Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 41 (2005) 361–381
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

An analysis of Web searching by European AlltheWeb.com users

Bernard J. Jansen [a,*], Amanda Spink [b,1]

[a] School of Information Sciences and Technology, The Pennsylvania State University, 2P Thomas Building, University Park, PA 16802, USA
[b] School of Information Sciences, University of Pittsburgh, 510 IS Building, 135 N. Bellefield Avenue, Pittsburgh, PA 15260, USA

Received 15 April 2003; accepted 15 July 2003
Available online 30 August 2003

Abstract

The Web has become a worldwide source of information and a mainstream business tool. It is changing the way people conduct the daily business of their lives. As these changes are occurring, we need to understand what Web searching trends are emerging within the various global regions. What are the regional differences and trends in Web searching, if any? What is the effectiveness of Web search engines as providers of information? As part of a body of research studying these questions, we have analyzed two data sets collected from queries by mainly European users submitted to AlltheWeb.com on 6 February 2001 and 28 May 2002. AlltheWeb.com is a major and highly rated European search engine. Each data set contains approximately a million queries submitted by over 200,000 users and spans a 24-h period. This longitudinal benchmark study shows that European Web searching is evolving in certain directions. There was some decline in query length, with extremely simple queries. European search topics are broadening, with a notable percentage decline in sexual and pornographic searching. The majority of Web searchers view fewer than five Web documents, spending only seconds on a Web document. Approximately 50% of the Web documents viewed by these European users were topically relevant. We discuss the implications for Web information systems and information content providers.

Keywords: Web searching; Session duration; Query language; Search engine evaluation

* Corresponding author. Tel.: +1-814-865-6459; fax: +1-814-865-6426.
E-mail addresses: jjansen@ist.psu.edu (B.J. Jansen), aspink@mail.sis.pitt.edu (A. Spink).
[1] Tel.: +1-412-624-5230; fax: +1-412-624-5231.

0306-4573/$ - see front matter
doi:10.1016/S0306-4573(03)00067-0

1. Introduction

The Web is changing the way many people locate information. As the Web is becoming a worldwide phenomenon, we need to understand what searching trends are emerging. These trends include how searchers utilize Web search engines in the search process and the viewing of Web documents. There is a growing body of Web research concerning how users interact with Web search engines (Spink, Jansen, Wolfram, & Saracevic, 2002). However, the majority of research in this area has focused on users of United States Web search engines. There is a need to understand what searching trends are emerging within different global regions. To our knowledge, there has been limited large-scale research examining the interactions of users with European Web search engines. Examining the Web searching behavior of different users from different world regions is an important area of research with potential to impact our understanding of global Web search and the design of Web search engines.

In this paper, we examine the interactions of the users of a major and predominantly European search engine. We report general searching characteristics and trends, including session duration, query length, languages, and result pages viewed. We also examine the number of Web documents viewed, and analyze the relationship between sessions, queries, and pages viewed. Finally, we evaluate the success of these searches by analyzing the topical relevance of documents retrieved and viewed.

We begin with a review of the literature, followed by the research design utilized to obtain and analyze this Web search engine data. We use these Web queries to isolate trends in searching and page viewing, also known as click through or page view data (i.e., the Web page/s a user visits when following a hyperlink from a search engine results page). This analysis includes the temporal aspects of Web page viewing. We discuss the implications of these results for Web search engine users and designers, and Web sites targeting the European market. We conclude with directions for future research.

2. Related studies

2.1. Web searching

There is a growing body of research examining the search patterns of users of predominantly US search engines (Jansen & Pooch, 2001; Jansen, Spink, & Saracevic, 2000; Silverstein, Henzinger, Marais, & Moricz, 1999; Spink et al., 2002). 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. Jansen et al. (2000) conducted an in-depth analysis of the user interactions with the Excite search engine. Silverstein et al. (1999) conducted a large study with a sample of more than a billion queries from the Alta Vista search engine. Spink et al. (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.

Overall, we see that Web searching sessions are very short as measured by 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. [2] From the studies cited, the majority of Web searchers, approximately 80%, view no more than 10–20 Web documents. 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. There has also been a focus on primarily US search engines, with much less study of European and other Web search systems. From previous research, indications are that over 80% of users of US based search engines are from the US (Spink, Bateman, & Jansen, 1999). This stream of research provides useful information and a methodology for considering the Web search process when evaluating search engine usage in other regions.

2.2. European Web studies

Limited research has focused on users of European Web search engines. Three studies have examined this area of Web searching (Cacheda & Viña, 2001a; Hölscher & Strube, 2000; Spink, Ozmutlu, et al., 2002). Hölscher and Strube (2000) examined European searchers on the Fireball [3] 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. [4] Table 1 provides the key results for the Fireball and BIWE studies.

