

This is the author-version of a paper published as:

Jansen, Bernard J. and Spink, Amanda (2006) How are we searching the world wide web? A comparison of nine search engine transaction logs. *Information Processing and Management* 42(1):pp. 248-263.

Copyright 2006 Elsevier

## **HOW ARE WE SEARCHING THE WORLD WIDE WEB?: A COMPARISON OF NINE SEARCH ENGINE TRANSACTION LOGS**

Bernard J. Jansen  
School of Information Sciences and Technology  
The Pennsylvania State University  
329F IST Building  
University Park, PA 16802  
Email: [bjansen@acm.org](mailto:bjansen@acm.org)

Amanda Spink  
School of Information Sciences  
University of Pittsburgh  
610 IS Building, 135 N. Bellefield Avenue  
Pittsburgh, PA 15260  
Email: [aspink@sis.pitt.edu](mailto:aspink@sis.pitt.edu)

### **Abstract**

The Web and especially major Web search engines are essential tools in the quest to locate online information for many people. This paper reports results from research that examines characteristics and changes in Web searching from nine studies of five Web search engines based in the U.S. and Europe. We compare interactions occurring between users and Web search engines from the perspectives of session length, query length, query complexity, and content viewed among the Web search engines. The results of our research shows (1) users are viewing fewer result pages, (2) searchers on U.S.-based Web search engines use more query operators than searchers on European-based search engines, (3) there are statistically significant differences in the use of Boolean operators and result pages viewed, and (4) one can not necessarily apply results from studies of one particular Web search engine to another Web search engine. The wide spread use of Web search engines, employment of simple queries, and decreased viewing of result pages may have resulted from algorithmic enhancements by Web search engine companies. We discuss the implications of the findings for the development of Web search engines and design of online content.

## 1. Introduction

The Web is now the primary source of information for many people (Cole, et al., 2003; Fox, 2002). Over 80% of Web searchers use Web search engines to locate online information or services (Nielsen Media, 1997). There is a critical need to understand how people use Web search engines. Amichai-Hamburger (2002) presents a review of the effect of the Web and the lack of awareness of the user in the design of Web systems and site content. The research reported in this article attempts to contribute to such a dialogue. Most research of Web searching provides little longitudinal, regional, or across system analysis. We need a clearer understanding of emerging Web searching trends across different global regions and between different Web search engines in order to design better searching systems.

This important research area directly impacts pay-per-click marketing, Web-site-optimization strategies, and Web and Intranet search engine design. It complements research such as that conducted by Liawa and Huangb (2003), who showed that individual experience, individual motivation, search engine quality, and user perceptions of technology acceptance are all factors affecting individual desire to use Web search engines.

In this paper, we present a comparison of nine major Web studies, four European and five U.S. based Web search engines, over a seven-year period. We provide a temporal comparison of differences in Web searching among and between U.S. and European-based Web searches as one might expect some divergence due to linguistics and interface factors (Spink et al., 2002b). We specifically investigate the interactivity between searchers and Web search engines, identifying changes in the complexity of Web search interactions. In addition, we present a longitudinal analysis of the types of information people are searching for on the Web.

We center our research analysis on the interactions between the user and the search engine. Interaction has several meanings in information searching, although the definitions generally encompass query formulation, query modification, and inspection of the list of results,

among other actions. Belkin, et. al. (1995) have extensively explored user interaction within an information session. Efthimiadis and Robertson (1989) present and categorize interaction at various stages in the information retrieval process from information seeking research. Bates (1990) presents four levels of interaction, which are move, tactic, stratagem, and strategy. Lalmas and Ruthven (1999) two groups of interaction, that which occurs across sessions and that which occurs within a session.

This within-session category is the type of interaction that we examine in this study. We consider an interaction as any specific exchange between the searcher and the system (i.e., submitting a query, clicking a hyperlink, etc.). We define a searching episode as a series of interactions within a limited duration to address one or more information needs. This duration is typically short, with Web researchers using between 5 and 120 minutes to define a session duration (He, Göker & Harper, 2002; Montgomery & Faloutsos, 2001; Silverstein, et al., 1999). The searcher may be multitasking (Spink, 2004) within a searching episode, or the episode may be an instance of the searcher engaged in successive searching (Lin, 2002; Spink, et al., 1998).

We begin with an extensive review of literature concerning the rapidly growing area of Web search engine research. We then present the data sets used in this study. We discuss the analysis, results, and implications of the results for the design of Web searching systems.

## **2. Related Studies**

There have been a few review articles on Web searching. Jansen and Pooch (2001) provide a review of Web transaction log research of Web search engines and individual Web sites through 2000. Hsieh-Yee (2001) reviews studies conducted between 1995 and 2000 on Web search behaviors. The researcher reports that many studies investigate the effects of certain factors on search behavior, including information organization and presentation, type of search task, Web experience, cognitive abilities, and affective states. Hsieh-Yee (2001) also notes that

many studies lack external validity. Bar-Ilan (2004) presents an extension and integrative overview of Web search engines and the use of Web search engines in information science research. Bar-Ilan (2004) provides a variety of perspectives including user studies, social aspects, Web structure, and search-engine evaluation. We extend these review articles in this section, setting the stage for our analysis.

Web searching studies fall into three categories: (1) those that primarily use transaction-log analysis, (2) those that incorporate users in a laboratory survey or other experimental setting, and (3) those that examine issues related to or affecting Web searching. In this paper, we focus on studies using transaction log analysis. Romano, et. al. (2003) present a methodology for general qualitative analysis of transaction log data. Wang, Berry, and Yang (2003) and Spink and Jansen (2004) also present detailed explanations of approaches to transaction log analysis.

In investigations of single Web sites, Yu and Apps (2000) use transaction log data to examine user behavior in the SuperJournal project. For 23 months (February 1997 to December 1998), the researchers recorded 102,966 logged actions, related these actions to four subject clusters, 49 journals, 838 journal issues, 15,786 articles, and three Web search engines. In another study covering the period from 1 January to 18 September, 2000, Kea, et. al. (2002) examined user behavior in Elsevier's ScienceDirect, which hosts the bibliographic information and full-text articles of more than 1,300 journals with an estimated 625,000 users. Loken, et. al. (2004) examined the transaction log data of the online self-directed studying of more than 100,000 students using a Web-based system to prepare for U.S. college admissions tests for several months of use. The researchers noted several non-optimal behaviors, including a tendency toward deferring study and a preference for short-answer verbal questions. The researchers discuss the relevance of their findings for online learning.

