The 4th International Conference on Internet Computing. Las Vegas, Nevada, p. 65-69. 23 – 26 June 2003.

An Analysis of Web Documents Retrieved and Viewed

Bernard J Jansen and **Amanda Spink**
School of Information Sciences and Technology
The Pennsylvania State University
004C Thomas Building, University Park PA 16802
Tel: (814) 865-4454 Fax: (814) 865-5604
E-mail: jjansen@ist.psu.edu, spink@ist.psu.edu

Abstract

The placement of Websites in ranked retrieval and the viewing patterns of Web search engine users is a crucial issue for Web site owners and Web search engines. However, little large-scale research has examined the viewing patterns of users of commercial Web search engines. The research results reported here address three questions, which are: (1) How many pages of results do Web search engine users' examine? (2) How many Web documents do Web search engine users' view when searching the Web?, and (3) How relevant are the Web documents that they are viewing? We present findings from large-scale research into the page viewing patterns of users of the FAST commercial Web search engine. Using data samples representing thousands of users, we examine common patterns concerning the number of pages of results viewed, the number of pages viewed and the relationship between the number of queries, the number of actual Web sites visited, and time between multiple site visits. The implications for Web search engines and services, Web sites and Web users are discussed.

Keywords: Web searching, pages viewed

INRODUCTION

The Web has dramatically changed the way people locate information. As the Web is becoming a worldwide phenomenon, we need to understand what searching trends are emerging, from how Web search engines are utilized in the search process to the viewing of Web documents. There is a growing body of Web research concerning how users interact with Web search engines. There are some reports on the number of result pages viewed. However, to our knowledge there has been no large-scale research examining the interactions between Web search engine users and the Web documents that they viewed.

In this article, we address this shortcoming in the literature by examining the page viewing activities of users of a major Web search engine. We examine general searching characteristics including the number of result pages viewed. We then examine the number of pages that users view, analyzing the relationship between sessions, queries, and pages viewed. We also explore the temporal relationships of these interactions.

We begin with a review of the literature, followed by the methodology utilized to obtain and analyze actual Web queries. We use these queries to examine trends in searching and page viewing or click through data (i.e., the web page/s a user visits when following a hyperlink from a search engine results page), including the temporal aspects of this viewing. We then discuss the implications of these results for Web search engine users, search engine designers, and the designers of Web sites. We conclude with the impact of our findings and directions for future research.

LITERATURE REVIEW

Little research has examined the results viewing patterns of Web search engine users. There is a growing body of literature in information science that examines how people search on the Web [1-4]. This research provides some insight into how people search for information on the Web, and provides a framework for considering the Web document viewing and search process. Jansen and Pooch [1] present an extensive review of the Web searching literature, reporting that Web searchers exhibit different search techniques than do searchers on other information systems. Hölscher and Strube [3] examine European searchers and report information on sessions, queries, and terms, noting that experts exhibit different searching patterns than novices. Jansen and colleagues [4] conducted an in-depth analysis of the user interactions with the Excite search engine. Spink and colleagues [2] analyzed trends in Web searching, reporting that Web searching has remained relatively stable over time, although they noted a shift from entertainment to commercial searching. This stream of research provides useful information and a methodology for examining Web searchers and their patterns of results viewing.

In general, Web searching sessions are very short as measured in number of queries. There has been less analysis of session temporal length, but it is assumed to be short. Users view a very limited number of results pages[1]. From our studies above, this implies that the majority of Web searchers, approximately 80%, view no more than 10 to 20 results. The page viewing characteristics of Web searchers have not been analyzed at any finer level of granularity. We do not know how many Web documents Web searchers actually view (i.e., Pages Viewed). In this study, we seek to address these issues by examining the page viewing patterns of actual Web search engine users.

RESEARCH QUESTIONS

More specifically, the overall research questions driving this study are:

(1) How many pages of results do Web search engine users' examine?

