An Analysis of Document Viewing Patterns of Web Search Engine Users

Bernard J. Jansen
School of Information Sciences and Technology
The Pennsylvania State University
2P Thomas Building
University Park PA 16802
Tel: (814) 865-6459 Fax: (814) 865-6426
E-mail: jjansen@ist.psu.edu

Amanda Spink
School of Information Sciences
University of Pittsburgh
610 IS Building, 135 N. Bellefield Avenue
Pittsburgh, PA 15260
Tel: (412) 624-9454 Fax: (412) 648-7001
Email: aspink@sis.pitt.edu

ABSTRACT

This chapter reviews the concepts of Web results page and Web page viewing patterns by users of Web search engines. It presents the advantages of using traditional transaction log analysis in identifying these patterns, serving as a basis for Web usage mining. The authors also present the results of a temporal of analysis of Web page viewing illustrating that the user – information interaction is extremely short. By using real data collected from real users interacting with real Web information retrieval systems, the authors aim to highlight one aspect of the complex environment of Web information seeking.

KEYWORDS: Relevance of Information, Human/computer interaction, User Information Satisfaction

INTRODUCTION

The Web has dramatically changed the way people locate information. One can define Web mining as

> the discovery of and analysis of useful information from the World Wide Web. This describes the automatic search of information resource available on-line, i.e., Web content mining, and the discovery of user access patterns from Web services, i.e., Web usage mining. (Cooley, Mobasher, & Srivastava, 1997).

Information viewing characteristics of users are central aspects of this view of Web mining. As the Web has become a worldwide phenomenon (Cole, Suman, Schramm, Lunn, & Aquino, 2003), we need an understanding what searching trends are emerging, including both how people utilize Web search engines in the search process to locate Web documents and how searchers are visiting and viewing the documents that the search engine locates.

There is a growing body of Web research concerning how users interact with Web search engines. There are also reports on the number of result pages viewed. When a Web search engine user submits a query, the search engine returns the results in “chucks”, of usually about 10 results. We refer to these “chucks” as results pages, and the search engine presents results within these pages to the user sequentially from the top most ranked results page to the maximum number of result pages retrieved by the search engine. However, there has been little large-scale research examining the pattern of interactions between Web search engine users and the actual Web documents presented by these results pages.

In this chapter, we summarize research on the results page viewing activities of users of Web search engines. We examine general searching characteristics including the number of result pages viewed. We then examine the number of Web documents that users view, analyzing the relationship between sessions, queries, and Web pages viewed. We also explore the temporal relationships of these interactions.

We begin with a review of the literature, followed by the methodology we utilized to analyze actual Web queries submitted by users to Web search engines. We use these queries to examine trends in searching. Specifically, we examine result pages accessed, 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. Click thru data shows great promise in the area of Web mining to isolate relevant content, identify searchers' usage patterns, and evaluate Web search engine system performance (Joachims, 2002). 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 directions of future research in this area.

BACKGROUND

There has been limited research examining the results pages and little analysis of the Web page 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 (Hölscher & Strube, 2000; Jansen & Pooch, 2001; Jansen, Spink, & Saracevic, 2000; Spink, Jansen, Wolfram, & Saracevic, 2002). This research provides 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 (2001) 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 (2000) examined European searchers and report information on sessions, queries, and terms, noting that experts exhibit different searching patterns than novices. (Jansen, Spink, & Saracevic, 2000) conducted an in-depth analysis of the user interactions with the Excite search engine. Spink, Jansen, Wolfram and Saracevic (2002) 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 pages viewing.

Focusing specifically on result page access patterns, Jansen, Spink and Pedersen (Under Review) present temporal results of results page viewing activities on the Alta Vista Web search engine. The queries examined for this study were submitted to Alta Vista on 8 September 2002 and span a 24-hour period. The queries were recorded in four transaction logs (*general*, *audio*, *image*, and *video*) and represent a portion of the searches executed on the Web search engine on this particular date. The original general transaction log contains approximately 3,000,000 records. Each record contains three fields: (1) *Time of Day*: measured in hours, minutes, and seconds from midnight of each day as recorded by the Alta Vista server; (2) *User Identification*:

an anonymous user code assigned by the Alta Vista server; and (3) *Query Terms*: terms exactly as entered by the given user.