Wen, Nie and Zhang (2001) conducted research on a Web-based version of the Encarta encyclopedia. The researchers investigated the use of click-through data to cluster queries for

question answering. The researches explored the similarity between two queries using the common user-selected documents between them. The results indicate that a combination of both keywords and user logs is better than using either method alone. Using a Lucent proxy server, Hansen and Shriver (2001) used transaction-log analysis to cluster search sessions and to identify highly relevant Web documents for each query cluster.

Continuing the rich tradition of using transaction logs to investigate the remote use of library systems (Peters, 1993). Chen and Cooper (2001) clustered users of an online library system into groups based on patterns of states using transaction logs data. The researchers defined 47 variables, using them to classify 257,000 sessions. Then they collapsed these 47 variables into higher order groupings, identifying six distinct clusters of users. In a follow-on study, Chen and Cooper (2002) used 126,925 sessions from the same online system, modeling patterns using Markov models. The researchers found that a third-order Markov model explained five of the six clusters.

In what appears currently to be one of the longest temporal studies, Wang, Berry and Yang (2003) analyzed 541,920 user queries submitted to an academic-Website-search engine during a four-year period (May 1997 to May 2001). Conducting analysis at the query and term levels, the researchers report that 38% of all queries contained only one term and that most queries are unique. Eiron and McCurley (2003) used 448,460 distinct queries from an IBM Intranet search engine to analyze the effectiveness of anchor text.

Rather than focusing on single Web sites, other researchers have investigated information searching on Web-search engines. Ross and Wolfram (2000) analyzed queries submitted to the Excite search engine for subject content based on the co-occurrence of terms. The researchers categorized more than 1,000 of the most frequently co-occurring term pairs into one or more 30 developed subject areas. The cluster analyses resulted in several well-defined high-level clusters of broad subject areas. He, Göker and Harper (2002) examined contextual information

from Excite and Reuters transaction logs, using a version of the Dempster–Shafer theory (Voorbraak, 1991) to identify search engine sessions. The researchers determined the average Web user session duration was about 12 minutes. Ozmutlu and Cavdur (Forthcoming) investigate contextual information using an Excite transaction log. The researchers explore the reasons underlying the inconsistent performance of automatic topic identification with statistical analysis and experimental design techniques.

Xie and O'Hallaron (2002) investigated caching to reduce both server load and user-response time in distributed systems by analyzing a transaction log from the Vivisimo search engine, from 14 January to 17 February 2001. The researchers report that queries have significant locality, with query frequency following a Zipf distribution. Lempel and Moran (2003) also investigated clustering to improve caching of search engine results using more than seven million queries submitted to AltaVista. The researchers report that pre-fetching of search engine results can increase cache-hit ratios by 50% for large caches and can double the hit ratios of small caches.

Pu (2000) explored the searching behavior of users searching on two Taiwanese Web search engines, Dreamer and Global Area Information Servers (GAIS). The average length of English terms on these two Web search engines is 1.0 term for Dreamer and 1.22 terms for GAIS. Baeza-Yates and Castillo (2001) examined approximately 730,000 queries from TodoCL, a Chilean search system. They found that queries had an average length of 2.43 terms. A lengthier analysis is presented in (Baeza-Yates & Castillo, 2000). Montgomery and Faloutsos (2001) analyzed more than 20,000 Internet users who accessed the Web from July 1997 through December 1999 using data provided by Jupiter Media Metrix (<http://www.jupiterresearch.com>). The researchers report users revisited 54 percent of URLs at least once during a searching session. They also report that browsing patterns follow a power law and the patterns remained stable throughout the period of analysis.

Rieh and Xu (2001) analyzed queries from 1,451,033 users of Excite collected on 9 October 2000. The researchers examined how each user reformulated his/her Web query over a 24 hour period. Out of the 1,451,033 users' logs collected, the researcher used various criteria to select 183 sessions for manual analysis. The results show that while most query reformulation involves content changes, about 15% of the reformulation relate to format modifications.

Huang, Chien and Oyang (2003) propose an effective term-suggestion approach for interactive Web search using more than two million queries submitted to Web search engines in Taiwan. The researchers propose a transaction log approach to relevant term extraction and term suggestion using relevant terms that co-occur in similar query sessions.

Jansen and Spink (2003) determined that the typical Web session was about 15 minutes from an analysis of click through data from AlltheWeb.com. The researchers report that the Web search engine users on average view about eight Web documents, with more than 66% of searchers examining fewer than five documents in a given session. Users on average view about two to three documents per query. Over 55% of Web users view only one result per query. 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 to the user's perception of relevance.

Beitzel, et al., (2004) examine hundreds of millions of queries submitted by approximately 50 million users to America Online (AOL) over a 7 day period from 26 December 2003 through 1 January 2004. During this period, AOL used results provided by Google. The researchers report that only about 2% of the queries contain query operators. The average query length is 2.2 terms, and 81% of users view only one results page. The researchers report changes in popularity and uniqueness of topically categorized queries across hours of the day.

Park, Bae and Lee (Forthcoming) analyzed transaction logs of NAVER, a Korean Web search engine and directory service. The data was collected over a one-week period, from 5

January to 11 January 2003 and contained 22,562,531 sessions and 40,746,173 queries. Users of NAVER implement queries with few query terms, seldom use advanced features, and view few results' pages. Users of NAVER had an average session length of 1.8 queries.

There is a growing breadth and depth in research concerning Web searching and interest in a variety of issues from interactions, cognitive processes, to algorithm enhancements, with a notable emphasis on clustering. There is an increasing common lexicon in the analysis and presentation of results, which permits the contrasting of results among this body of research. However, there has been little comparison of findings across studies. Therefore, we do not know if these finding have external validity across the larger Web population and among the various Web search engine user groups. It is this issue that we address in this research by comparing results at key levels of analyses across a set of Web searching studies that provided significant data.

