A Query-Level Examination of End User Searching Behaviour on the Excite Search Engine

Dietmar Wolfram

University of Wisconsin– Milwaukee

Abstract

This study presents an analysis of selected characteristics of a large set of queries submitted by end users to the Excite search service. Characteristics examined include the frequency distribution of queries, browsing persistence at the query level based on query size, and browsing persistence at the session level based on query session length. Findings reveal that users submit largely unique queries and engage in similar, non-persistent browsing habits. Users are not browsing much beyond the first one or two pages of results regardless of the effort they put into their queries or the number of queries submitted during a search session.

1. Introduction

End user access to and use of Information Retrieval (IR) systems has greatly increased in recent years due to the development of Internet-based search services and wider availability of Internet access. In addition to traditional bibliographic-based IR systems such as online public access catalogs in libraries and specialized databases offered through online vendors, users are now able to access immense quantities of electronic resources through the Internet. To better understand how users interact with IR systems, researchers have undertaken user-based studies of prototype and commercial systems. Much of the user-oriented IR research to date, however, has taken place in laboratory environments where users are closely monitored and are often assigned information seeking tasks. These controlled studies typically make use of small samples of users and may not reflect how typical end users interact with IR systems.

Internet search services provide excellent environments in which to unobtrusively study end user searching on a large scale. Many of these services process millions of queries daily by large numbers of users from around the world. Despite the growing popularity of Internet services for general information searching, until recently very few studies have been undertaken that report on information searching behaviours using these services. One of the primary reasons for the lack of published research in this area has been the limited availability of data sets to researchers outside of the search service companies. Understandably, these companies wish to protect the privacy of their users while also guarding strategic information on their products in a highly competitive environment. Excite@Home has been one of the few commercial search services to make available several data sets to members of the research community. One of the more recent sets consists of more than one million queries submitted to Excite on a single day. Selected aspects of this data set are reported here.

The investigation of the growth and use of resources on the Internet has been increasing in importance over the last several years. Studies have investigated general search habits (Huberman, Pirolli, Pitkow & Lukose, 1998; Lawrence & Giles, 1998; Pitkow & Kehoe, 1996) and search engine coverage of the World Wide Web (Lawrence & Giles, 1999). Evaluations of previous Excite data sets have been undertaken by Jansen, Spink & Saracevic (1998, in press), Spink, Bateman & Jansen (1998, 1999), and Spink, Chang, Goz & Jansen (1999). These studies have examined searcher behaviours using an earlier set of data consisting of approximately 50,000 Excite queries. The authors found empirical regularities in the usage of query terms and the query size that are consistent with more traditional informetric studies. They also found that fewer than 10% of queries submitted used Boolean operators. In examining the use of query reformulation and relevance feedback, Spink *et al*. (1999) concluded that, although relevance feedback was not widely used, the query patterns appear to indicate that this feature was useful to users much of the time. In a more recent

study, Silverstein, Henzinger, Marais & Moricz (1999) analyzed approximately 285 million search sessions submitted to the AltaVista search service over a six-week period. The authors' findings were similar to the Excite studies, where most queries were short, browsing of retrieved items was brief, and little query modification was used during a search session.

The purpose of the present study is to analyze specific characteristics of user search behaviours and query construction. Basic query and user characteristics of the one million query Excite data set are examined elsewhere (Spink, Wolfram, Jansen & Saracevic, submitted; Ross & Wolfram, in press; Wolfram, 1999). Spink *et al*. (submitted) report that searchers use few search terms, rarely modify queries, view few Web pages, and made infrequent use of advanced search features. These previous studies have investigated regularities in the way terms are used within queries, but the regularities of the content of the queries themselves have not been analyzed for the Excite data set. Interactions between selected query and search habit characteristics are also presented here. By combining query attributes such as the number of terms used per query, the number of pages browsed, and queries submitted per user session, one might investigate whether query construction or session lengths influence browsing behaviour. Of particular interest are the intersections between combinations of characteristics to determine if search behaviours differ among users or if query characteristics produce different browsing habits.

The following questions are examined:

1. *Is there a discernable pattern of frequency of use of specific queries across all users in the data set?* Frequency distributions of query terms reveal a traditional inverse relationship with lengthy tails found in many informetric phenomena. Can the same be said for complete queries, or are the search interests of users so diverse that a relatively short-tailed distribution is observed, indicating that few queries are submitted with high frequency?