For the temporal analysis, the researchers compared the results from this analysis to results from a 1998 study of Alta Vista searchers. Silverstein, Henzinger, Marais and Moricz (1998) used a transaction log with several fields including: (1) *Time Stamp*: measured in milliseconds from 1 January 1970, (2) *Cookie*: which is the cookie filename used to identify a user computer, and (3) *Query*: terms exactly as entered by the given use. The queries in the 1998 study were submitted to Alta Vista during the period 2 August through 13 September 1998. The total transaction log contained 993,208,159 requests, just under a billion records.

Table 1: Overview results for data analysis of 1998 and 2002

	Alta Vista 1998		Alta Vista 2002	
Sessions	285,474,117		369,350	
Queries	993,208,159		1,073,388	
Terms				
Unique			369,350	9.5%
Total			1,073,388	100%
Results Pages Viewed				
1 page	718,615,763	85.2%	781,483	72.8%
2 pages	63,258,430	7.5%	139,088	13.0%
3+ pages	13,674,409	7.3%	150,904	14.1%

Table 1 shows an increase in the percentage of users viewing more than the first results page, which when combined with other increased interactions may indicate an increased persistence in locating relevant results.

Table 2 presents a more detailed view of the result pages viewing of Alta Vista Web users.

Table 2: Results Pages Viewed for 2002

Number of Results Pages Viewed	Occurrences	%	Occurrences	%
1	846,213,351	85.2%	781,483	72.8%
2	74,490,612	7.5%	139,088	13.0%
3	29,796,245	3.0%	60,334	5.6%
4	42,707,951 *	4.3%	27,196	2.5%
5			16,898	1.6%
6			11,646	1.1%
7			6,678	0.6%
8			4,939	0.5%
9			3,683	0.3%
>=10			21,398	2.0%

Note: For the 1998 figure, calculated based on distinct queries only, 153,645,993.

* Number and percentages are for session of 4 and more.

Jansen, Spink, and Pedersen (Forthcoming) also presents result page viewing of searchers using a multimedia ontology.

Table 3 presents these results.

Table 3: Result Page Viewing of General, Audio, Image, and Video Searching in 2002.

	General	Audio	Image	Video
Sessions	369,350	3,181	26,720	5,789
Queries	1,073,388	7,513	127,614	24,265
Terms				
Unique	297,528 (9.5%)	6,199 (33.4%)	71,873 (14.1%)	8,914 (19.1%)
Total	3,132,106 (100%)	18,544 (100%)	510,807 (100%)	46,708 (100%)
Results Pages Viewed				
1 page	781,483 (72.8%)	5,551 (73.9%)	80,455 (63.0%)	13,357 (55.0%)
2 pages	139,088 (13.0%)	1,070 (14.2%)	14,498 (11.1%)	3,905 (16.1%)
3+ pages	150,904 (14.1%)	892 (11.9%)	32,661 (25.65)	1,949 (28.9%)

When comparing among the four types of searching (general, audio, image, and video) in Table 3, we see that video searchers viewed more results pages than other searchers, with only 55% of video searchers viewing only one results page.

There has been less research focusing on European users of Web search engines, relative to users of U.S. search engines. Three studies have examined this area of Web searching (Cacheda & Viña, 2001a; Hölscher & Strube, 2000; Spink, Ozmutlu, Ozmutlu, & Jansen, 2002). Hölscher and Strube (2000) examined European searchers on the Fireball search engine, a predominantly German search engine, and reported on the use of Boolean and other query modifiers. The researchers note that experts exhibit different searching patterns than novice users. Cacheda and Viña (2001a; 2001b) reported statistics from a Spanish Web directory service, BIWE.

Table 4 provides the key results for the FIREWALL and BIWE studies.

Table 4: Results Pages Comparison of FIREWALL and BIWE study.

	Fireball Study		BWIE Study	
Sessions	Not Reported		71,810 *	
Queries	451,551		105,786	
Terms				
Unique	Not Reported		18,966	16%
Total	Not Reported		116,953	
Results Pages Viewed				
1 page	9261367	60%	48,831	68%
2 pages	6545887	40%	9,335	13%
3+ pages			13,644	19%

* Data reported using 71,810 initial queries.