### **3. Research Questions**

We present the results from a comparative analysis across Web search engines focusing on following research questions:

1. What are the trends and differences in the number of one query sessions?
2. What are the trends and differences in the number of one-term queries?
3. What are the trends and differences in the number of result pages viewed?
4. What are the trends and differences in search topics?

In the next section, we present our research methodology.



## 4. Research Design

### 4.1 Data Collection

We utilize nine studies from currently published or forthcoming articles that provide significant data from searching on Web search engines. The nine studies we compare in this paper are shown chronologically in Table 1.

[Place Table 1 Here]

The nine studies include: (1) a 1997 study of the Excite Web search engine (Jansen, Spink & Saracevic, 2000), (2) a 1998 study of the Fireball Web search engine (Hölscher & Strube, 2000), (3) a 1998 study of the AltaVista Web search engine (Silverstein, et al., 1999), (4) a 1999 study of the Excite Web search engine (Wolfram et al., 2001), (5) a 2000 study of the BWIE Web search service (Cacheda & Viña, 2001a, 2001b), (6) a 2001 study of the AlltheWeb.com Web search engine (Spink, et al., 2002b), (7) a 2001 study of Excite Web search engine (Spink, et al., 2002a), (8) a 2002 of the AlltheWeb.com (Spink, et al., 2002b), and (9) a 2002 study of AltaVista (Jansen & Spink, Forthcoming). Collectively, the nine studies represent 287,212,000 (nearly 300 million) Web searching sessions and 1,015,126,814 (over 1 billion) queries that people submitted to the Web search engines.

If one views the studies from the geographical perspective of the Web search engine, there is a European and an U.S. grouping. For the analysis of European Web searching trends, we examined results from four studies over a five year period from three Web search engines. Fireball (<http://www.fireball.com>) is a predominantly German Web search engine. BWIE (<http://www.biwe.com/>) is a Spanish Web search service, and AlltheWeb.com (<http://www.alltheweb.com>) is a Web search engine based in Norway.

Our analysis of U.S.-based Web search engines covers five studies and data samples over a six period from two Web search engines. Excite (<http://www.excite.com>) was a major Web search engine at the time of the studies and is now a meta-search service. AltaVista

(<http://www.altavista.com>) was an independent Web search engine from 1998 through 2002 and is now a Web search engine within the Yahoo! Search (<http://www.yahoo.com>) network. Other published studies did not provide a rich enough data set for comparison at the time of the study. We could not obtain data from other Web search engines in either Europe or the U.S. (e.g., Google, MSN) at the time of the study.

#### 4.2 Data Analysis

We compare the changes in session length, query length, operator usage, and number of results pages viewed across these nine studies.

- *Session length* is the number of queries that a searcher submits in one episode with a Web search engine. We define an episode as the period from the first recorded time stamp to the last recorded time stamp on the search engine server from a particular searcher in a particular day.
- *Query length* is the number of terms in a query.
- *Term* is a series of alpha-numeric characters separated by white space or other delimiter.
- *Operator usage* is the number of Boolean or other operators in a query (i.e., *AND*, *OR*, *MUST APPEAR*, *PHRASE*).
- A *results page* is the set of usually 10 ranked uniform resources locators (URL) of Web documents (i.e., organic results) and other information (i.e., sponsored results) that a search engine presents to the user in response to a query.
- A *results page viewed* is the viewing of a results page by a searcher while trying to locate relevant documents.

The nine studies all use large-scale Web transaction logs that contain records of the interactions between searchers and the particular Web search engine. Web transaction logs

allow for the analysis of aggregate Web search characteristics and trends, and are beneficial for understanding aspects of the *real search process* (i.e., a real user with a real information need using a working system and content). However, data on individual identities is typically not in a Web transaction log. A Web transaction log also does not record the reasons for the search, the searcher's motivations, or other qualitative aspects of user. In addition, client-side caching may result in incomplete data logging of the number of identical Web queries from users. However, Web transaction logs have the advantage of unobtrusively recording real interactions by real users in the pursuit of real information needs in the complex Web information environment. This natural interaction in such a realistic environment is difficult to recreate in a laboratory setting (Dumais, 2002).

Web transaction logs follow a standard format and usually contain at least the following fields: (1) *Time of Day*: measured in hours, minutes, and seconds from some daily time mark, (2) *User Identification*: an anonymous user code assigned by the server representing the Internet Protocol address of the client's computer, and (3) *Query*: terms entered by the user. (4) *Results Page*: a code representing a set of URLs and result abstracts returned by the Web-search engine in response to a query.

## 5. Results

We present the results of our comparative analysis at the session, query, and results page levels of analysis from 1997 to 2002 across the 9 data sets. Since the absolute numbers of sessions, queries, and results pages vary for each study, we use the percentages for comparison.

## 5.1 Sessions

At the session level, we analyze the percentage of sessions with only one query (i.e., a searcher submits one query and then departs) on each Web search engine. The trend in the percentage of one query sessions will inform us whether or not the number of queries per user is increasing or decreasing. Figure 1 displays the results of this session analysis.

[Place Figure 1 Here]

All figures in this paper follow a similar layout. The x-axis is the year of the study. The y-axis is the measured percentage for a particular metric. The dark bar columns show the data points for the European studies. The light bar columns show the data points for the U.S. studies. There is a label on the columns identifying each study (i.e., ATW – AlltheWeb.com, AV – AltaVista, BWIE – BWIE, EX – Excite, FB – Fireball).

Figure 1 shows that for the U.S. Web search engines, it does not appear that the complexity of interactions is increasing as indicated by longer sessions (i.e., users submitting more Web queries). We conducted a Chi-Square goodness of fit procedure to evaluate whether or not the percentage of one query session across Web search engines was significantly different. A Chi-Square test indicated only marginally significance difference among the Web search engines in terms of percentage of one query sessions ( $\text{Chi-Square}(6) = 11.09$ ,  $p = 0.086$ ). However, if the 1998 AltaVista dataset is removed, there is no significant difference among the remaining search engine data sets ( $\text{Chi-Square}(5) = 2.505$ ,  $p = 0.776$ ). This would indicate that the temporal cut-off used for analysis in the 1998 AltaVista study (Silverstein, et al., 1999) was too short.