2. *Does the number of terms in a query influence browsing persistence?* The size of a query may influence how much time a user is willing to invest in browsing,

measured by the number of pages of retrieved hits viewed. Do longer queries result in different browsing habits than brief queries?

3. *Does the number of queries users submit in a search session influence browsing persistence?* A search session consists of all the unique queries associated with a specific user identifier. Do the browsing habits differ between lengthy and brief search sessions?

Examination of these search characteristic combinations will shed light on whether differences exist in user search behaviours, which may inform systems design research. Future systems may then incorporate features that accommodate these differences, if they exist.

2. Methodology

Raw data from the Excite data set were imported into a Microsoft Access database. The data included numeric identifiers for searchers/machines, time information on when each query was submitted, and the full queries entered by users. The 1,025,910 queries represent a subset of queries submitted to the Excite search engine on a single day and only include submissions where an identifier, stored as a cookie on a user machine, was available. Queries submitted from machines on which the browser cookie facility was disabled were not included.

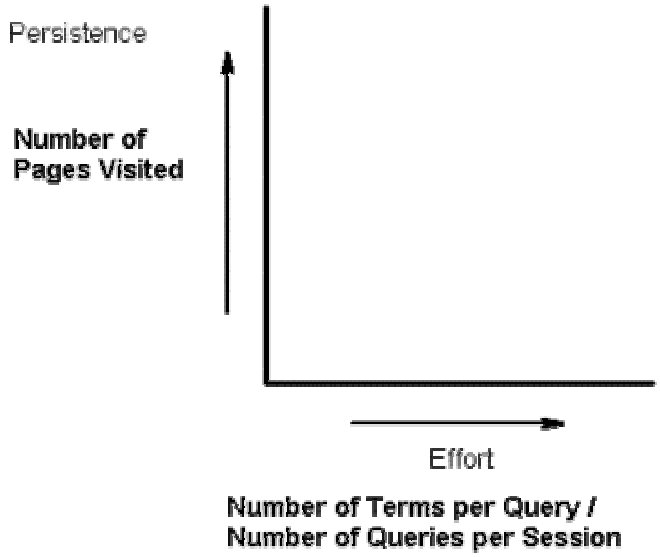
Queries were initially parsed for individual terms based on alphanumeric starting characters. Terms were delimited using spaces and other non-alphanumeric characters. Non-alphanumeric characters that were part of a URL or email address (‘ /’ , ‘ .’ , ‘ @’ , ‘ :’) as well as hyphens were not treated as delimiters. Occurrences of the words ‘ and’ , ‘ or’ , and ‘ not’ were treated as terms since the context of their use as Boolean operators or as parts of phrases was not readily determinable without a query by query analysis. Other modifiers allowed by Excite for inclusion or exclusion of terms (+, -) were ignored. Repeated queries (i.e. identical queries submitted by a user appearing in succession) were assumed to represent requests to view the next page of results

associated with the initial query. These were not treated as separate queries, but rather as extensions of the same query.

To date, studies of end user web searching have looked at search characteristics individually. In the present study, the author investigates the intersection of two characteristics to determine if query characteristics (effort) influence browsing habits (persistence) (Figure 1). The number of terms in a query provides an indication of searcher effort invested in the formulation of a query. The number of pages viewed represents searcher persistence in browsing returned hits. By looking at both characteristics for each query, a determination can be made whether a relationship exists between the two.

The number of terms associated with each query and the number of pages browsed for each unique query were tabulated. The total number of pages browsed beyond the minimum for each query within each session was also tabulated. From the set of non-repeating queries for each user/machine identifier, unique queries were identified (regardless of the identifier) and tabulated. From the original query set, 363,282 unique queries were identified. Visual inspection of the plotted data sets revealed highly skewed distributions, so non-parametric methods were used, where appropriate, to analyze differences in the data distributions.

Figure 1: User Query Formulation and Browsing Characteristics



3. Results

3.1 Distribution of Query Frequencies

The twenty-five most frequently appearing queries submitted by different user ids appear in Table 1. Twenty of the most frequent queries consist of a single word, while the remaining five consist of two words. The most frequently input query was input by just over 1% of all users within the data set. Eleven of the top 25 queries represent requests for ‘ adult-oriented’ material, with other queries relating to net-based topics, leisure activities, general information, and information about specific people. The frequency of use of queries quickly drops off.