Table 4 shows that users of European search engines view even fewer results pages than users of U.S. search engines.

Most of the existing Web searching literature focuses on human searching and page viewing. However, much searching is now done using automated processes such as agents and meta-searching tools. Jansen, Spink and Pedersen (2003a; 2003b) conducted two studies of agent searching on Web search engines. In the first study (Jansen, Spink, & Pederson, 2003a), the queries examined were submitted to Alta Vista on 8 September 2002. The researcher culled the agent submissions from the original transaction logs. The researchers examined sessions with over 10,000 queries.

Table 5 displays the results of this analysis.

Table 5: Agents Searching Characteristics for Top Agents

	Number	Percentage
Sessions	22	
Queries	219,718	
Terms *Unique*	277,902	60%
Total	459,537	
Results Pages Viewed Per Query		
1 page	18,8747	86%
2 pages	17,155	8%
3+ pages	13,816	6%

Agents exhibit the same characteristic as human Web searchers, a very low tolerance for wading through a lot of results. In fact, Web agents appear to have an even lower tolerance for viewing a large number of results. For 86% of the agent's, only the first set of results were viewed, which is 30% higher than human Web searchers.

The researchers conducted a follow-on study (Jansen, Spink, & Pederson, 2003b) with a larger set of agent submissions. The results are displayed in Table 6.

Table 6: Aggregate results for general search trends

	Agent Searching Data During Interactions with Alta Vista	
Sessions	2,717	
Queries	896,387	
Terms		
Unique	570,214	17.7%
Total	3,224,840	
Results Pages Viewed		
1 page	760,071	85%
2 pages	67,755	8%

Table 6: Aggregate results for general search trends

	Agent Searching Data During Interactions with Alta Vista	
3+ pages	68,561	8%

Table 6 shows for over two thousand and five hundred agents, most still viewed only one results page.

Spink, Jansen, Wolfram, and Saracevic (2002), as part of a body of research studying Web searcher and Web search engine interaction analyzed three data sets culled from more than one million queries submitted by more than 200,000 users of the Excite Web search engine, collected in September 1997, December 1999, and May 2001. This longitudinal benchmark study shows that public Web searching is evolving in certain directions, specifically in the area of result pages viewed.

Table 7 shows the results pages aspect of this study.

Table 7: Comparative statistics for Excite Web query data sets.

Variables	1997	1999	2001
Result pages viewed per query			
1 page	28.6%	42.7%	50.5%
2 pages	19.5%	21.2%	20.3%
3+ pages	51.9%	36.1%	29.2%

Generally, from Table 7, we see that users are viewing fewer results pages in 2001 relative to 1997. Over 50% of the users by 2001 viewed no more than only one results page.

VIEWING OF RESULTS PAGES AND WEB DOCUMENTS

In general, from a summation of this literature, 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. The studies cited previously illustrate that the majority of Web searchers, approximately 80%, view no more than 10 to 20 results.

However, 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 chapter, we present research results to address these issues by examining the page viewing patterns of actual Web search engine users.

More specifically, the overall research questions driving this study (Jansen & Spink, 2003), are:

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

(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 at the time owned by FAST. From this analysis, we could determine the number of result 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, we evaluate a subset of click thru data from this transaction log to determine whether or not the Web document contained relevant information.

Data Collection

The queries examined for this study were submitted to FAST, 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). In our analysis, we generally use the procedure and terminology outlined in Jansen and Pooch (2001).

Briefly, the metrics we used and addressed by Jansen and Pooch (2001) include:

(1) Session. The *session* is the entire sequence of queries entered by a searcher. We identified a searcher as a unique *User Identification* and applied no temporal cut-off. We attempted to exclude sessions from softbots using numerical limitation. However, currently, there is no way to precisely identify all of these automated searches (Silverstein, Henzinger, Marais, & Moricz, 1999).

(2) Query. A set of queries compose a sessions. We define a *query* as a string of zero or more characters submitted to a search engine. This is a mechanical definition as opposed to an information seeking definition (Korfhage, 1997). We refer to the first query by a particular searcher as an *initial query*. A subsequent query by the same searcher that is identical to one or more of the searcher's previous queries is a *repeat query*.