In 2002, approximately 47% of searchers on AltaVista submitted only one query, down from 77% in 1998. In the 1998 study, however, a session was artificially limited to five minutes. Subsequent research has shown that the typical Web session is about fifteen minutes (He, Göker & Harper, 2002; Jansen & Spink, 2003). Therefore, the 1998 AltaVista study probably over estimates the number of one query sessions. The downward trend also appears with

Excite users from 1999 to 2002, dropping from 60% to 55%, although not a significant decrease.

The data analysis methods were similar for all Excite studies and did not impose a session time limit.

The session data for European users is available from two Web search engines, BWIE and AlltheWeb.com. For these European Web search engines, there is also no significant change in one query sessions. So, for session length, the trend appears to be one of stability, with no differences among search engines.

## 5.2 Queries

At the query length level, we analyze the percentage of queries with only one term. The percentage of one term queries will inform us whether or not the length of queries is increasing or decreasing. Figure 2 displays the results for the analysis of Web query lengths.

[Place Figure 2 Here]

A Chi-Square test did indicate a significant difference among the Web search engines in terms of percentage of one term queries (Chi-Square (7) = 26.43,  $p = 0.01$ ). However, if the 1998 Fireball dataset is removed, there is no significant difference among the remaining search engine data sets (Chi-Square (6) = 3.72,  $p = 0.714$ ). This would indicate that there is something in the Fireball user base, content, or system that differentiates it from users of the other Web search engines.

For the U.S.-based Web-search engines the percentage of one-term queries is holding steady, within a range of 20% to 29% of all queries. Using data from 1999 onward, the trend with U.S.-based Web-search engines appears to be of one-term queries declining as a percentage of all queries, dropping from 30% to 20%.

For the Europe-based Web-search-engine users, the trend appears to be one of little change, although there is a spike in 2002 with AlltheWeb.com users. Otherwise, we see a

percentage of one-term queries on these European-based Web-search engines within a range of about 25% to 35%, excluding the 1998 Fireball study.

### *5.3 Query Operators*

We also analyze the percentage of Web queries containing searching operators. The trend in the percentage of queries with searching operators will inform us whether or not the complexity of query structure is increasing or decreasing.

Based on the use of advanced operators, the complexity of interaction appears to be at least remaining stable. Figure 3 shows the results for query operator usage on the various Web search engines.

[Place Figure 3 Here]

The usage of query operators appears to be search-engine dependent, and there is a notable regional difference. A Chi-Square test indicated significant difference among the US Web search engines in terms of percentage of usage of query operators (Chi-Square (4)= 16.383,  $p = 0.01$ ). A Chi-Square test indicated no significant difference among the three Excite search-engine data sets in terms of percentage of usage of query operators (Chi-Square (2)= 0.258,  $p = 0.879$ ). A Chi-Square test indicated no significant difference among the two AltaVista search-engine data sets in terms of percentage of usage of query operators (Chi-Square (1)= 1.33,  $p = 0.244$ ). This indicates that there is a search engine dependency in terms of the use of query operators with a particular search engine system.

For the AltaVista Web search engine, the usage of query operators has held steady at approximately 20%. For the Excite Web search engine, the usage increased steadily from 1997 to 2001, although not a statistically significant variation between data sets.

For the European-based Web search engines, the usage also varied among the three Web search engines, but these searchers seldom use advanced operators. A Chi-Square test

indicated no significant difference among the four European search data sets in terms of percentage of usage of query operators (Chi-Square (3)= 4.4,  $p = 0.221$ ), with the usage was extremely low on all.

The most notable feature of operator usage is the rather large gap between usage on the U.S. and European-based Web search engines. The usage of query operators on the U.S.-based Web search engines varied from 11% to 20%. The usage on the European-based Web search engines varied from 2% to 10% and held fairly stable at under 5% from 1998 to 2001.

#### *5.4 Results Pages*

We analyze the percentage of users viewing only one results page. This trend will inform us how persistent searchers are when locating information or services on the Web. Overall, it appears that Web searchers are tending to view fewer documents per Web query, which might indicate a move to less complex interactions. Figure 4 presents results-page-viewing findings.

[Place Figure 4 Here]

We see that the percentage of searchers viewing only one results page is increasing for users of both U.S. and European based Web search engines. The percentage of searchers viewing only the first results page has increased from 29% in 1997 to 73% in 2002 for U.S. based Web search engines users. Again, the 1998 AltaVista study limited sessions to five minutes, which probably increased the percentage of sessions with only one page result. For European searchers, the variability ranged from 60% to 83%, although there was a dip to 76% in 2002.

A Chi-Square test indicated significant difference among the Web search engines in terms of percentage of single result page viewing (Chi-Square (8)= 45.743,  $p = 0.01$ ). A Chi-Square test indicated a significance difference among the three Excite Web search engine data sets in terms of percentage of single result page viewing (Chi-Square (2)= 6.049,  $p = 0.05$ ). A Chi-Square test indicated no significance difference among the two AltaVista search engine data

sets in terms of percentage of single result page viewing (Chi-Square (1)= 0.911,  $p = 0.34$ ). A Chi-Square test indicated no significance difference among the four European search data sets in terms of percentage of percentage of single result page viewing (Chi-Square (3)= 4.136  $p = 0.247$ ). Therefore, there was trend among Excite users to view fewer result pages. Excite users viewed more result pages than users of other Web search engines. However, as time processed, the tendency was to view fewer.

### 5.5 Topical Classification

For the six Web query data sets that we had access to, we qualitatively analyzed a random sample of approximately 2,600 queries from each in order to determine trends in the type of information people are searching for on the Web. We classified each query into eleven non-mutually exclusive, general topic categories developed by Spink, Jansen, Wolfram and Saracevic (2002a). At least two independent evaluators manually classified queries from each data set independently. The evaluators then met and resolved discrepancies.

Table 2 and 3 display the topical evaluation results for European and U.S. based Web search engines, respectively.

[Place Table 2 Here]

For searching on AlltheWeb.com, *People, Places or Things* category remained the top ranked category with a large percentage increase from 2001 to 2002, accounting for over forty percent of queries. *Commerce, Travel, Employment or Economy* and *Computers, Internet or Technology* accounted approximately 25% of the queries. Noticeably percentage decreases occurred in *Computers or Internet, Entertainment or recreation*, and *Sex or Pornography*. A Chi-square goodness of fit test indicates a significant difference between the Web search engine data sets based on category of *People, Place or Things* (Chi-Square (3)= 5.554  $p = 0.05$ ).