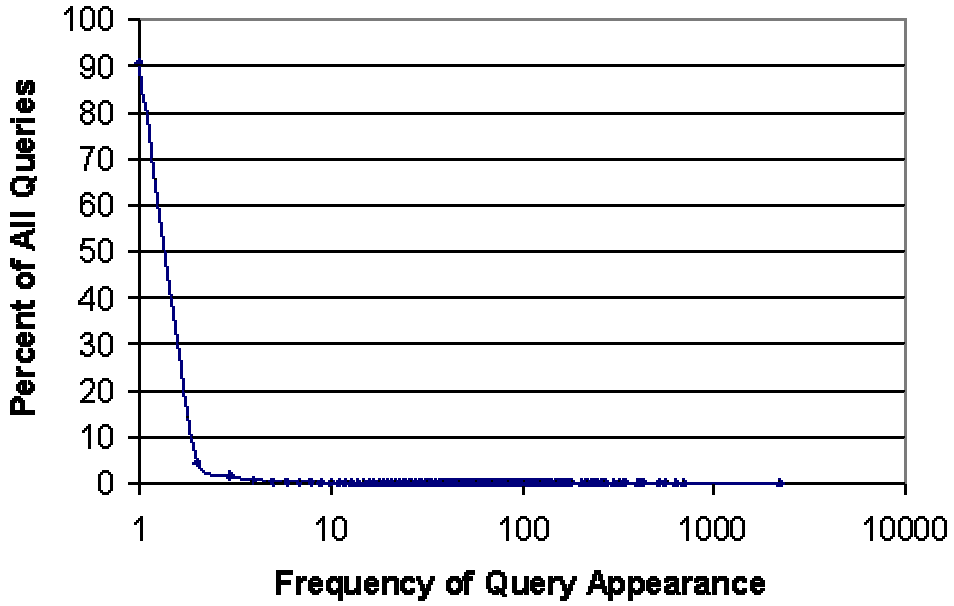
Table 1: Twenty-five Most Frequently Occuring Questions

Query	Frequency	Query	Frequency
sex	2235	warez	304
yahoo	707	music	304
playboy	637	persian kitty	273
chat	563	games	273
porn	515	nude	263
princess diana	429	jobs	261
xxx	417	pamela anderson	249
p****	408	horoscope	247
porno	403	weather	244
jokes	345	maps	244
hotmail	338	penthouse	242
chat rooms	321	beastiality	241
		erotic stories	237

**** = expletive

Figure 2 summarizes the distribution of query frequencies by percentages. The vast majority of queries in the dataset (90.6%) appear only one time, indicating that searchers are entering largely unique queries, and are not just limiting themselves to a small number of popular queries. The distribution has a comparatively short tail, with the 25 most frequently entered queries representing 2.1% of all queries within the data set. The most popular queries are similar to those found by Silverstein *et al.* (1999) in their study examining queries submitted to AltaVista. Fourteen of the top 25 queries were common to both lists.

Figure 2: Distribution of Query Frequencies



3.2 Query Size and Page Browsing

The overall distribution of terms used per query is highly skewed with a mode of two terms and a mean of 2.4 terms per query indicating most queries are quite short (see Spink, *et al.*, submitted). Frequency of occurrence values for each query size were converted to percentages to allow visual comparisons to be made between the distributions of different query term sizes. The small numbers of queries submitted containing more than 10 terms were combined into a single category (‘ >10’). Since Excite only processes the first 10 terms of lengthy queries submitted, which could influence searcher browsing behaviours, these queries were best combined.