(3) Term. A *term* is a string of characters separated by some delimiter such as a space or some other separator. In our analysis, we used a blank space as the separator.

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 (URL) 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.

Data Analysis

With these four 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 URL 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 submitted by the same agent to give the unique queries in order to analyze 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 (Saracevic, 1975). The reviewers received training regarding the judgment process and were given instructions for determining relevance. Inter-rater agreement across the three raters was found to be quite high (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 8 presents an overview of the analysis.

Table 8: Overview Data

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 (Montgomery & Faloutsos, 2001; Silverstein, Henzinger, Marais, & Moricz, 1999), leading us to believe that the data from this transaction log represents searches submitted by the typical population of Web users. Jansen and Pooch (2001) also noted similarities among users of a variety of Web search engines.

Number of Result Pages Viewed

From an analysis of Table 8, 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 result pages viewed parallels each other with about equal percentages of queries submitted and Result Pages viewed. This may imply some relationship between the sufficiency (Jansen, Spink, & Saracevic, 1998) of the retrieved results relative to the user's information need. For example, if the results from the first query were relevant and satisfied the information there would be no need for the user to submit additional.

Table 9 presents a more in-depth analysis of the number of pages viewed per query submitted.

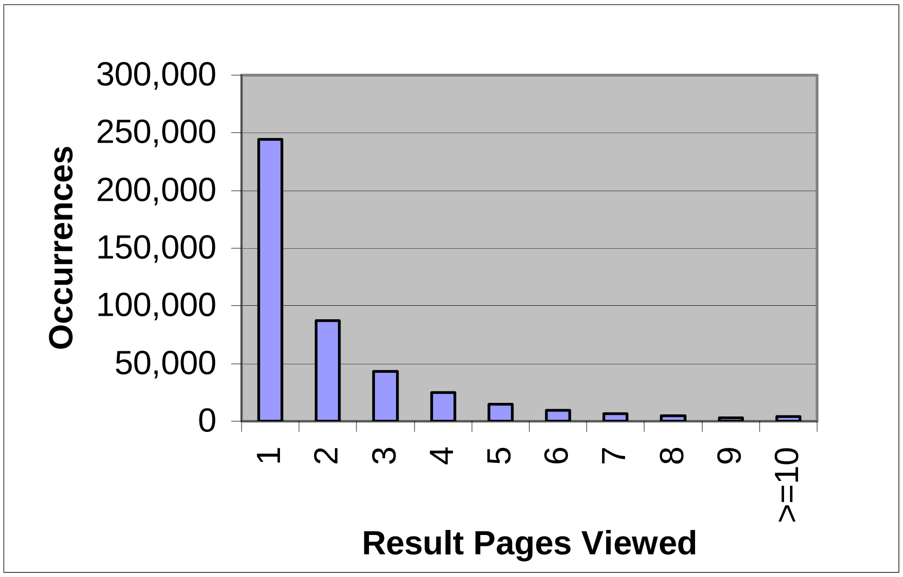
Table 9: Results Pages Viewed

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	4,391	2.8%

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

Figure 1 displays the trend in results page viewing using the data displayed in Table 9.

Figure 1: Viewing of Results Pages.



Web Documents Viewed By Users

Although most users viewed only the first one or two results pages, this does not tell us the actual number of Web pages they actually visited (i.e., pages viewed). They may have viewed all results presented, or they may have viewed none. To address this issue, Table 10 shows the number of pages viewed per session.