[1] When a Web search engine user submits a query, the search engine returns the results in "chucks", of usually about 10 results. These "chucks" are referred to as *Results Pages* and are presented to the user sequentially from the top most ranked results page to the maximum number of results retrieved by the search engine.

(2) How many Web documents do Web search engine users' view when searching the Web?

(3) How relevant are the Web documents that they are viewing?

To address the first research question, we obtained, and quantitatively analyzed, actual queries submitted to AlltheWeb.com, a major Web search engine owned by FAST. From this analysis, we could determine the number of results pages the searcher viewed. In addition to capturing the user's query, we also captured the Web document that the user viewed for each query, which addresses the second research question. For the third research question, a subset of queries from this transaction log was submitted to a major Web search engine. The retrieved sites were evaluated to determine whether or not they contained relevant materials.

Data Collection

The queries examined for this study were submitted to FAST[2], a major Web search engine on 6 February 2001 and spans a 24-hour period. They were recorded in a transaction log and represent a portion of the searches executed on the Web search engine on this particular date. The transaction log held a large and varied set of queries (over one million records).

Each record within the transaction log contains three fields: (1) *Time of Day*: measured in hours, minutes, and seconds from midnight of each day as logged by the Web server; (2) *User Identification*: an anonymous user code assigned by the FAST server; (3) *Query Terms*: terms exactly as entered by the given user, and (4) *Page Viewed*: the uniform resource locator that the searcher viewed after entering the query. With these fields, we located a user's initial query and recreated the chronological series of actions by each user in a session. In our analysis, we generally use the procedure and terminology outlined in [1].

Data Analysis

With these three fields, we located the initial query and recreated the chronological series of actions in a session. A term is any series of characters separated by white space. A query is the entire string of terms submitted by a searcher in a given instance. A session is the entire series of queries submitted by a user during one interaction with the web search engine. A results page is the chuck of results presented by the search engine. The Web page is the Web document located at the uniform resource locator presented by the Web search engine in the results page.

When a searcher submits a query, then views a document, and returns to the search engine, the FAST server logs this second visit with the identical user identification and query, but with a new time (i.e., the time of the second visit). This is beneficial information in determining how many of the retrieved results the searcher visited from the search engine, but unfortunately it also skews the results in analyzing how the user searched on system.

To address the first research question, we collapsed the data set by combining all identical queries [1] submitted by the same agent to give the unique queries in order to analysis sessions, queries and terms and pages of results viewed.

For the second research question, we utilized the complete un-collapsed sessions in order to obtain an accurate measure of the temporal length of sessions and the number of pages visited.

For the third research question, we randomly selected 530 records from the transaction log. Each record contained the query submitted by the Web search engine user and the Web page viewed after the user submitted that query. Three independent raters reviewed these 530 queries for relevance, assigning a binary relevance judgment of 1 (for relevant) or 0 (for not relevant) based on the rater's interpretation of the query.

Relevance is a standard measure utilized in information retrieval to evaluate the effectiveness of a query based on the documents retrieved [5]. The reviewers received training regarding the judgment process and were given written instructions for determining relevance. Agreement across the three raters was calculated using r_{wg}, and was found to be quite high (r_{wg}=0.95). From these relevance rankings, we were able to calculate relative precision (i.e., the ratio of the number of relative documents retrieved to the number of documents retrieved at a certain point in the results listing).

RESULTS

General Searching Characteristics

Table 1 presents an overview of the analysis.

Table 1: Overview of Transaction Log

Sessions	153,297	
Queries	451,551	
Terms		
Unique	180,998	13%
Total	1,350,619	
Session size		
1 query	81,036	53%
2 queries	28,117	18%
3+ queries	44,144	29%
Pages of Results		
1 page	244,441	54%
2 pages	86,976	19%
3+ pages	43,509	27%

Overall, the relationship between the number of sessions and queries, the ratio of unique terms relative to the total number of terms, and the percentages of pages viewed correspond closely to that reported in other Web searching studies [6, 7], leading us to believe that the data from this transaction log represents searches submitted by the typical population of Web users.

