Web Query Characteristics and their Implications on Search Engines

Jason Zien, Jörg Meyer, John Tomlin
IBM Almaden Research Center
650 Harry Rd., K57/B2
San Jose, CA 95120-6099
(jasonz,jmeyer,tomlin)@almaden.ibm.com

Joy Liu
U.C. Berkeley
2635 Etna St.
Berkeley, CA 94704
joyliu@uclink4.berkeley.edu

ABSTRACT

The rapid growth of the World Wide Web presents significant challenges to the implementers of high performance Web search engines. There are two dimensions to this growth, the rapidly increasing number of web pages and the growing population of users. Understanding the search behavior of users is critical to the overall design of a web search engine. We study the characteristics of a large Web query log and assess the impact of these queries on search engines. In particular, we will look at the properties of vocabulary growth, and term occurrences. We also analyze the temporal qualities of search queries and classify them into three categories: hot, popular, and unpopular.

Keywords

query, logs, search, engine, caching, Web, performance

1. INTRODUCTION

We gathered a total of 50,538,653 web queries from the WebCrawler search engine from March 22, 2000 to May 26, 2000 using their Search Ticker [3]. A web query is defined as the exact string typed in by a user searching for data, and may contain one or more terms as well as special query operators. We found that each query had an average of 3.3 terms (we did not filter out stop words) which is roughly the same as the 3.34 found by Spink [5] (though they did filter out stop words). This is significantly higher than the 2.2 term average reported by Kirsch [4], and is in part due to the fact that over 17.9% of the queries were natural language questions.

2. VOCABULARY GROWTH

The vocabulary used by search engine users can impact a search engine in two ways. The first way is the reduction of the index size by pruning useless terms. The second way is through caching of frequently used terms or queries to improve performance. It would be ideal if the vocabulary (either the terms or the exact queries) used by search engine users stayed within some small subset or grew very slowly.

Let V be the vocabulary size, n be the number of vocabulary elements (tokens, terms, query strings, etc.) in a data set, and K and β be constants which are dependent on the particular text of the data set. It has been generally observed that vocabulary size grows according to the following formula: $V = Kn^{\beta}$ where β is typically between 0.4 and 0.6, so vocabulary in a document collection typically grows proportionally to the square root of the words in the documents. This is known as Heaps' Law [1]. We found that Heap's Law also applies to web queries. We considered two cases:

1. Each query was taken to be a single word in the vocabulary, shown in the upper curve of Figure 1. The curve is almost linear (a curve fit gives the equation $V = 1.53n^{0.95}$). This is very discouraging news, as it means that almost all of the queries are unique in our query logs.

2. Each term in the query (ignoring capitalization and ignoring modifiers) was considered to be a single word in the vocabulary. The lower curve in Figure 1 shows a much smaller slope than the previous case (a curve fit gives the equation $V = 6.63n^{0.69}$), growing just slightly faster than the square root of the text size. There is thus great potential for the caching of query terms and their associated posting lists.

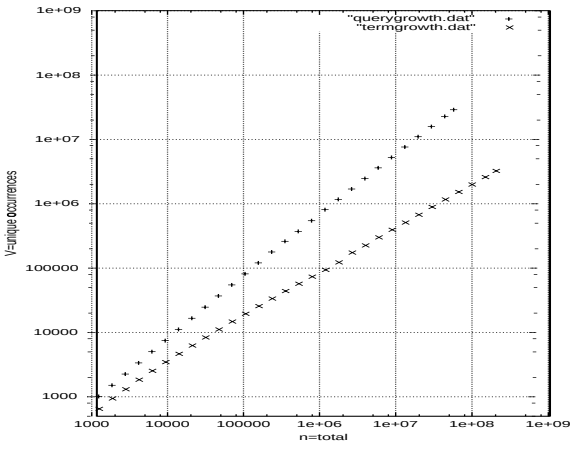

Figure 1: Query vocabulary growth (upper curve) and term vocabulary growth (lower curve).

Indeed, the potential for caching of query terms is evident by looking at the occurrence frequencies of the most popular terms (Table 1). The top 3.2% (100,000) terms make up over 95.1% of the total terms in the queries. Another implication is, dropping a substantial number of terms in a search engine's dictionary and index should have relatively little effect on the user.

Top X Terms	% of Total	% of Distinct
100	35.9	.0032
1,000	58.1	.032
10,000	83.5	.32
100,000	95.1	3.2

Table 1: Percentage of terms accounted for by the most frequent query terms.