Overall, the resulting distributions appear quite similar (see Figure 3). Mean and median numbers of pages visited for each query size appear in Figure 4. Visual inspection of the percentages for each query size reveals a similar distribution for the number of pages viewed, indicating that browsing persistence does not seem to be affected by the amount of effort (query size) invested by the user, although there does appear to be a small decrease in the average number of pages viewed with an increase in the number of terms. The higher mean value in the number of pages viewed for queries of greater than 10 terms could be attributed to the smaller number of queries in this group, resulting in a larger fluctuation of the sample mean from the population mean. The non-normal distribution of the data, large differences in the cell sizes, and large numbers of ties in the ranks of the data make traditional parametric (one-way ANOVA) and even non-parametric (Kruskal-Wallis test) analysis of the set unfeasible. A chi-square test of the counts of pages browsed across the terms per query categories with data for 10 or more pages browsed collapsed into a single category was conducted to determine if there were any significant differences in the proportional allocation of pages browsed for each query size. The resulting χ^2 total of 466 (90 d.f.) is significant at $\alpha = 0.01$, indicating sizeable departures in the distributions that are not apparent from the visual inspection of the percentage allocations. The largest departures between observed and expected cell values arose for the smaller number of terms per query and the smaller number of pages browsed (Figure 5), representing the cells with the largest number of queries. With many tens of thousands of queries in each of these cells, even small departures from the expected values quickly become large by the squaring of the difference. Observed values were not consistently above or below expected values, indicating that there was no pattern to the differences between the observed and expected values.

Figure 3: Distribution of Pages Browsed by Query Size (Percentages)