Number of Result Pages Viewed

From Table 1 some patterns emerge. Some 53% of the users entered one query and about 54% of the users viewed only one page of results. The relationship between the number of queries submitted and the number of results pages viewed

[2] http://www.alltheweb.com/

parallels each other with about equal percentages of queries submitted and results pages viewed. This may imply some relationship between the sufficiency [8] of the retrieved results relative to the user's information need.

Table 2 presents a more in-depth analysis of the number of pages viewed per query submitted.

Table 2: Number of Results Pages Viewed

Number of Results Pages Viewed	Occurrences	Percentage
1	24,4441	54.1%
2	86,976	19.3%
3	43,509	9.6%
4	24,880	5.5%
5	14,999	3.3%
6	9,706	2.1%
7	6,583	1.5%
8	4,570	1.0%
9	3,219	0.7%
10	2,479	0.5%
>10	1,912	2.3%

There is a sharp decrease in the number of viewings between the first and second and the second and third pages of results, with very few users viewing more than four or five pages of results. In line with results from previous Web studies, Web users have a low tolerance for wading through large numbers of results.

Web Documents Viewed By User

Although most users viewed only the first one or two pages of results, this does not tell us the actual number of Web pages they actually visited. They may have viewed all results presented or they may have viewed none. To address this issue, Table 3 shows the number of results viewed per session.

Table 3: Pages Viewed Per Session

Number of Results Viewed	Occurrences	Percentage
1	42,499	27.62%
2	22,997	14.95%
3	15,740	10.23%
4	11,763	7.65%
5	9,032	5.87%
6	7,157	4.65%
7	5,746	3.73%
8	4,563	2.97%
9	3,869	2.51%
10	3,308	2.15%
> 10	26,062	16.94%

The mean number of Web results viewed was 8.2, with a standard deviation of 26.9. Previous studies report that most Web searchers rarely few more than the first result page, which is usually 10 results. While 10 results is in line with the average, our analysis shows that over 66% of searchers examine fewer than 5 results in a typical session and almost 30% view only one document in a given session.

Web Documents Viewed By Query

This low number of viewed results hold when we move from the session level of analysis to the query level. Table 4 presents the number of Web results viewed per query.

Table 4: Results Viewed Per Query

Number of Results Viewed	Occurrences	Percentage
1	274,644	54.3%
2	95,532	18.9%
3	47,770	9.4%
4	27,625	5.5%
5	16,800	3.3%
6	11,024	2.2%
7	7,653	1.5%
8	5,231	1.0%
9	3,802	0.8%
10	2,975	0.6%
> 10	12,498	2.5%

The mean number of results viewed per query is 2.5, with a standard deviation of 3.9. FAST users viewed 5 or less documents per query over 90% of time. The largest number of users by far viewed only one result per query, just fewer than 55%.

Session Duration

Table 5 presents the session duration, as measured from the time the first query is submitted until the user departs the search engine for the last time (i.e., does not return).

Table 5: Session Duration

Session Duration	Occurrences	Percentage
< 5 minutes	55,966	26.2%
5 to 10 minutes	13,275	6.2%
10 to 15 minutes	41,987	19.7%
15 to 30 minutes	19,314	9.1%
30 to 60 minutes	30,955	14.5%
1 to 2 hours	8,691	4.1%
2 to 3 hours	21,901	10.3%
3 to 4 hours	2,635	1.2%
> 4 hours	18,605	8.7%

With this definition of search duration, we can measure the total user time on the search engine and the time spent viewing the first and all subsequent Web documents, except the final document. Unfortunately, this final viewing time is not available since the Web search engine search records the time stamp. Naturally, the time between visits from the Web document to the search may have not been entirely spent viewing the Web document.

