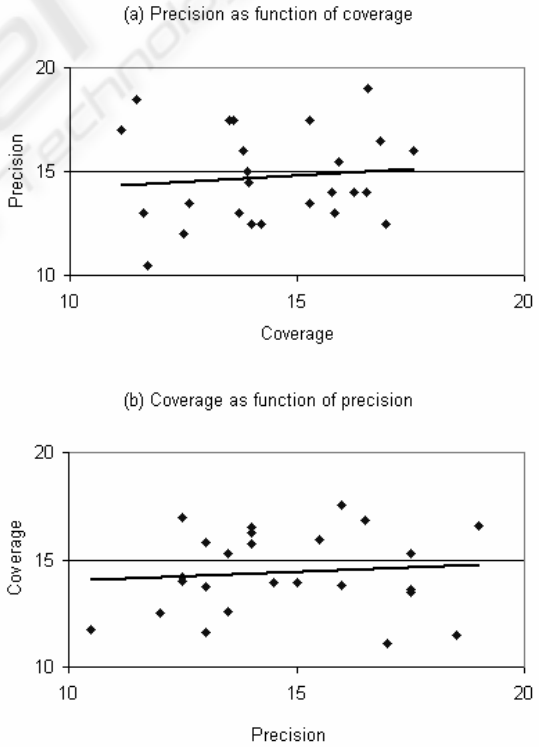

Figure 2: Observed relationships between precision and coverage induced by the variations in the query quality.

4.3 Characteristics of a Good Query

The number of attempts made by a volunteer to improve the query had palpable benefit to the precision of the query over the initial naïve query. Average increase in precision noticed from a single attempt to improve the query is 7.2, from two attempts the increase is 9.3, and from three attempts it is 13. A part of this improvement is attributable to the extra choices that become available from multiple attempts to pick the best case. At the same time, a committed volunteer is likely to make more attempts at improving the precision of the query; thus contributing to the observed trend. A quadratic trend line showing precision as function of attempts to improve is shown in Figure 3.

The number of terms (T) in a query is one of the primary characteristic of a Boolean query. The expected precision of a query increases nearly linearly with the number of terms in query (T): $P = 2.28T+7.4$ for number of terms $T<5$ (see Figure 4(c)). Thereafter, there is a drop in the average precision for T=5.

We had no case of a volunteer's best query with 6 or more terms. We believe that human users begin to have difficulties in effectively organising Boolean queries with 5 or more terms. A previously reported transaction log based analysis (Jones, 1998) has reported that less than 10% of user queries in the log records had 5 or more terms.

Table 2: Average query precision as a function of terms in query and number of attempts to improve query.

Terms in query	2	3	4	5	>5
Query improved 1 time	11	14.3	16	15	No case
Query improved 2 times	No case	13.7	16.8	13.3	
Query improved 3 times	Too few cases				

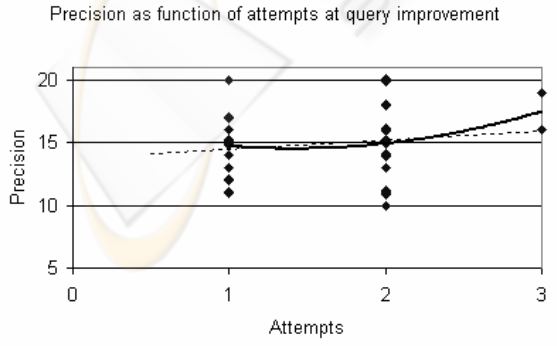

Figure 3: Average precision as a function of number of attempts to improve the query.

Figure 4 and Table 2 provide an evidence of a 3-way connection between the terms in a user query, number of attempts made to improve the query and the average precision of the queries. The average precision improves with the number of terms (T) up to 4 and then drops sharply as human ability to organise Boolean query with many terms declines.