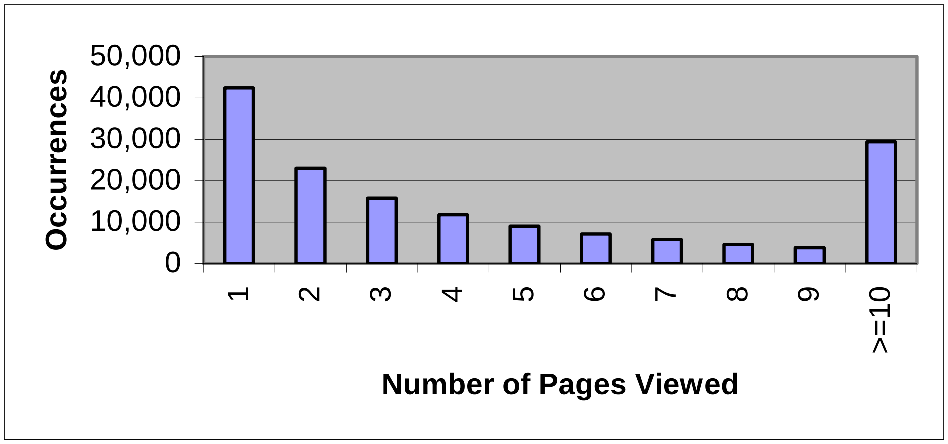
Table 10: Pages Viewed Per Session

Number of Pages 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	29,370	19.09%

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 visit more than the first results page, which usually displays 10 results. While 10 results is in line with the average, our analysis shows that over 66% of searchers examine fewer than 5 pages in a typical session and almost 30% view only one document in a given session.

Figure 2 displays the trend in page viewing using the data from Table 10.

Figure 2: Viewing Web Pages.



Web Documents Viewed By Query

This low number of viewed pages holds when we move from the session level of analysis to the query level.

Table 11 presents the number of pages viewed per query.

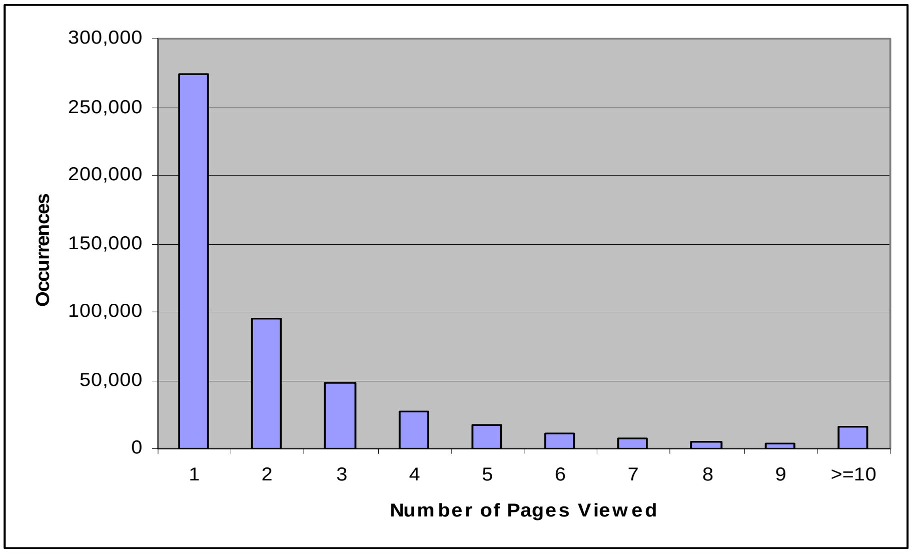
Table 11: Pages Viewed Per Query

Number of Pages 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	15,473	3.1%

The mean number of pages 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 Web page per query, just fewer than 55%.

Figure 3 displays the trend in page viewing by query using the data from Table 11.

Figure 3: Viewing of Web Pages by Query.



Session Duration

Table 12 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 12: 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 server 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 12. 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 (He, Göker, & Harper, 2002). 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 13.

Table 13: 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 views skewed our mean. Over 75% of the users viewed the retrieved Web document for less than 15 minutes. Nearly 40% of the users viewed the retrieved Web document for less than 3 minutes. Perhaps more surprisingly, 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 (CyberAtlas, 2002).

RELEVANCE OF PAGES VIEWED

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 14.

Table 14: 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	

Table 14: Relevance Results for Pages Viewed

Relevance Score	Number of Documents	Percentage

Three independent raters viewed 530 URLs and evaluated 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 (Spink, Bateman, & Jansen, 1999). Assuming a relevance score of 2 or higher indicates a relevant document, Web users would generally need to view about two documents to find a relevant one.