However, this may not be a significant issue as shown from the data in Table 5. The mean session duration was 2 hours, 21 minutes and 55 seconds, with a standard deviation of 4 hours, 45 minutes, and 36 seconds. However, we see that the longer session durations skewed our result for the mean. Fully 52% of the sessions were less than 15 minutes. This is inline with earlier reported research on Web session length [9]. Over 25% of the sessions were less than 5 minutes.

Document Viewing Duration

While session length has been address, what has not been previously reported in the literature is the duration of pages viewed by Web search engine users, which is presented in Table 6.

Table 6: Duration of Page Views

Page View Duration	Occurrences	Percentage
< 30 seconds	46,303	13.9%
30 to 60 seconds	16,754	5.0%
1 to 2 minutes	48,059	14.5%
2 to 3 minutes	16,237	4.9%
3 to 4 minutes	47,254	14.2%
4 to 5 minutes	15,203	4.6%
5 to 10 minutes	47,254	14.2%
10 to 15 minutes	14,047	4.2%
15 to 30 minutes	41,215	12.4%
30 to 60 minutes	9,054	2.7%
> 60 minutes	30,592	9.2%

The mean time spent viewing a particular Web document was 16 minutes and 2 seconds, with a standard deviation of 43 minutes and 1 second. However, some lengthy page viewed skewed our mean. Over 75% of the users viewed the retrieved Web document for less than 15 minutes. More surprisingly, nearly 40% of the users viewed the retrieved Web document for less than 3 minutes. Just fewer than 14% of the users viewed the Web document for less than 30 seconds. These results for Web document viewing are substantially less than has been previously reported, using survey data [10].

RELEVANCE OF VIEWED PAGES

This portion of the study involved using a random subset of records from the FAST transaction log, which included the Web site the searcher actually visited. Three independent raters visited the sites and evaluated the Web document to determine relevance. This analysis helps address the question of whether search sessions are short because the searchers are finding the information that they need or that they are not finding the information they need and just giving up or going elsewhere. The results are reported in Table 7..

Table 7: Relevance Results for Pages Viewed

Relevance Score	Number of Documents	Percentage
3	199	37.5%
2	74	14.0%
1	103	19.4%
0	154	29.1%
	530	

We had the three independent raters view 530 URLs and evaluate these pages for relevance based on their interpretation of the query submitted. Each rater assigned a relevance Web document a rating of 1. A non-relevant page received a rating of 0. So, the maximum score a Web page could receive was 3, meaning that all three reviewers rater the page relevant.

Approximately 52% of the time, two or more rater evaluated a page to be relevant. Over 48% of the time, two or more raters evaluated a page to be not relevant. These percentages, taking in total, represent precision for this set of results retrieved by this search engine. This confirms earlier survey data that users were finding relevant finding on Web search engines [11].

DISCUSSION OF RESULTS

There are some clear patterns concerning the number of results pages viewed by FAST users. Approximately 54% of the users view only one page of results. This result is similar to the percentage of users that enter only one query (53%) and the percentage of relevant documents (52%). The similarity among these percentages would seem to indicate several things. One, the information needs of a majority of Web searchers are not extremely complex, given they require only one query. Two, Web search engines appear to do a good job of indexing and ranking Web documents in response to these queries, based on the majority of users viewing only one results page. Three, it appears that on average about 50% of the documents viewed will be relevant, implying that the typical Web user will have to view about two Web documents to find a relevant document. This is supported by our analysis of Web documents viewed, with 43% of users in our sample viewing two or fewer Web documents.

From our results, Web search engine users on average view about 8 Web documents. However, our analysis shows that over 66% of searchers examine fewer than five with more than one in three Web searchers viewing only one document in a given session. Users on average view about 2 to 3 documents per query. Over 55% of Web users view only one result per query.

Not only are the session lengths of Web search engines users short in terms of number of queries submitted and documents viewed, but also they are also short temporally. Over half the sessions were less than 15 minutes and about twenty-five percent of the sessions were less than 5 minutes.