The researchers report on number of page results viewed, queries, operator usage, and terms.

Spink, Ozmutlu, Ozmutlu, and Jansen (2002) compared Excite (American) and AlltheWeb.com (European) search engine users. They found that AlltheWeb.com users tended to create longer sessions and search more for information on people and places, rather than the Excite user focus on e-commerce. A summary of their results is displayed in Table 2.

In general however, there has been limited research on European based Web searchers. European users may interact differently with Web search systems relative to their US counterparts. Their searching topics may differ. They may have different preferences in viewing results. In this study, we seek to address these issues by examining the searching patterns of actual Web searchers using a predominately European Web search engine.

3. Research questions

The research questions driving this study are:

(1) What are the trends in Web searching characteristics by European users of the AlltheWeb.com search engines?
(2) How many Web documents do AlltheWeb.com European Web search engine users' view, and how long do they spend viewing these documents?
(3) How topically relevant are the Web documents they are viewing?

[2] 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 the search engine presents.

[3] http://www.fireball.de/

[4] http://www.biwe.com/index.html/

Table 1
Comparison of Fireball and BIWE study results

	Fireball study[a]		BIWE study[b]	
Sessions	Not reported		71,810	
Queries	451,551		105,786[c]	
Terms				
Unique	Not reported		18,966	16%
Total	Not reported		116,953	
Mean terms per query	1.66		Not reported	
Terms per query				
1 term	8,873,001	55%	Not reported	
2 terms	5,005,653	31%	Not reported	
3+ terms	2,374,248	14%	Not reported	
Session size	Not reported		Not reported	
Results pages viewed				
1 page	9,261,367	60%	48,831	68%
2 pages	6,545,887[d]	40%	9335	13%
3+ pages			13,644	19%
Boolean queries	414,461	3%	33,302	5%
Terms not repeated in data set	Not reported		9356	8%

[a] Hölscher and Strube (2000).
[b] Cacheda and Viña (2001a).
[c] Data reported using 71,810 initial queries.
[d] Statistics reported for first page and all other pages.

These issues are important for the examination of European Web searching, as the Web becomes a more global tool for information searching.

4. Research design

4.1. Data collection

We obtained, and quantitatively analyzed, actual queries submitted to AlltheWeb.com, [5] a major European Web search engine at the time of the study owned by FAST. Since the study, an outside company has purchased the FAST corporation (Kane, 2003). According to AlltheWeb.com personnel, most European users of AlltheWeb.com are from Norway and Germany. All queries were submitted to the European Web site for the AlltheWeb.com search engine. The queries examined for this study were submitted to AlltheWeb.com on 6 February 2001 and 28

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

Table 2
Comparative results from Excite and AlltheWeb.com study[a]

Variables	Excite	AlltheWeb.com
Sessions	262,025	153,848
Queries	1,025,910	451,551
Terms	1,538,120	1,350,619
Mean terms per query	2.6	2.4
Terms per query		
1 term	26.9%	25%
2 terms	30.5%	36%
3+ terms	42.6%	39%
Mean queries per session	2.3	2.9
Session size		
1 query	55.4%	53%
2 queries	19.3%	18.9%
3+ queries	25.3%	29%
Mean pages viewed per query	1.7	2.2
% of use of 100 most frequently occurring query terms	22%	14%

[a] Spink, Ozmutlu, et al. (2002).

May 2002, each spanning a 24-h period. The queries were recorded in transaction logs and represent a portion of the searches executed on the Web search engine on these particular dates. The transaction logs hold 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 AlltheWeb.com server. The AlltheWeb.com server software derives this code using the Internet Protocol (IP) address of the searcher's machine. The code is unique and persistent.
(3) *Query terms*: terms exactly as entered by the given user.
Additionally, the 2001 transaction log contained:
(4) *Language*: the user preferred:
The 2002 transaction log contained:
(5) *Page viewed*: the uniform resource locator (URL) that the searcher visited after entering the query.

The transaction log contained searches from individuals, common user terminals, automated processes, and agents. We were interested in only those queries submitted by individuals for this study. From the transaction log, we therefore culled a subset of queries that we deemed were likely submitted by an individual. To do this, we separated all sessions with less than 101 queries into a separate transaction log, which we used for this research. We chose 101 queries because it is

almost 50 times greater than the reported mean search session (Jansen et al., 2000) for human Web searchers.