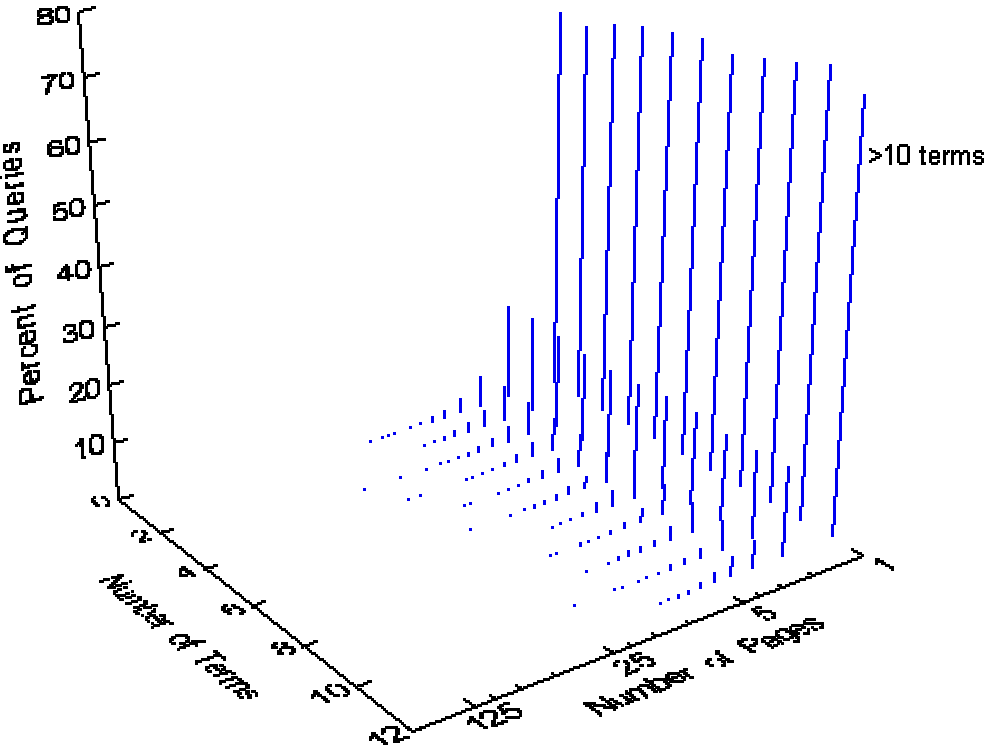


Figure 4: Mean and Median Number of Pages Browsed by Query Size

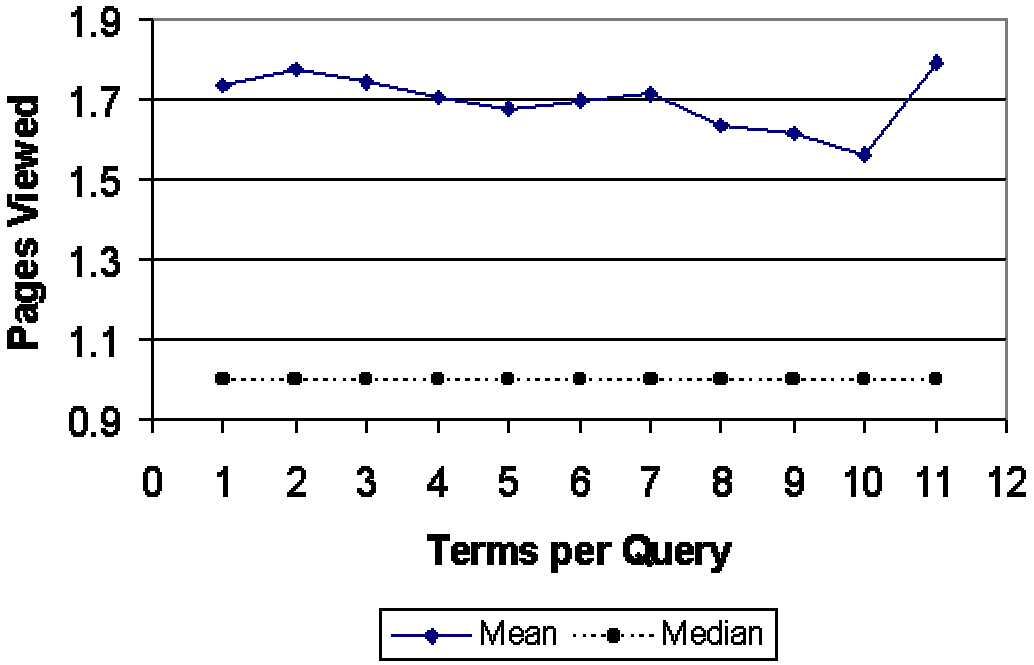
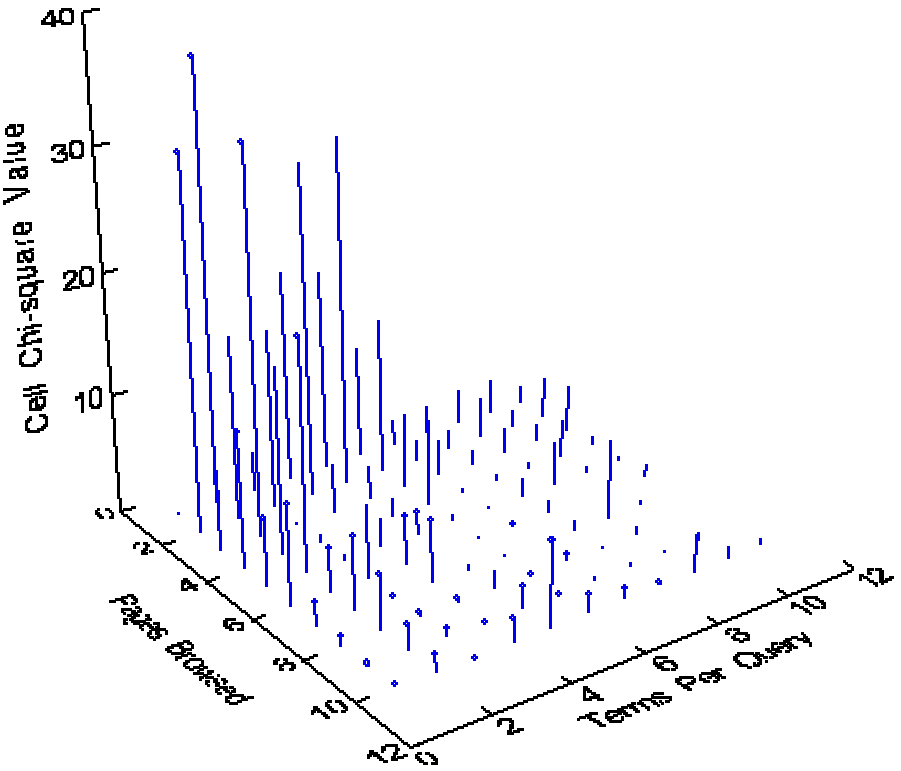


Figure 5: Chi-square Values for Cell Comparisons for Pages Browsed by Query Size



3.3 Session Duration and Page Browsing

The number of pages viewed beyond the minimum number for each session also provides an indication of user persistence during information searching. Figure 6 summarizes the distributions of the number of pages viewed beyond the minimum number for search sessions consisting of one to 15 queries. Beyond 15 queries the number of observations becomes too small to provide a reasonable comparison. Mean values for additional pages viewed per query for different session lengths are presented in Figure 7. The distribution of pages visited for different session sizes are clearly different. The median and mode of the pages visited increases with additional queries, while the mean remains relatively constant. The differences in the forms of the distributions prevent the use of traditional parametric and non-parametric tests.