SUMMARY

There are some clear patterns concerning the number of result pages viewed by FAST users. Approximately 54% of the users view only one results page. This finding 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 that a person views 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 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 significant implications for Web search engines and Web page designers. 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 popular search engine and one of the largest search engines on the Web in terms of both document collections. 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 (Jansen & Pooch, 2001) 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. Third, 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. Finally, it is possible that the click thru data we used in the relevance evaluation is not representative of the total transaction log.

FUTURE TRENDS

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 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. More importantly, a precision of approximately fifty percent would indicate there is room for continued improvement in Web search engine design.

ACKNOWLEDGEMENT

We thank Excite, AlltheWeb.com and AltaVista providing the Web query data sets without which we could not have conducted this research.

REFERENCES

Cacheda, F., & Viña, Á. (2001a). Experiences retrieving information in the world wide web. In *Proceedings of the 6th IEEE Symposium on Computers and Communications*, pp. 72-79. Hammamet, Tunisia. July.

Cacheda, F., & Viña, Á. (2001b). Understanding how people use search engines: A statistical analysis for e-business. In *Proceedings of the e-Business and e-Work Conference and Exhibition 2001*, pp. 319-325. Venice, Italy., October.

Cole, J. I., Suman, M., Schramm, P., Lunn, R., & Aquino, J. S. (2003, February 2003). *The ucla internet report surveying the digital future year three* [website]. UCLA Center for Communication Policy. Retrieved 1 February, 2003, from the World Wide Web: http://www.ccp.ucla.edu/pdf/UCLA-Internet-Report-Year-Three.pdf.

Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web mining: Information and pattern discovery on the world wide web. In *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, pp. 558-567. Newport Beach, CA. November.

CyberAtlas. (2002). *November 2002 internet usage stats* [website]. Nielsen//NetRatings Inc. Retrieved 1 January, 2002, from the World Wide Web: http://cyberatlas.internet.com/big_picture/traffic_patterns/article/0,,5931_1560881,00.html.

He, D., Göker, A., & Harper, D. J. (2002). Combining evidence for automatic web session identification. *Information Processing & Management, 38*(5), 727 - 742.

Hölscher, C., & Strube, G. (2000). Web search behavior of internet experts and newbies. *International Journal of Computer and Telecommunications Networking, 33*(1-6), 337-346.

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. (2003). An analysis of web information seeking and use: Documents retrieved versus documents viewed. In *Proceedings of The 4th International Conference on Internet Computing*, pp. 65 - 69. Las Vegas, Nevada. 23 - 26 June.

Jansen, B. J., Spink, A., & Pederson, J. (2003a). Monsters at the gates: When softbots visit web search engines. In *Proceedings of the 4th International Conference on Internet Computing*, pp. 620-626. Las Vegas, Nevada. 23 - 26 June 2003.

Jansen, B. J., Spink, A., & Pederson, J. (2003b). Web searching agents: What are they doing out there? In *Proceedings of the 2003 IEEE International Conference on Systems, Man & Cybernetics*, pp. 10 - 16. Washington, D.C., USA., 5-8 October.

Jansen, B. J., Spink, A., & Pederson, J. (Forthcoming). A view from above: A temporal analysis of web searching on alta vista. *Information Processing & Management*.

Jansen, B. J., Spink, A., & Pederson, J. (Under Review). Trend analysis of alta vista web searching. *Journal of the American Society for Information Science and Technology*.

Jansen, B. J., Spink, A., & Saracevic, T. (1998). Searchers, the subjects they search, and sufficiency: A study of a large sample of excite searches. In *Proceedings of the 1998 World Conference on the WWW and Internet*, pp. Orlando, FL.

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.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of 8th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp. 133 - 142. Edmonton, Alberta, Canada.

Korfhage, R. (1997). *Information storage and retrieval*. New York, NY: Wiley.

Montgomery, A., & Faloutsos, C. (2001). Identifying web browsing trends and patterns. *IEEE Computer, 34*(7), 94-95.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society of Information Science, 26*(6), 321-343.

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., Ozmutlu, S., Ozmutlu, H. C., & Jansen, B. J. (2002). U.S. Versus european web searching trends. *SIGIR Forum, 32*(1), 30 - 37.