Given that there is no way to accurately identify individual from non-individual searchers, most researchers relying on transaction logs for data collection must either ignore it (Cacheda & Viña, 2001a) or assume some temporal or interaction cut-off (Montgomery & Faloutsos, 2001; Silverstein et al., 1999) to the session. Using a cut-off of 100 queries, we were satisfied that we had retrieved a subset of the transaction log that contained queries submitted primarily by human searchers in a non-common user terminal, but also broad enough not to introduce bias by too low of a cut-off threshold.

4.2. Data analysis

Using the time stamp field and user identification code, we located the initial query and recreated the chronological series of actions in a session. A session is the entire series of queries submitted by a user during one interaction with the Web search engine. A query is the entire string of terms submitted by a searcher in a given instance. A term is any series of characters separated by white space. A results page is the chuck of results presented by the search engine, usually Web sites or Web pages. The language is that selected by the user, with the default on the search engine being *ANY*. The Web page viewed is the Web document located at the URL recorded and presented by the Web search engine in the results page.

When a searcher submits a query, views a document from the results listing, and returns to the search engine, the AlltheWeb.com 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 user to give us the unique queries in order to analysis sessions, queries, languages and terms, and pages of results viewed. We could then located a user's initial query and recreated the chronological series of actions by each user in a session. As outlined in Jansen and Pooch (2001), an initial query is the first query submitted by a particular user in a session.

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. The *page viewed* field permitted us to addresses what Web document the user 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 the Web document visited from each of these 530 queries for topical relevance, assigning a binary relevance judgment of 1 (for relevant) or 0 (for not relevant) based on the rater's interpretation of the query.

Topical 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 topical relevance judgment process and were given instructions for determining topical relevance. We calculated agreement across the three raters using r_{wg}, and we found it to be quite high ($r_{wg} = 0.95$). From these topical relevance rankings, we were able to calculate

relative precision (i.e., the ratio of the number of relevant documents retrieved to the number of documents retrieved at a certain point in the results listing).

From this analysis, we could determine the trends in Web searching over an approximately one-year period. In order to facility comparison with other studies, we generally use the procedure and terminology outlined in Jansen and Pooch (2001).

5. Results

In the following sections, we report the results of our analysis.

5.1. Aggregate results

We present the aggregate results for the analysis reported in Table 3.

Overall, we see a trend toward greater simplicity and increased variability of terms. The percentage of 1-term queries increased from 25% in 2001 to 33% in 2002. The number of users modifying queries decreased approximately 6%, from 47% of all users in 2001 to 41% in 2002. Concerning overall session length, the percentage of shorter sessions trended higher, with 53% of users with a one-query session in 2001 to 59% in 2002. The percentage of queries containing Boolean operators remained at 1%, which is low, even by Web standards. There was an increase in the percentage of users viewing more than the first results page, although this may be a result of the more naïve searching rather than an increased persistence in locating relevant results.

Based on the term analysis, there was a broadening of topics. The number of unique terms increased (from 12% to 15%) and the percentage of terms not repeated in the data set also increased, from 7% in 2001 to 10% in 2002. There was a corresponding decrease in the percentage of usage represented by the top 100 most frequently occurring terms, from 15% to 14%. Taken together, these may indicate a broadening of search interests of AlltheWeb.com users.

In the following sections, we examine the results of our analysis in more detail at three levels of granularity, the session, query, and term levels of analysis.

5.2. Sessions

5.2.1. Session length

We report session length analysis in Table 4.

An increase in single query sessions resulted in shorter sessions for all except a small (about 4%) group of very persistent users. This trend parallels what is reported from analyses of US search engines, namely a move toward great simplicity in searching (Spink et al., 2002), but differs from the longer sessions lengths found by Spink, Ozmutlu, et al. (2002).

5.2.2. 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) for the 2001 data set. Unfortunately, the login time were not properly recorded for the 2002 data set.

Table 3
Comparative results statistics for AlltheWeb.com data sets

	AlltheWeb.com 2001		AlltheWeb.com 2002	
Sessions	153,297		345,093	
Queries	451,551		957,303	
Terms				
Unique	180,998	13%	340,711	15%
Total	1,350,619		2,225,141	
Mean terms per query	2.4		2.3	
Terms per query				
1 term	113,447	25%	316,514	33%
2 terms	161,541	36%	312,498	33%
3+ terms	176,563	39%	328,291	34%
Mean queries per user	3.0		2.8	
Users modifying queries	72,261	47%	142,649	41%
Session size				
1 query	81,036	53%	202,444	59%
2 queries	28,117	18%	55,664	16%
3+ queries	44,144	29%	86,985	25%
Results pages viewed				
1 page	373,559	83%	730,363	76%
2 pages	42,957	10%	125,420	13%
3+ pages	30,839	7%	101,520	11%
Boolean queries	6745	1%	9355	1%
Terms not repeated in data set	100,649	7%	212,040	10%
Use of 100 most frequently occurring terms	196,390	15%	303,176	14%

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. 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 engine 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 h, 21 min and 55 s, with a standard deviation of 4 h, 45 min, and 36 s. However, we see that the longer session durations skewed our result for the mean, masking significant details. Fully 52% of the sessions were less than 15 min. This is inline with earlier reported research on Web session length (He, Göker, & Harper, 2002). Perhaps even more surprisingly, over 25% of the sessions were less than 5 min.