In retrospect, the resulting differences in the distributions are not surprising. With the extra queries being undertaken, added opportunities exist for users to view additional pages. Wider fluctuations in the average values for additional pages viewed for sessions consisting of more than 10 queries are undoubtedly due to

the smaller sample of queries in these categories, resulting in larger variations in values. With no apparent decrease in the average number of additional pages viewed for lengthy query sessions, it appears that users are investing roughly the same effort into each query regardless of the session length. Since the average number of additional pages viewed is less than one in each case, one can conclude that users are not willing to invest significant extra effort in their page viewing, regardless of the number of queries or information needs they wish to fill.

Figure 6: Distribution of Pages Viewed by Session Duration (Percentages)

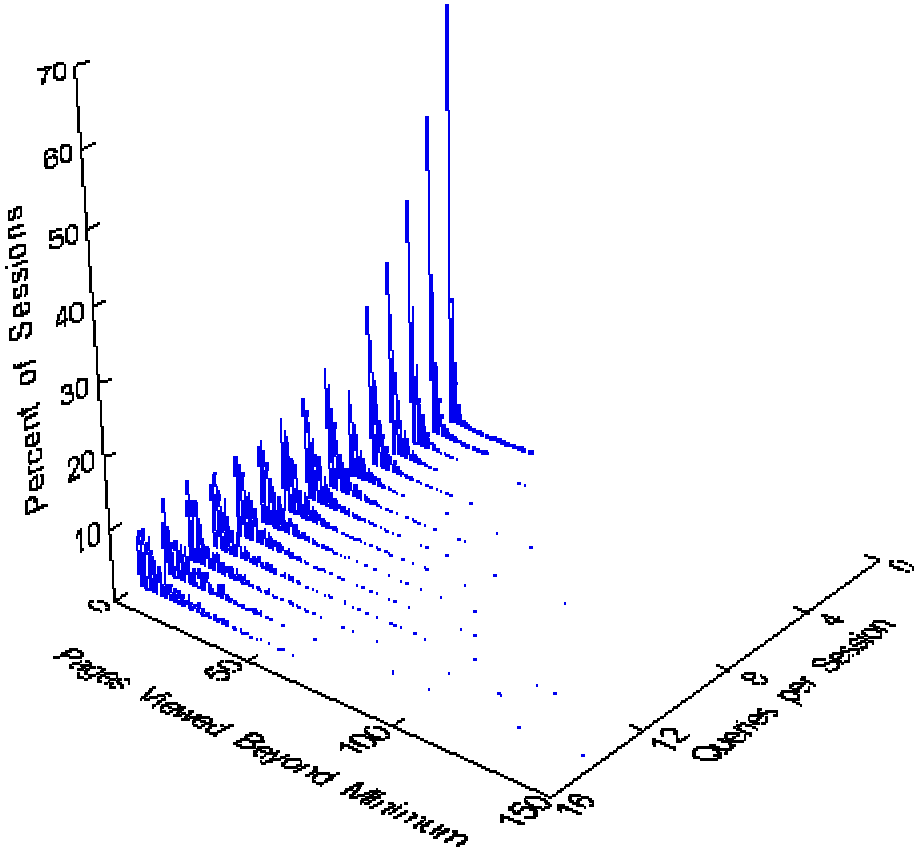
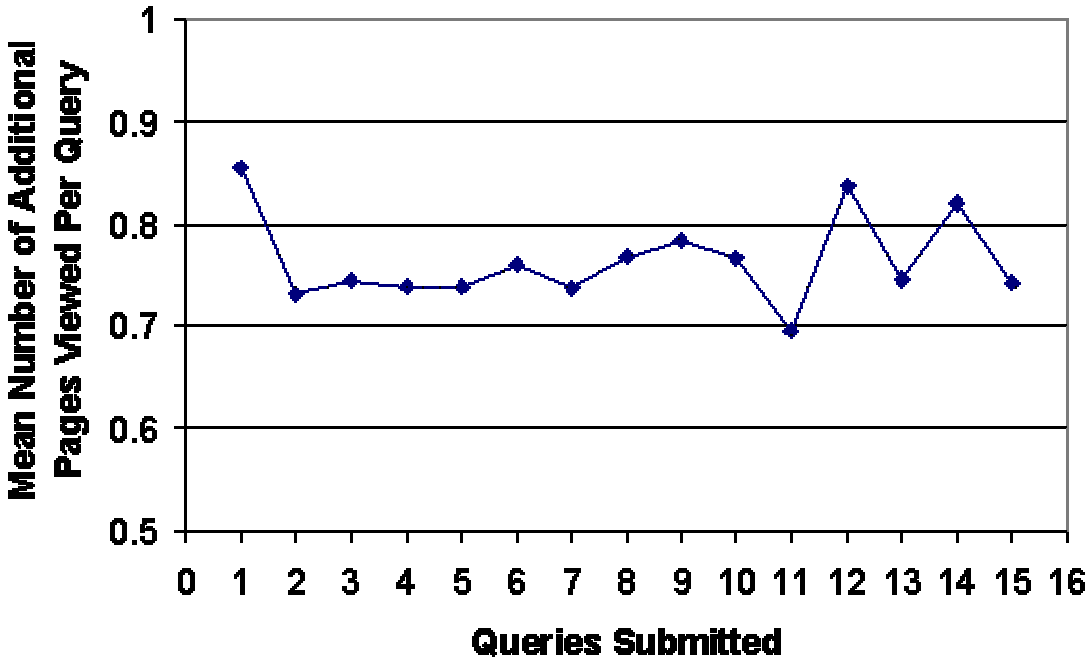