3. TERM-DOCUMENT DISTRIBUTIONS

Do people most frequently ask about the terms that are the most written about on the web? We studied this question by gathering statistics on the query terms and estimated posting list sizes. We compiled a list of the occurrence frequencies of the 10,000 most frequently asked query terms (ignoring capitalization and query operators). For each term, we performed a query against the search engine AllTheWeb in July 2000 (which at the time indexed 340 million unique pages) and extracted the total number of documents found, which corresponds approximately to the size of the posting list for a term.

A scatter plot of the term occurrences versus document occurrences is shown in Figure 2. There is an overall trend for terms that occur frequently in queries to also be terms that occur frequently in documents. Caching of only hundreds to thousands of the top term postings would be feasible, however, because of the Zipfian distribution of the term frequencies, even small caches may lead to significant improvements in performance as shown in Table 1. Since there is a direct correspondence between query term occurrences and document occurrences, it is reasonable to drop terms from the index that occur very infrequently in documents, which is already done in practice [2].

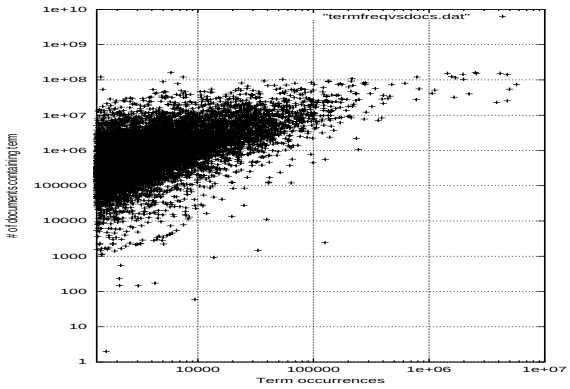

Figure 2: Term and Document Distributions

4. QUERY TERM CLASSIFICATION

Typically, terms are classified into two implicit categories: those that occurred frequently (the ones that were ranked in the top X) and those that did not (unpopular). We show that there are two distinct subclasses of terms with high frequency. First, there are the *popular* terms, that consistently occur with high frequency. Second, there are *hot* terms, which occur with high frequency for a short time period. To demonstrate that distinction, we need to do a time analysis of term occurrences. We break down our entire time period into smaller time intervals and identify the frequently occurring query terms in each time interval. Then we produce an overall view which shows a graph of the number of frequently occurring terms that occurred in 1 through T time intervals.

We chose to break our 64 day time period into four hour intervals, so $T = 384$. We chose $f = 101$ as the minimum cut-off frequency of the terms. A term was counted for an interval if it appeared at least 101 times during that interval. Figure 3 shows the distribution of frequently occurring terms. The y axis indicates the total number of distinct terms which were popular for x time intervals. The curve is a bathtub curve. The left end of the x axis shows that there are a significant number of terms that were popular for brief durations of time (hot terms) while the right end of the x axis shows that there were a significant number of terms that occurred frequently during almost every time interval (popular terms). It is also interesting to note that 80% of the frequently occurring terms occur in 20% of the range (the left 10% and right 10% of the x axis). A good caching algorithm must not only capture popular terms, but also dynamically capture terms that suddenly become popular and also throw out terms that just as quickly lose their popularity.

Please refer to our research report [6] for a complete version of this paper.

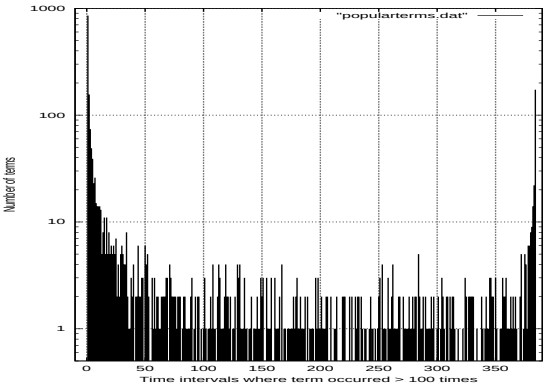

Figure 3: Hot terms are on the left and Popular Terms are on the right

5. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, New York, NY, 1999.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Seventh International World Wide Web Conference (WWW7)*, April 1998.

[3] Excite@Home. Webcrawler search ticker. *http://www.webcrawler.com/cgi-bin/SearchTicker*.

[4] S. Kirsch. Searching the internet, sigir '98 keynote. *http://www.skirsch.com/presentations/sigir.ppt*, August 1998.

[5] A. Spink, J. Bateman, and Major B. J. Jansen. Searching heterogeneous collections on the web: Behavior of excite users. *Information Research*, 4(2), October 1998.

[6] J. Zien, J. Meyer, and J. Tomlin. Web query characteristics and their implications on search engines. Technical report, IBM Research Division, San Jose, CA, Novemeber 2000. RJ 10199.