Table 4
Occurrences and percentages of session length for AlltheWeb.com 2001 and 2002

Session length	2001		2002	
	Occurrences	%	Occurrences	%
1	81,036	52.9	202,444	58.7
2	28,117	18.3	55,664	16.1
3	14,445	9.4	27,307	7.9
4	8335	5.4	17,440	5.1
5	5100	3.3	10,046	2.9
6	3534	2.3	7059	2.0
7	2431	1.6	4461	1.3
8	1833	1.2	3476	1.0
9	1290	0.8	2532	0.7
⩾10	7176	4.7	14,664	4.2
Average session length	3.0		2.8	

Table 5
Occurrences and percentage of AlltheWeb.com session duration

Session duration	2001	
	Occurrences	%
<5 min	55,966	26.2
5–10 min	13,275	6.2
10–15 min	41,987	19.7
15–30 min	19,314	9.1
30–60 min	30,955	14.5
1–2 h	8691	4.1
2–3 h	21,901	10.3
3–4 h	2635	1.2
>4 h	18,605	8.7
Mean	2 h and 22 min	
Standard deviation	4 h and 37 min	

5.3. Language preferences

We analyzed the 2002 data to determined language preferences by user, with results reported in Table 6.

These Web users appear to not be concerned with specifying the language of the Web documents they retrieve, with the vast majority of searchers accepting the AlltheWeb.com default of *ANY*. The use of the particular query terms themselves itself may provide the needed selectivity for most Web users (i.e., the use of a query containing terms in Russian, for example, will retrieve primarily Web documents written in Russian). It has been reported that the majority of AlltheWeb.com customers are German (Spink, Ozmutlu, et al., 2002); however, the top non-English

Table 6
AlltheWeb.com terms, queries and sessions analysis by language

Language	Queries	Mean terms per query	Sessions	Mean queries per session	Total terms
Entire data set	957,303	2.3	345,093	2.8	2,225,141
Default[a]	874,168	2.3	313,987	2.8	2,025,072
French	53,047	2.3	33,959	1.6	119,511
Spanish	13,293	2.6	8455	1.6	34,352
German	8650	1.9	5765	1.5	16,721
Italian	4839	2.3	3386	1.4	11,310
Russian	1337	9.5	1124	1.2	12,671
English	526	5.0	194	2.7	2620
Japanese	499	3.1	292	1.7	1523
Portuguese	443	2.3	338	1.3	1026
Polish	161	1.7	5	32.2	276
Afrikaans	70	1.5	45	1.6	106
Dutch	67	1.7	8	10.2	116
Swedish	40	1.9	6	6.7	77
Danish	24	1.2	6	4.0	28
Turkish	19	1.7	12	1.6	33
Catalan	15	1.7	14	1.1	26
Arabic	7	9.1	7	1.0	64
Norwegian	6	1.5	6	1.0	9
Portugal	4	4.3	1	4.0	17
Hebrew	3	3.0	1	3.0	9
Korean	2	2.5	2	1.0	5
Albanian	2	1.5	1	2.0	3
Ukrainian	1	2.0	1	1.0	2
Greek	1	2.0	1	1.0	2
Latin*	1	0.0	1	1.0	–
Basque*	1	0.0	1	1.0	–
Other[b]	77	–	–	–	–

* The queries were blank.
[a] The default language selection is ANY.
[b] Non-language options such as DOMAIN and ALL.

language preference was French, followed by Spanish, with German at a distance third. Italian and Russian also had fairly high rates of occurrences.

5.4. Queries

5.4.1. Query length

We report query length analysis in Table 7.

Query lengths of 1–3 terms inclusive account for 83% of all queries in 2001 and 84% of all queries in 2002. The percentage of queries with 1 term has increased by 8%. After 3 terms, there is a sharp decline in the frequency of occurrences, dropping to almost minimal occurrences after 4 terms per query. Similar to trends in sessions, this trend with European Web searchers parallels that reported from analysis of US search engines (Spink et al., 2002).