The mean time spent viewing a particular Web document was just over 16 minutes. However, 75% of the users spent less than 15 minutes viewing the retrieved Web document. Twenty percent of the Web users view a Web document for less than a minute. These results would seem to indicate that the initial impression of a Web document is extremely important as Web searchers are typically not going to spent a great deal of time combing the document to find the relevant information.

From our analysis, it appears that generally the precision Web users can expect is about 50%, meaning that one out of every two of the Web documents viewed will be relevant to their query. Given the large number of documents that most Web search engines retrieve, fifty percent is rather high. However, note that this analysis is for Web documents viewed, not documents retrieved. This is has great implications for Web search engines and Web page designer. It is clear the Web search engine users are making relevance determination based solely on the document summary that is displayed in the search engines results page.

This study contributes to the Web searching literature in several important ways. First, the data comes from users submitting real queries and viewing actual Web pages. Accordingly, it provides a realistic glimpse into how users search, without the self-selection issues or altered behavior that can occur with lab studies or survey data. Second, our sample is quite large, with approximately 150,000 users. Third, we obtained data from a very popular search engine and conducted our relevance analysis on one of the largest search engines on the Web in terms of both document collection and number of unique visitors to ensure that our results were generalizable. Finally, it provides a detailed examination of the Web document viewing patterns and viewing duration of Web users.

As with any research, there are limitations that should be recognized. The sample data comes from one major Web search engine, introducing the possibility that the queries do not represent the queries submitted by the broader Web searching population. However, Jansen and Pooch [1] have shown that characteristics of Web sessions, queries, and terms are very consistent across search engines. Another potential limitation is that we do not have information about the demographic characteristics of the users who submitted queries, so we must infer their characteristics from the demographics of Web searchers as a whole. Finally, we do not have information about the browsing patterns of the users once they leave the search engine to visit a Web document. It is possible that they are browsing using the hypermedia structure of the Web. However, given that the duration between departing and returning to the search engine, this is unlikely in most situations.

CONCLUSION

Our results provide important insights into the current state of Web searching and Web usage. The short sessions lengths, combined with short queries have been puzzling issues for designers of Web information systems. This does not seem to be a successful strategy to maximize recall or precision, the standard metric for information retrieval system performance. However, it appears that Web search engine users are finding relevant information with this searching strategy. This may indicate the need for new metrics for evaluation of Web information systems.

ACKNOWLEDGEMENT

We thank FAST.com and especially Per Gunan Auran for providing the Web query data sets.

REFERENCES

[1] B. J. Jansen and U. Pooch, "Web User Studies: A Review and Framework for Future Work," *Journal of the American Society of Information Science and Technology*, vol. 52, pp. 235-246, 2001.

[2] 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.

[3] 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.

[4] 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.

[5] T. Saracevic, "Relevance: A Review of and a framework for the thinking on the notion in information science," *Journal of the American Society of Information Science*, vol. 26, pp. 321-343, 1975.

[6] 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.

[7] A. Montgomery and C. Faloutsos, "Identifying web browsing trends and patterns," *IEEE Computer*, vol. 34, pp. 94-95, 2001.

[8] B. J. Jansen, A. Spink, and T. Saracevic, "Searchers, the Subjects they Search, and Sufficiency: A Study of a Large Sample of EXCITE Searches," In Proc. of the 1998 World Conference on the WWW and Internet, pp. Orlando, FL, 1998.

[9] D. He, A. Göker, and D. J. Harper, "Combining Evidence for Automatic Web Session Identification," *Information Processing & Management*, vol. 38, pp. 727 - 742, 2002.

[10] CyberAtlas, "November 2002 Internet Usage Stats", Retreived from the World Wide Web on 1 January 2002 from http://cyberatlas.internet.com/big_picture/traffic_patterns/article/0,,5931_1560881,00.html.

[11] A. Spink, J. Bateman, and B. J. Jansen, "Searching the Web: A Survey of Excite Users," *Journal of Internet Research: Electronic Networking Applications and Policy*, vol. 9, pp. 117-128, 1999.