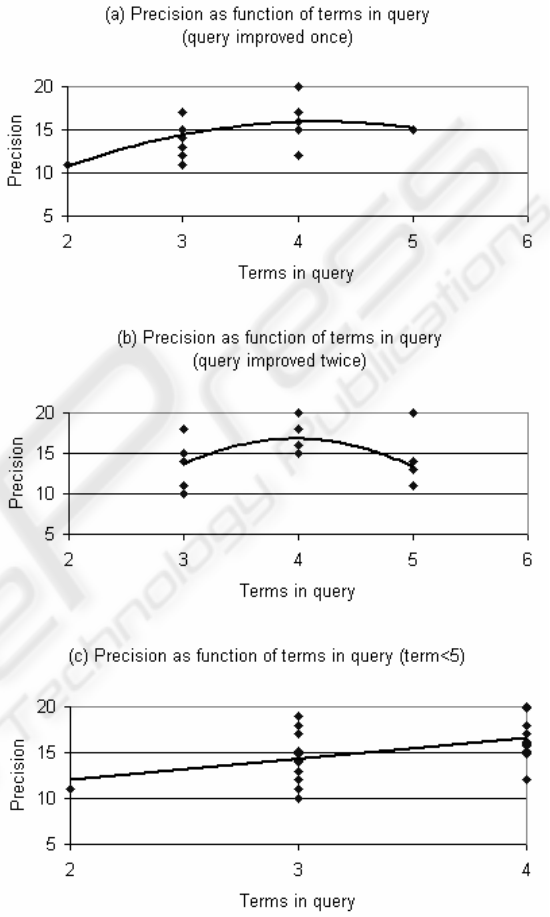

Figure 4: Precision of queries as function of terms in query. Two common cases of number of attempts to improve query are shown separately in (a) single attempt and (b) two attempts. Figure (c) shows linear relationship between number of terms and precision for up to 4 terms

The better-quality of the queries with 3 or 4 terms, evident in Figure 4, is further elaborated in Table 3 which shows the fractions of queries showing below average and above average performance within various groupings. First row shows the cases where the volunteer made only one attempt to improve the query and the final precision of the query was below average. Arguably in plain English, the specifying expression translates to *volunteer found the search difficult*. In this row the fraction of queries with below average precision

decreases with the number of terms in the query. The bottom row in the table shows that the fraction of queries with above average precision improves with the number of terms till it hits the high mark where humans start to be overwhelmed by the size of Boolean expression.

We suggest that a query with fewer than 3 terms alert us to the possibility that the searcher is finding it difficult to identify appropriate domain terms (or jargons) for the query. More than 4 terms makes it difficult to organise the terms in an effective Boolean query. The best performing query sizes, however, do not coincide with the most common query size. The most common size of the queries as reported in (Jones et al, 1998) is 2 terms and it accounts for about third of all queries.

Other researchers have reported that domain-savvy searchers use a small number of domain specific terms in their search query. Our observation is not inconsistent with those findings. To further test our inference, we grouped the volunteer queries into three nearly equal size groups based on their performance relative to the synthesised queries.

Queries with precision up to 2 units below the corresponding synthesised query were marked *good*. Those that had precision 5 or more units below the synthesised queries were marked *poor*. The group in the middle had 14 cases. Table 4 shows the distribution of terms in the two groups.

The proportion of queries in good group with 4 terms is about three times as high as in the poor performing group.

Table 3: Fraction of queries with stated precision characteristics as function of terms in query.

Terms in query	2	3	4	5
(precision of user query < average precision) among queries with single improvement attempt	100%	44%	20%	0%
Query precision > average precision	0%	50%	93%	29%

5 CONCLUSIONS

Four terms and above average coverage emerges as a good predictor of a successful Boolean web search query. Indeed, all samples with this property in our data have above-average precision of 15 or more. One-half of these queries achieve perfect precision score of 20.

To further improve odds for success choose 4-terms, above average coverage *with several attempts to improve query* Minimum precision delivered by these queries in our survey is 19.

Table 4: Distribution of terms in two groups of volunteer queries marked good and poor.

Number of terms	Good queries	Poor queries
Count	13	12
2	0%	8%
3	31%	42%
4	54%	17%
5	15%	33%
Total	100%	100%

REFERENCES

Baeza-Yates, R., and Riberio-Neto, B. (1999). *Modern Information retrieval*, Addison-Wesley, Reading, Ma.

Brin, S. and Page, L. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks* 30(1-7) pp. 107-117

Broder, A.Z. (2002). A taxonomy of web search. *SIGIR Forum* 36(2) pp. 3-10

Chakrabarti, S. (2003). *Mining the Web: Discovering knowledge from hypertext data*, Morgan Kaufmann Publishers, Amsterdam

Hölscher, C. and Strube, G. (2000). Web search behavior of Internet experts and newbies. *Computer Networks* 33(1-6) pp. 337-346

Jansen, B.J. (2000). The effect of query complexity on web searching results, *Information Research*, 6(1)

Jones, S., Cunningham, S.J. and McNab, R. (1998). Usage Analysis of a Digital Library. In: *Third ACM Conf. on Digital Libraries, Pittsburgh, PA, USA*. June 23-26.

Jones, S., Cunningham, S.J., McNab, R.J., Boddie, S.J. (2000). A transaction log analysis of a digital library. *Int. J. on Digital Libraries* 3(2) pp. 152-169

Malhotra, V., Patro, S. and Johnson, D. (2005). Synthesise web queries: Search the Web by examples, *Int. conf. on enterprise information systems, Maimi, Florida.*

Manning, C.D. and Schtze, H. (1999). *Foundations of statistical natural language processing*, MIT press, Cambridge, MA

Powers, D.M.W. (2003). Recall and precision versus the bookmakers, *Joint International conference on cognitive science*. pp. 529-534

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

Spink, A. (2002). A user-centered approach to evaluating human interaction with Web search engines: an exploratory study. *Inf. Process. Manage*. 38(3).

Witten, I.H. and Frank, E. (2000). *Data mining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann Publishers, San Francisco.