[Place Table 3 Here]

On the U.S. based Web search engines. Queries for *People, Place or Things* account for nearly half of the queries in 2002, with *Commerce, Travel, Employment or Economy* and *Computers, Internet or Technology* accounting for another 25% of the queries. There appears to be a steady rise in searching for *People, Place or Things* and *Commerce, Travel, Employment or Economy*, with decreased searching for *Sex and pornography* and *Entertainment or recreation*. A Chi-squared goodness of fit test indicated significant differences among the Web search engines data sets based on distribution of queries among categories in the areas of *People, places, or things* (Chi-Square (3)= 39.317  $p = 0.01$ ), *Entertainment or recreation* (Chi-Square (3)= 13.80  $p = 0.01$ ), and *Sex and pornography* (Chi-Square (3)= 10.892  $p = 0.05$ ). There was a marginally significant difference with the category of *Commerce, travel, employment, or economy* (Chi-Square (3) = 4.136  $p = 0.06$ ). There was no significant difference among the datasets in the other categories.

## 6. Discussion

As the Web is becoming a worldwide phenomenon, we need to understand better the emerging trends in Web searching given the tremendous influence Web search engines have on directing traffic to online information and services. Our findings indicate that the interactions between Web search engines and searchers are not becoming more complex, and in some respects, are becoming less complex. Our comparative analysis also indicates that finding from a study focusing on one Web search engine can not be applied wholesale to all Web search engines.

Sessions lengths are not increasing as measured by number of queries. The percentage of one term sessions is remaining stable over time and across Web search engines. There was a

difference with the 1998 AltaVista study, but this appears to be caused by an artificially short session duration that the researchers used. Query lengths are also not increasing as measured by number of terms. There was a statistical difference in the percentage of one term queries on the German Fireball Web search engine, which may be due to linguistic differences with the other Web search engines. The percentage of single-term queries is holding steady, and the use of query operators is also remaining stable. Web search engines in the future may better leverage the implicit feedback from this interaction to provide more personalized results (Callan & Smeaton, 2003). However, the use of query operators between Web search engines varies significantly, so in this area findings from one study can not necessary be applied to predict behaviors on other Web search engines.

The viewing of only the first page of results is extremely high, and it significantly increased over time on the Excite Web search engine. This may indicate increasing simplicity in interactions. It may also be an indication of the increasing ability of Web search engines to retrieve and rank Web documents more effectively. There is certainly a need for more studies that focus on the Web document and virtual document (Watters, 1999) level of analysis.

The trend toward view fewer result pages with Excite users may be related to a changing user base during the time of the study as the Web population dramatically increased during this time. Excite was the second most popular Web site in 1997 (Munarriz, 1997), and was the fifth most popular in 1999 and 2001 as measured by number of unique visitors (Cyber Atlas, 1999, 2001).

There are both similarities and differences between usage on U.S. and European-based Web search engines. Searchers on both are similar in session length, query length, and number of results pages viewed. Additionally, the use of Web query operators on both is fairly stable. However, the usage of these advanced Web-query operators is much higher on U.S.-based Web search engines than on their European counterparts. In investigating this difference, we ruled out size of content collections (they are all immense), user bases (they all number in the

millions), or algorithmic sophistication (they are all similar in performance tests). Fireball and BWIE did not prominently display the advanced Web searching options; however, it may be that users of these Web search engines just do not use query operators. This increases the criticality of keyword and phrase selection for Web providers targeting these users.

Fireball is a general purpose Web search engine, but, BWIE is also a search directory. A search directory supplements query matching of the entire content collection with directory-based search (c.f., Yahoo <http://www.yahoo.com> or Open Directory <http://dmoz.org/>). The idea behind directory services is to provide additional organization to the content. However, some research has shown that directory-based searching does not improve searching performance and also takes longer (Dennis, Bruza & McArthur, 2002). There are variations of the search directory including specialized or niche Web search engines that provide content within a specific Web search engines, including computer science literature (CiteSeer <http://www.researchindex.com>), e-commerce (Froogle <http://froogle.google.com/>), or personal information (c.f., <http://www.switchboard.com>). Some Web search engines provide clustering (Vivisimo <http://vivisimo.com/>), which one can view as an automated, real time, and virtual directory service.

AlltheWeb.com has extensive advanced Web search features, however. Additionally, the results of the 2002 AlltheWeb.com data set do not conform to the results from studies of the other European based Web search engines. One possible reason may be that AlltheWeb.com is attracting searchers outside of its traditional European market. From our analysis of the AlltheWeb.com transaction log, nearly 90% of the query requests are in English, with 6% French, 1% each Spanish, German, Italian, and a variety of other languages making up the rest. Further research will be needed to isolate the effects of linguistic differences.

Web searching topics are changing. There was a decrease in sexual searching as a percentage of overall Web searching on both European and U.S. based Web search engines. The overall trend is towards using the Web as a tool for information or commerce, rather than

entertainment. This trend is more pronounced with U.S. as opposed to European searchers. This analysis certainly confirms survey and other data that the Web is now a major source of information for most people (Cole, et al., 2003; Fox, 2002). There is increased use of the Web as an economic resource and tool (Lawrence & Giles, 1999; Spink, et al., 2002a), and people use the Web for an increasingly variety of information tasks (Fox, 2002; National Telecommunications and Information Administration, 2002).

The decreased level of interaction of Web searches may be unwelcome news for Web-search engine developers and for those providing Web-based information content, products, and services. Web users appear unwilling to invest additional effort to locate relevant Web content. The trend towards viewing only the first results page is a challenge for those seeking to draw visitors to their Web sites or for Web search engines attempting to generate revenue via ad impressions. Users have a low tolerance of viewing any results past the first page. They prefer to reformulate the Web query rather than wade through result listings. Placement within the first page of Web search engine results of an accurate abstract appears to be a determining factor in drawing traffic to a particular Web site.

We continue to conduct ongoing analysis of Web searching trends to provide a valuable insight into this important and critical area of human computer interaction and electronic commerce.