Table 7
AlltheWeb.com query lengths

Query length	2001		2002	
	Occurrences	%	Occurrences	%
0	3682	0.8	2905	0.3
1	113,447	25.1	316,514	33.1
2	161,541	35.8	312,498	32.6
3	101,276	22.4	181,270	18.9
4	43,473	9.6	78,162	8.2
5	16,498	3.7	32,233	3.4
6	6493	1.4	13,287	1.4
7	2619	0.6	6286	0.7
8	1137	0.3	8225	0.9
9	581	0.1	1812	0.2
⩾10	804	0.2	4111	4.0

5.5. Page results viewed

Table 8 presents a more in-depth analysis of the number of pages viewed per query submitted.

There was 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. The percentage of European searchers viewing only one page of results is significantly higher (5–25%) than reported in previous research (Jansen et al., 2000; Silverstein et al., 1999). European users appear to have a low tolerance for wading through large numbers of results.

5.6. Terms

We present a term analysis in Table 9.

Table 8
AlltheWeb.com results pages viewed

Results pages viewed	2001		2002	
	Occurrences	%	Occurrences	%
1	373,559	83.5	730,363	76.3
2	42,957	9.6	125,420	13.1
3	13,602	3.0	37,270	3.9
4	6027	1.3	21,375	2.2
5	3481	0.8	13,510	1.4
6	1955	0.4	8488	0.9
7	1339	0.3	5464	0.6
8	912	0.2	3512	0.4
9	639	0.1	2277	0.2
10	542	0.1	1615	0.2
>10	2342	0.5	1170	0.8

Table 9
AlltheWeb.com top occurring terms and frequencies

Term	Frequency	
	2001	2002
Free	8583	9691
Sex	4513	6784
Download	5566	5997
Software	2031	3838
UK	3534	3549
Windows	2216	3252
New	2240	2994
Hotel	2433	2991
MP3	2303	2909
Video	1574	2793
Crack	1660	2731
Nude	2439	2689
Pictures	3539	2552
Web	1336	2513
Home	939	2235
World	1304	2192
Online	1438	2189
Internet	1341	2133
CD	1420	2113
Music	1612	2041
Girls	1449	2005
Canada	905	1928
Photo	1208	1876
How	1415	1871
Car	1025	1852
Pics	2110	1848
XP	17	1815
Map	1574	1705
Games	1307	1639
School	1470	1615
Lyrics	1901	1503
University	1551	1193
History	1370	1072
Linux	1413	894

From both transaction logs, we extracted the top terms, removing the terms without content (*and*, *or*, *de*, *la*, *le*, etc.). We then took the top 25 terms from each year. For better identification of trends, if a term appeared in one list and not the other we added terms and the frequency of occurrence to each list for those terms. The combined list is what is presented in Table 9.

Three trends present themselves from the term level of analysis. First, all of the top terms are English language terms, despite AlltheWeb.com being primarily a European search engine. Second, technology terms dominate the top term usage list, with terms such as *Internet*, *Linux*, *software*, *Web*, *Windows*, and *XP*. These types of terms certainly seem to stand out as indicators of a major topic for AlltheWeb.com searchers. Third, another topical area for AlltheWeb.com

searchers is entertainment, with terms such as *CD*, *games*, *MP3*, *lyrics*, *music*, and *video*. These topical areas have held fairly constant over the analysis period.

5.6.1. Term co-occurrence

Although a term analysis is useful, it is sometimes difficult to determine the specific usage of a term intended by a searcher within the framework of a particular query. In these cases, a term co-occurrence is more helpful (Leydesdorff, 1989; Wolfram, 1999). We present in Table 10 a term co-occurrence for the 2001 data set in a correlation matrix fashion.

From the term co-occurrence analysis, the predominance of technical searching is even clearer. Nearly half of the top occurring term pairs are technology related (e.g., *Windows mac*, *Windows os*, and *bug fixes*). Business related pairs are a distance second.

In Table 11, we present the term co-occurrence for 2002 in a correlation matrix fashion.

In Table 11, we do not see the clustering that was displayed in Table 10 for the 2001 data set. This diversity reinforces our findings with the initial term analysis that these Web users are searching for an increasingly variety of topics and domains. Business and entertainment pairs have replaced technology as the predominant grouping.

5.7. Topical query classification

We classified a random sample of approximately 2500 English language queries each from the 2001 and 2002 data sets, into 11 non-mutually exclusive, general topic categories previously derived by Spink et al. (2002). Table 12 presents the results of this classification.

People, *places or things* category remained the top ranked category with a large percentage increase from 2001 to 2002. Percentage drops occurred in several other categories, most noticeably *computers or Internet* and *sex or pornography*. The category rankings changed somewhat. The *sex or pornography* category, for example, dropped from 4th to 6th place. This decrease in sexual searching as a percentage of overall Web searching parallels that reported in studies of US searching (Spink et al., 2002). This analysis confirms that reported by Spink, Ozmutlu, et al. (2002) who found little European commercial searching compared to the large shift to e-commerce searching in the US. It also parallels the increase in commercial content on the Web (Lawrence & Giles, 1999).