Figure 7: Mean Number of Additional Pages Viewed per Query by Session Duration



4. DISCUSSION

Topic content of complete queries cannot be generalized based on the most frequently occurring queries. They represent only a small percentage of all queries submitted. Although the largest single category of topics appears to be sexual in nature, this cannot be extrapolated to the complete query set. Ross & Wolfram (in press) concluded from their study of topic content of queries based on co-occurring term pairs in the Excite set that the frequency of sexually-oriented queries may be due to searchers' use of a terse vocabulary for this topic area, which results in the same queries being submitted numerous times. Other topics areas with a richer vocabulary may be distributed across a broader range of queries, with a smaller number of submissions for each when considered separately. The huge number of unique queries appearing in the data set makes manual investigation of query topic content cumbersome.

The precise reasons for the number of pages browsed cannot be determined, raising one of the limitations of the present study and for query log analysis studies in general. Users may have simply abandoned their search, or were satisfied with the results they received and discontinued viewing additional pages of links. Other studies that have surveyed Internet users on their search habits

(Logan & Driscoll-Eagan, 1998; Spink, Bateman & Jansen, 1998) indicate that users stop searching for a variety of reasons, including frustration or satisfaction with the results. Larger samples of users are needed to confirm or expand upon these findings. Based on the uniformity of browsing habits, it is clear that users expect quick retrieval and are not willing to engage in additional browsing to satisfy their information needs, or it is possible that users are coming up with sufficient numbers of relevant hits in the first few pages, making additional browsing unnecessary. This seems unlikely when one considers the potentially large number of hits returned for the many queries that consist of only one or two terms. Single term queries by their very nature are not very specific unless they relate to known item search topics such as URLs, for example. The overall browsing pattern remains largely the same whether queries contain few terms or many terms. The slight decrease observed in the mean number of pages viewed for queries with larger numbers of terms may be attributable to the greater specificity within at least some of these queries, resulting in relevant items appearing closer to the top of the ranked lists. Or, the query specificity results in a smaller number of potentially relevant hits for viewing. The end result in each case is that fewer additional pages will be viewed. The decline in the number of pages browsed for these higher-term queries must be verified with additional data sets before more conclusive statements may be made.