## 7. References

- Amichai-Hamburger, Y. (2002). Internet and personality. *Computers in Human Behavior*, 18(2002), 1-10.
- Baeza-Yates, R., & Castillo, C. (2000, October). *Relating web characteristics [in Spanish]* [Web Site]. University of Chile. Retrieved 15 July, 2002, from the World Wide Web: <http://www.todocl.cl/stats/rbaeza.pdf>.
- Baeza-Yates, R., & Castillo, C. (2001). Relating web structure and user search behavior. In *Proceedings of the 10th World Wide Web Conference*, pp. 1-2. Hong Kong, China. 1-5 May.

- Bar-Ilan, J. (2004). The use of web search engines in information science research. In B. Cronin (Ed.), *Annual review of information science and technology* (Vol. 33, pp. 231-288). Medford, NY, USA: Information Today.
- Bates, M. J. (1990). Where should the person stop and the information search interface start? *Information Processing and Management*, 26(5), 575-591.
- Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., & Frieder, O. (2004). Hourly analysis of a very large topically categorized web query log. In *Proceedings of the 27th annual international conference on Research and development in information retrieval*, pp. 321 - 328. Sheffield, U.K., 25-29 July.
- Belkin, N., Cool, C., Stein, A., & Theil, S. (1995). Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems With Applications*, 9(3), 379-395.
- 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.
- Callan, J., & Smeaton, A. (2003). *Personalisation and recommender systems in digital libraries. Joint nsf-eu delos working group report*. Joint NSF-EU DELOS Working Group Report. Retrieved 1 January, 2002, from the World Wide Web: <http://www-2.cs.cmu.edu/~callan/Papers/Personalisation03-WG.pdf>.
- Chen, H.-M., & Cooper, M. D. (2001). Using clustering techniques to detect usage patterns in a web-based information system. *Journal of the American Society for Information Science and Technology*, 52(11), 888-904.
- Chen, H.-M., & Cooper, M. D. (2002). Stochastic modeling of usage patterns in a web-based information system. *Journal of the American Society for Information Science and Technology*, 53(7), 536-548.
- Cole, J. I., Suman, M., Schramm, P., Lunn, R., & Aquino, J. S. (2003, February). *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>.
- Cyber Atlas. (1999). *U.S. Top 50 internet properties, December 1999, at home/work combined* [Website]. CyberAtlas. Retrieved 1 July, 2000, from the World Wide Web: <http://cyberatlas.internet.com>.
- Cyber Atlas. (2001). *U.S. Top 50 internet properties, may 2001, at home/work combined* [Website]. CyberAtlas. Retrieved 1 July, 2000, from the World Wide Web: <http://cyberatlas.internet.com>.

- Dennis, S., Bruza, P., & McArthur, R. (2002). Web searching: A process-oriented experimental study of three interactive search paradigms. *Journal of the American Society for Information Science and Technology*, 53(2), 120-133.
- Dumais, S. T. (2002, 7-11 May). *Web experiments and test collections* [Presentation]. Retrieved 20 April, 2003, from the World Wide Web: <http://www2002.org/presentations/dumais.pdf>.
- Efthimiadis, E. N., & Robertson, S. E. (1989). Feedback and interaction in information retrieval. In C. Oppenheim (Ed.), *Perspectives in information management* (pp. 257-272). London: Butterworths.
- Eiron, N., & McCurley, K. (2003). Analysis of anchor text for web search. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 459-460. Toronto, Canada. 28 July - 1 August.
- Fox, S. (2002, July 2002). *Search engines* [website]. The Pew Internet & American Life Project. Retrieved 15 October 2002, 2002, from the World Wide Web: <http://www.pewinternet.org/reports/toc.asp>.
- Hansen, M. H., & Shriver, E. (2001). Using navigation data to improve ir functions in the context of web search. In *Proceedings of the tenth international conference on Information and Knowledge Management*, pp. 135 - 142. Atlanta, Georgia, USA. October.
- 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.
- Hsieh-Yee, I. (2001). Research on web search behavior. *Library & Information Science Research*, 23(1), 168-185.
- Huang, C.-K., Chien, L.-F., & Oyang, Y.-J. (2003). Relevant term suggestion in interactive web search based on contextual information in query session logs. *Journal of the American Society for Information Science and Technology*, 54(7), 638 - 649.
- 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. (Forthcoming). An analysis of web searching by european alltheweb.Com users. *Information Processing and Management*.
- 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.

- Kea, H.-R., Kwakkelaar, R., Taic, Y.-M., & Chen, L.-C. (2002). Exploring behavior of e-journal users in science and technology: Transaction log analysis of Elsevier's sciencedirect onsite in Taiwan. *Library & Information Science Research*, 24(1), 265–291.
- Lalmas, M., & Ruthven, I. (1999). A framework for investigating the interaction in information retrieval. In *Proceedings of 9th European-Japanese Conferences on Information Modeling and Knowledge Bases*, pp. 222-239. Iwate, Japan. 24-28 May.
- Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the web. *Nature*, 400, 107-109.
- Lempel, R., & Moran, S. (2003). Predictive caching and prefetching of query results in search engines. In *Proceedings of the twelfth international conference on World Wide Web*, pp. 19 - 28. Budapest, Hungary.
- Liaw, S.-S., & Huang, H.-M. (2003). An investigation of user attitudes toward search engines as an information retrieval tool. *Computers in Human Behavior*, 19(2003), 751–765.
- Lin, S.-J. (2002). Design space of personalized indexing: Enhancing successive web searching for transmuting information problems. In *Proceedings of the Eighth Americas Conference on Information Systems*, pp. 1092 - 1100. Dallas, Texas. 9-11 August.
- Loken, E., Radlinski, F., Crespi, V. H., Millet, J., & Cushing, L. (2004). Online study behavior of 100,000 students preparing for the sat, act, and gre. *Journal of Educational Computing Research*, 30(3), 255-262.
- Montgomery, A., & Faloutsos, C. (2001). Identifying web browsing trends and patterns. *IEEE Computer*, 34(7), 94-95.
- Munarriz, R. A. (1997). *How did it double?* Daily Double. Retrieved 10 November, 2002, from the World Wide Web: <http://www.fool.com/DDouble/1997/DDouble970812.htm>.
- National Telecommunications and Information Administration. (2002). *A nation online: How americans are expanding their use of the internet*. Washington, D.C.: U.S. Department of Commerce.
- Nielsen Media. (1997). *Search engines most popular method of surfing the web* [web site]. Commerce Net / Nielsen Media. Retrieved 30 August, 2000, from the World Wide Web: <http://www.commerce.net/news/press/0416.html>.
- Ozmutlu, H. C., & Cavdur, F. (Forthcoming). Application of automatic topic identification on excite web search engine data logs. *Information Processing & Management*, In Press, Corrected Proof.
- Park, M., Bae, J., & Lee, S. (Forthcoming). End user searching: A web log analysis of naver, a korean web search engine. *Library & Information Science Research*, 27(2).
- Peters, T. (1993). The history and development of transaction log analysis. *Library Hi Tech*, 42(11), 41-66.