5.8. Web documents viewed per session and query

Although most searchers viewed only the first one or two pages of results, this does not tell us the number of Web documents they actually visited. They may have viewed all documents presented or they may have viewed none. To address this issue, Table 13 shows the Web documents viewed per session. Initial results for were presented in Jansen and Spink (2003).

The mean Web documents 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 (Spink et al., 2002). While 10 documents is in line with the average, our analysis shows that over 66% of searchers examine fewer than five documents in a typical session and almost 30% view only one document in a given session.

Table 10
Frequency of term co-occurrence for top 25 AlltheWeb.com terms for 2001

	Agree-ment	Alterna-tive(s)	Bug	Cell(s)	Com-muni-cations	Fix	Fuel(s)	Hat	Joint	Linux	Mac	Me	Micro-soft	Ms	Nt	Oper-ating	Os	Power	Red	Sys-tem	Tech-nol-ogy
Agree-ment	–																				
Alterna-tive(s)		–																			
Bug			–																		
Cell				–																	
Cells					–																
Commu-nications						–															
Fix			564				–														
Fuel(s)		1054		1380				–													
Hat									–												
Joint										–											
Linux											–										
Mac												–									
Me													–								
Microsoft														–							
Ms															–						
Nt																–					
Operating																	–				
Os																		–			
Power		459					690												–		
Red								328													
Suse																				–	
System																					–
Technol-ogy																					
Venture									315	342											
Windows	317										403	410	1041	642	632	420	405			408	
Wireless					321																394

Table 11
Frequency of term co-occurrence for top 25 AlltheWeb.com terms for 2002

	2000	Basic	Cup	Down-load	Engine	Estate	For	Free	His-tory	Job	Map	Pics	Pic-tures	Real	Re-sume	Sale	Search	Sex	Skills	Soft-ware	Uni-versity	Visual	Win-dows	World
2000	–																							
Basic		–																						
Cup			–																					
Download				–																				
Engine					–																			
Estate						–																		
For							–																	
Free				611				–																
History									–															
Job										–														
Map											–													
Pics								499				–		–										
Pictures													–											
Real						1162								–										
Resume										1059					–									
Sale							995									–								
Search					411												–							
Sex				619				552										–						
Skills										482					410				–					
Software								332							325					–				
University																					–			
Visual		453																				–		
Windows	583																						–	
World			714																					–

Table 12
Comparison of AlltheWeb.com general topic categories

Rank	2001 (2503 English queries)	2002 (2525 English queries)
1	22.5% People, places or things	41.5% People, places or things
2	21.8% Computers or Internet	16.3% Computers or Internet
3	12.3% Commerce, travel, employment, or economy	12.7% Commerce, travel, employment, or economy
4	10.8% Sex or pornography	9.5% Entertainment or recreation
5	9.1% Entertainment or recreation	4.9% Health or sciences
6	7.8% Health or sciences	4.5% Sex or pornography
7	4.8% Society, culture, ethnicity or religion	2.6% Government
8	4.7% Performing or fine arts	2.5% Unknown
9	2.9% Education or humanities	2.3% Education or humanities
10	2.7% Government	2.1% Society, culture, ethnicity or religion
11	0.6% Unknown or other	1.1% Performing or fine arts

Table 13
Pages viewed per AlltheWeb.com session

Documents viewed	Occurrences	%
1	42,499	27.6
2	22,997	14.9
3	15,740	10.2
4	11,763	7.6
5	9032	5.8
6	7157	4.6
7	5746	3.7
8	4563	2.9
9	3869	2.5
10	3308	2.1
>10	26,062	16.9

The low number of documents viewed also holds when we move from the session level of analysis to the query level. Table 14 presents the Web documents viewed per query.

The mean documents viewed per query were 2.5, with a standard deviation of 3.9. AlltheWeb.com users viewed five or less Web documents per query over 90% of time. The largest percentage of users by far, just fewer than 55%, viewed only one document per query.

Prior research on Web searching has not reported the duration of viewing of Web documents by Web search engine users. We present this information for this data sample in Table 15.