What are the implications of the present findings? Users are quite homogenous in their browsing habits, but largely unique in their query submissions. Preprocessing of frequently entered queries in anticipation of their submission to provide better system responsiveness does not appear to be warranted, unless the query consists of a single word that is among the most frequently entered terms (Wolfram, 1999). Response customization based on query size or session persistence also does not appear to be needed. By providing more hits per page beyond the typical 10 hits, users will be introduced to additional sites of potential relevance without requiring additional effort. This will, however, increase the amount of data to be transmitted to the user’ s browser for each page. Without

additional data on user motivation for discontinuing searches, conclusions are at best speculative.

5. Conclusions

Analysis of query-level regularities in submissions to search engines can provide a better understanding of user search characteristics that may not be evident in a term-level analysis. An analysis of query-level data reveals that users are engaged in many search topics with little repetition in query content among users. Browsing habits of users are largely uniform, regardless of the investment in individual query formulation, or the number of query topics undertaken per user session. Without access to the users themselves to understand their motivations for their querying and browsing strategies, only educated guesses can be made. However, researchers may still investigate the factual characteristics of user searches. Larger studies that take into account user satisfaction with results and comparisons of query sizes may reveal additional insights into user search motivation and perceptions of the systems they are searching.

With additional query data sets, longitudinal studies may be undertaken to determine if user query formulation and browsing habits change over time. With the increasing availability of natural language interfaces that encourage longer queries, changes seem likely. A more recent set of Excite data consisting of 2.5 million queries is currently being examined and will be compared to the one million query data set for changes in query formulation, term usage, and browsing habits over time.

Acknowledgements

The author would like to thank Excite@Home for making the query data accessible and Nancy Ross for research assistance.

References

Huberman, Bernardo A., Peter L. T. Pirolli, James E. Pitkow, & Rajan M. Lukose. 1998. Strong regularities in World Wide Web surfing. *Science* 280 (April 3): 95-97.

Jansen, Bernard J., Amanda Spink, Judy Bateman, & Tefko Saracevic. 1998. Real life information retrieval: A study of user queries on the web. *SIGIR Forum* 32 (1): 5-17.

Jansen, Bernard J., Amanda Spink, & Tefko Saracevic. In Press. Real life, real users, and real needs: A study of user queries on the web. *Information Processing and Management*.

Lawrence, Steve & C. Lee Giles. 1998. Searching the World Wide Web. *Science* 280 (April 3): 98-100.

Lawrence, Steve & C. Lee Giles. 1999. Accessibility of information on the web. *Nature* 400 (July 8): 107-109.

Logan, Elisabeth, & Lori L. Driscoll-Eagan. 1998. Is searching the Internet really different? Search process models for two electronic environments. In *Proceedings of the 26th Annual Conference of the Canadian Association for Information Science*, ed. Elaine G. Toms, D. Grant Campbell & Judy Dunn. Toronto: CAIS.

Pitkow, James E., & Colleen M. Kehoe. 1996. Emerging trends in the WWW user. *Communications of the ACM* 39 (6): 106-108.

Ross, Nancy C. M. & Dietmar Wolfram. In Press. End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine. *Journal of the American Society for Information Science*.

Silverstein, Craig, Monika Henzinger, Hannes Marais, & Michael Moricz. 1999. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(3).

Spink, Amanda, Judy Bateman, & Bernard J. Jansen. 1999. Searching the web: A survey of Excite users. *Internet Research*. 9(2), 117-128.

Spink, Amanda, Judy Bateman, & Bernard J. Jansen. 1998. Searching the web: A survey of Excite users. 1998. Users' searching behavior on the Excite web search engine. In *National Online Meeting Proceedings*, ed. Martha Williams. New York: Information Today.

Spink, Amanda, Carol Chang, Agnes Goz, & Bernard J. Jansen. 1999. Users' interactions with the Excite web search engine: A query reformulation and

relevance feedback analysis. In *Proceedings of the 27th Annual Conference of the Canadian Association for Information Science*, ed. James Turner. Toronto: CAIS.

Spink, Amanda, Dietmar Wolfram, Major B. J. Jansen & Tefko Saracevic. Submitted. A large study of users' queries on the web.

Wolfram, Dietmar. 1999. Term Co-occurrence in Internet search engine queries: An analysis of the Excite data set. In *Proceedings of the 27th Annual Conference of the Canadian Association for Information Science*, ed. James Turner. Toronto: CAIS.