- Pu, H. T. (2000). An exploratory analysis on search terms of network users in taiwan [in chinese]. *Central Library Bulletin*, 89(1), 23-37.
- Rieh, S. Y., & Xu, H. (2001). Patterns and sequences of multiple query reformulation in web searching: A preliminary study. In *Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology*, pp. 246-255.
- Romano, N. C., Donovan, C., Chen, H., & Nunamaker, J. F. (2003). A methodology for analyzing web-based qualitative data. *Journal of Management Information Systems*, 19(4), 213-246.
- Ross, N., & Wolfram, D. (2000). End user searching on the internet: An analysis of term pair topics submitted to the excite search engine. *Journal of the American Society for Information Science*, 51(10), 949-958.
- 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. (2004). Multitasking information behavior and information task switching: An exploratory study. *Journal of Documentation*, 60(3), 336-345.
- Spink, A., & Jansen, B. J. (2004). *Web search: Public searching of the web*. New York: Kluwer.
- Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002a). 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. (2002b). U.S. Versus european web searching trends. *SIGIR Forum*, 32(1), 30 - 37.
- Spink, A., Wilson, T., Ellis, D., & Ford, F. (1998, April 1998). Modeling users' successive searches in digital environments. *D-Lib Magazine*.
- Voorbraak, F. (1991). On the justification of dempster's rule of combination. *Artificial Intelligence*, 48(1), 171-197.
- Wang, P., Berry, M., & Yang, Y. (2003). Mining longitudinal web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology*, 54(8), 743-758.
- Watters, C. (1999). Information retrieval and the virtual document. *Journal of the American Society for Information Science*, 50(11), 1028-1029.
- Wen, J.-R., Nie, J.-Y., & Zhang, H.-J. (2001). Clustering user queries of a search engine. In *Proceedings of the 10th International Conference on World Wide Web*, pp. 162 - 168. Hong Kong. 1-5 May.
- 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.



- Xie, Y., & O'Hallaron, D. (2002). Locality in search engine queries and its implications for caching. In *Proceedings of the Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies*, pp. 307-317. New York City, New York, USA. 23-27 June.
- Yu, L., & Apps, A. (2000). Studying e-journal user behavior using log files: The experience of superjournal. *Library & Information Science Research*, 22(3), 311–338.

Table 1. Aggregate data from Web search engine studies from 1997 through 2002.

Study No.	1	2	3	4
	Excite	Fireball	AltaVista	Excite
Region	U.S.	European	U.S.	U.S.
Data Collection	Tuesday 16 Sep. 1997	1-31 Jul. 1998	2 Aug. - 13 Sep. 1998	Wednesday 1 Dec. 1999
Sessions	211,063	Not Reported	285,474,117	325,711
Queries	1,025,908	16,252,902	993,208,159	1,025,910
Terms	1,277,763	Not Reported	Not Reported	1,500,500

	5	6	7	8	9
	BWIE	AlltheWeb.com	Excite	AlltheWeb.com	AltaVista
Region	European	European	U.S.	European	U.S.
Data Collection	3-18 May 2000	Tuesday 6 Feb. 2001	Monday 30 Apr. 2001	Tuesday 28-May-02	Sunday 8 Sep. 2002
Sessions	83,232	153,297	262,025	345,093	369,350
Queries	71,810	451,551	1,025,910	957,303	1,073,388
Terms	116,953	1,350,619	1,538,120	2,225,141	1,073,388

**Figure 1.** Percentage of single query sessions.

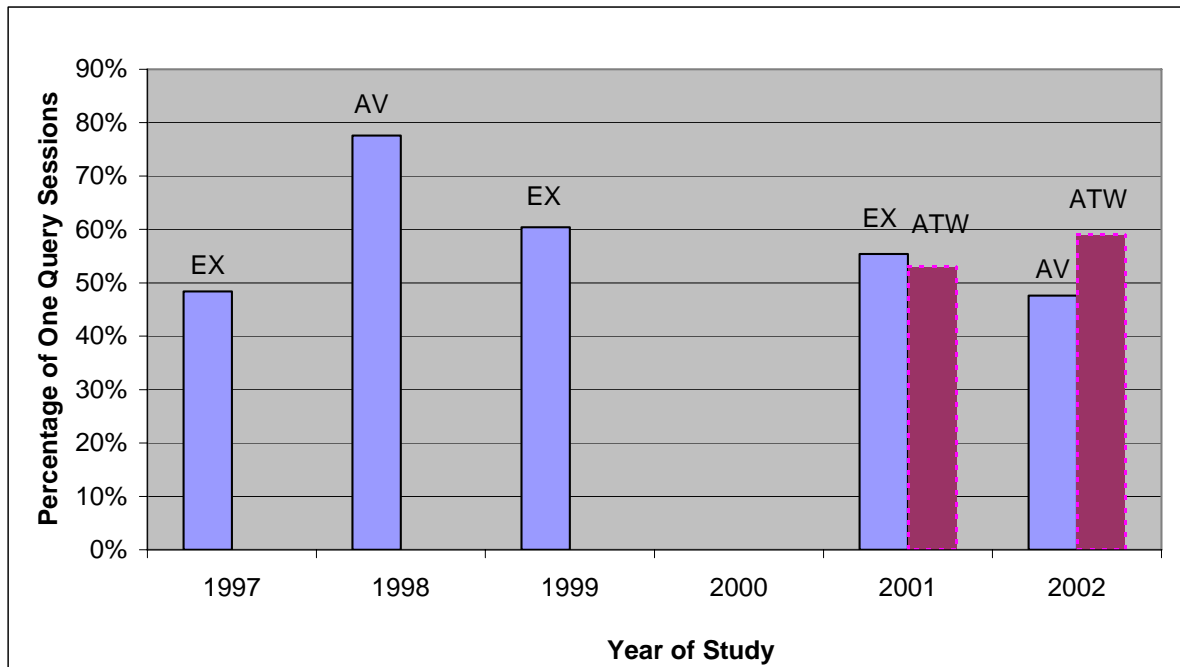


Figure 2. Percentage of one-term queries.