The mean time spent viewing a particular Web document was 16 min and 2 s, with a standard deviation of 43 min and 1 s. However, some lengthy document views skewed our mean. Over 75% of the users viewed the retrieved Web documents for less than 15 min. More surprisingly, perhaps, nearly 40% of the users viewed the retrieved Web document for less than 3 min. Fewer than 14% of the users viewed the Web document for less than 30 s. These results for Web document viewing time are substantially less than has been previously reported using survey data (Cyber Atlas, 2002). These results suggest the need for Web site designers to place more emphasis on the presentation of Web documents given the short assessment time by searchers. Again, the time be-

Table 14
Results viewed per AlltheWeb.com query

Results viewed	Occurrences	%
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	7653	1.5
8	5231	1.0
9	3802	0.8
10	2975	0.6
>10	12,498	2.5

Table 15
Duration of page views by AlltheWeb.com users

Page view duration	Occurrences	%
<30 s	46,303	13.9
30–60 s	16,754	5.0
1–2 min	48,059	14.5
2–3 min	16,237	4.9
3–4 min	47,254	14.2
4–5 min	15,203	4.6
5–10 min	47,254	14.2
10–15 min	14,047	4.2
15–30 min	41,215	12.4
30–60 min	9054	2.7
>60 min	30,592	9.2

tween visits from the Web document to the search engine may have not been entirely spent viewing the Web document.

5.9. Topical relevance of documents viewed

This portion of the study used a random subset of records from the 2001 transaction log that included the Web site the searcher actually visited. Three independent raters visited the sites and evaluated the Web document to determine topical relevance. Topical relevance is a relevance based on a direct topic matching between the search terms used and the terms in the retrieved document, not necessarily related to the user's information seeking stage or information need (Greisdorf & Spink, 2001). Our analysis explores the question of whether search sessions are short because the searchers are potentially finding topically relevant information. The results are reported in Table 16.

We had the three independent raters view 530 URLs and evaluate these pages for topical relevance based on their interpretation of the query submitted. Each rater assigned a topical

Table 16
Topical relevance results for pages viewed by AlltheWeb.com users

Topical relevance score	Number of documents	%
3	199	37.5
2	74	14.0
1	103	19.4
0	154	29.1
Total Web documents reviewed	530	

relevance Web document a rating of 1. A non-relevant page received a rating of 0. So, the maximum topical relevance score a Web page could receive was 3, meaning that all three reviewers rated the document as topically relevant.

Approximately 52% of the time, two or more raters evaluated a page to be topically relevant. Approximately 48% of the time, two or more raters evaluated a page to be not topically 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 are finding topically relevant information on Web search engines (Spink et al., 1999) despite the simplistic searching methods.

6. Discussion

Our study identified some interesting searching patterns by AlltheWeb.com users. Web searching by these European users trended toward greater simplicity from 2001 to 2002. Queries decreased in length and sessions were shorter. Sessions were temporally short, about 15 min on average. About 25% of the sessions were less than 5 min. Boolean usage was almost non-existent. The range of topics searched for increased, and the users employed a greater variety of terms.

These searchers are generally unconcerned with specifying the preferred language of the retrieved Web documents, although within some linguistically groups this did occur (e.g., French and Spanish). This may be due to the relatively high rate of transference of terms from these to other languages. Therefore, linguistically terms from these languages will appear in Web documents that are written primarily in other languages. Searching for pornography decreased slightly as a percentage of overall Web searching and an increase in commercial searching was not apparent. Spink, Ozmutlu, et al. (2002) also found a similar result.

Web searchers do not appear to make lengthy judgments on the relevance of information retrieved. Approximately 75% of the users spent less than 15 min viewing the retrieved Web documents. Twenty percent of the Web users viewed a Web document for less than a minute. These results seem to indicate that Web searchers are typically not spending a great deal of time combing the documents to find 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 topically relevant to their query. However, note that this analysis is for Web documents viewed, not documents retrieved.

The results of this study have several implications. For search engine designers, there is still work to be done. Although, search engines are currently helping people find information, with a pre-

cision of about 50% for documents viewed, there is certainly room for improvement. For information content providers, the abstract that appears in the results listing seems to have significant impact on attracting or dissuading searchers from visiting the site. Also, Web documents must be well-designed, easy to load, and relevant information ease to find, given the short amount of time users spend on a particular Web site. For search engine users, it appears that about one in two documents viewed will be relevant, indicating a need for persistence in looking needed information.

6.1. Strengths

This study contributes to the Web searching literature in several important ways. First, the data comes from real users submitting real queries and viewing actual Web pages. Accordingly, it provides a realistic glimpse into European public Web searching, without the self-selection issues or altered behavior that can occur with lab studies or survey data. Second, our sample is quite large, with between 150,000 and 350,000 users per data set. Third, we obtained data from a very popular European based search engine. Finally, it provides a detailed examination of the Web document viewing patterns and viewing duration of Web users.

6.2. Limitations

As with any research, there are limitations that should be recognized. First, the query data comes from only one major European Web search engine, introducing the possibility that the queries do not represent the queries submitted by the broader European or global Web searching population. However, Jansen and Pooch (2001) suggest that characteristics of Web sessions, queries, and terms are very consistent across search engines.

Second, 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. Similarly, we do not have information about the demographic characteristics of the users who submitted queries, and there is no knowledge of the underlying cognitive motivation concerning the searcher's information.

Finally, there are limitations related to the use of transaction log analysis as a research tool. The identification of a user is dependent on the logging software of the search engine and the IP address of the searcher's computer. We used a numerical limit to define the upper boundary of a session (i.e., 100 queries), while other researchers have ignored the issue (Cacheda & Viña, 2001a) or utilized a temporal cut-off (Montgomery & Faloutsos, 2001; Silverstein et al., 1999). We believe that existing research supports a numerical rather than a temporal boundary.

In previously published research, there is a high degree of consistency at the session and query levels of analysis across multiple Web studies (Abdulla, Liu, & Fox, 1998; Cacheda & Viña, 2001a; Croft, Cook, & Wilder, 1995; Hölscher & Strube, 2000; Jansen et al., 2000; Montgomery & Faloutsos, 2001; Selberg & Etzioni, 1997; Silverstein et al., 1999; Spink et al., 2002; Wolfram, Spink, Jansen, & Saracevic, 2001). The similarities exist even with researchers studying various search engines and utilizing a variety of analytical methods, definitions, and metrics.

However, there is only one previously published research study that we could locate examining temporal lengths of Web sessions (He et al., 2002). He et al. (2002) report a session length of about

12 min. Based on survey data, Cyber Atlas (2002) reports a session length of over 32 min. In their studies, Silverstein et al. and fellow researchers (1999) use a session length of 5 min, and Montgomery and Faloutsos (2001) uses a session length of 2 h. Our research supports a session length of about 15 min. We believe more research in this area is needed to clearly define the temporal length of Web sessions.

7. Conclusion

Our results provide important insights into the current state of European Web searching and Web usage. The short sessions lengths combined with short queries of many Web searchers are puzzling issues for designers of Web information systems. This does not seem to be a successful strategy to maximize recall or precision, the standard metrics for information retrieval system performance. However, it appears that Web search engine users are finding topically relevant information with this searching strategy. Our research also highlights the need for further research comparing users of US and European Web search engines.

Acknowledgements

We thank AlltheWeb.com and especially Per Gunan Auran for providing the Web query data sets without which this research could not have been conducted.

References

Abdulla, G., Liu, B., & Fox, E. (1998). Searching the World-Wide Web: implications from studying different user behavior. In *Proceedings of the world conference of the World Wide Web, Internet, and Intranet, Orlando, FL* (pp. 1–8).

Cacheda, F., & Viña, Á. (2001a). Experiences retrieving information in the World Wide Web. In *Proceedings of the 6th IEEE symposium on computers and communications, July, Hammamet, Tunisia* (pp. 72–79).

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, October, Venice, Italy* (pp. 319–325).

Croft, W. B., Cook, R., & Wilder, D. (1995). Providing government information on the Internet: experiences with THOMAS. In *Proceedings of the digital libraries conference, Austin, TX* (pp. 19–24).

Cyber Atlas (2002). *November 2002 Internet usage stats* [Web site]. Nielsen//NetRatings Inc. Retrieved 1 January, 2003, from the World Wide Web: http://cyberatlas.internet.com/big_picture/traffic_patterns/article/0,5931_1560881,00.html.

Greisdorf, H., & Spink, A. (2001). Median measure: an approach to IT systems evaluation. *Information Processing and Management, 37*(6), 843–857.

He, D., Göker, A., & Harper, D. J. (2002). Combining evidence for automatic Web session identification. *Information Processing and 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, 23–26 June, Las Vegas, NV* (pp. 65–69).

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.

Kane, M. (2003). *Overture to buy search services* [Electronic Journal]. CNET News.com. Retrieved 1 March, 2003, from the World Wide Web: http://rss.com.com/2100-1023-985850.html?type = pt&part = rss&tag = feed&subj = news.

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the Web. *Nature, 400*, 107–109.

Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy, 18*, 209–223.

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.

Selberg, E., & Etzioni, O. (1997). The MetaCrawler architecture for resource aggregation on the Web. *IEEE Expert, 12*(1), 11–14.

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.

Wolfram, D. (1999). Term co-occurrence in Internet search engine queries: an analysis of the Excite data set. *Canadian Journal of Information and Library Science, 24*(2/3), 12–33.

Wolfram, D., Spink, A., Jansen, B. J., & Saracevic, T. (2001). Vox Populi: the public searching of the Web. *Journal of the American Society of Information Science and Technology, 52*(12), 1073–1074.