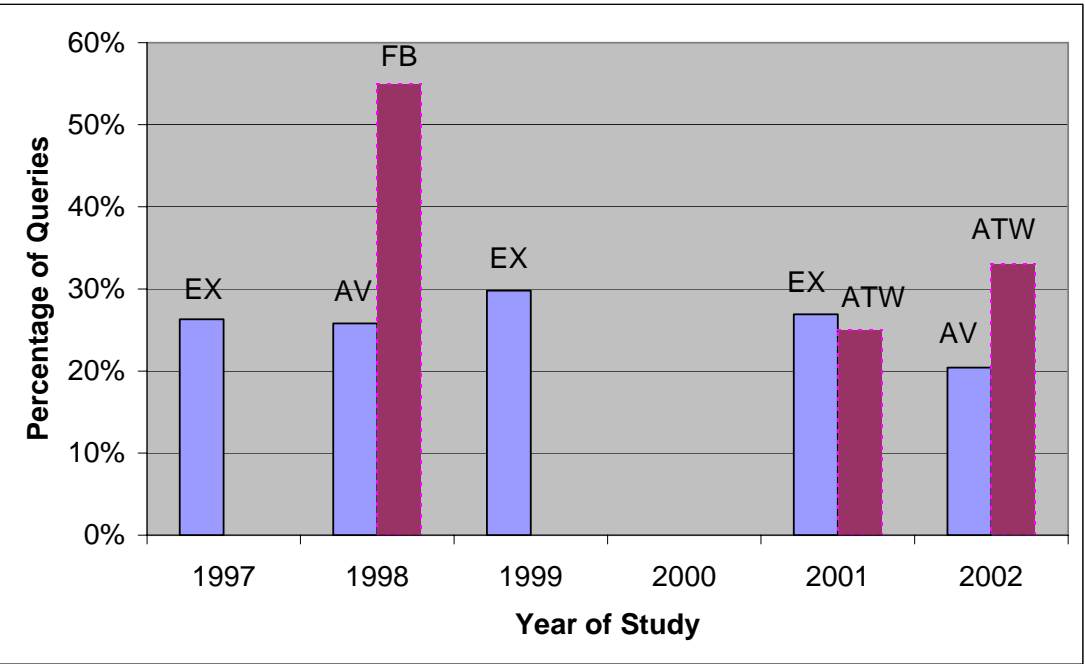


Figure 3. Percentage of operator usage.

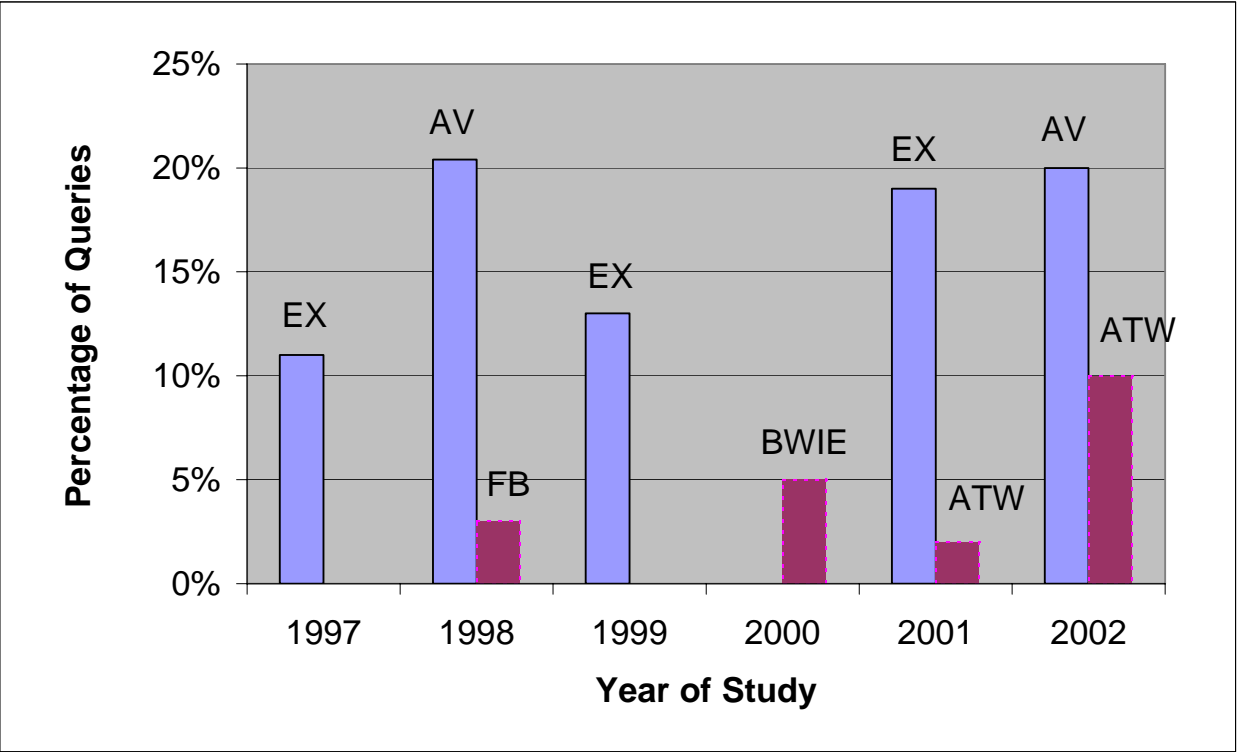


Figure 4: Percentage of single result page viewing.

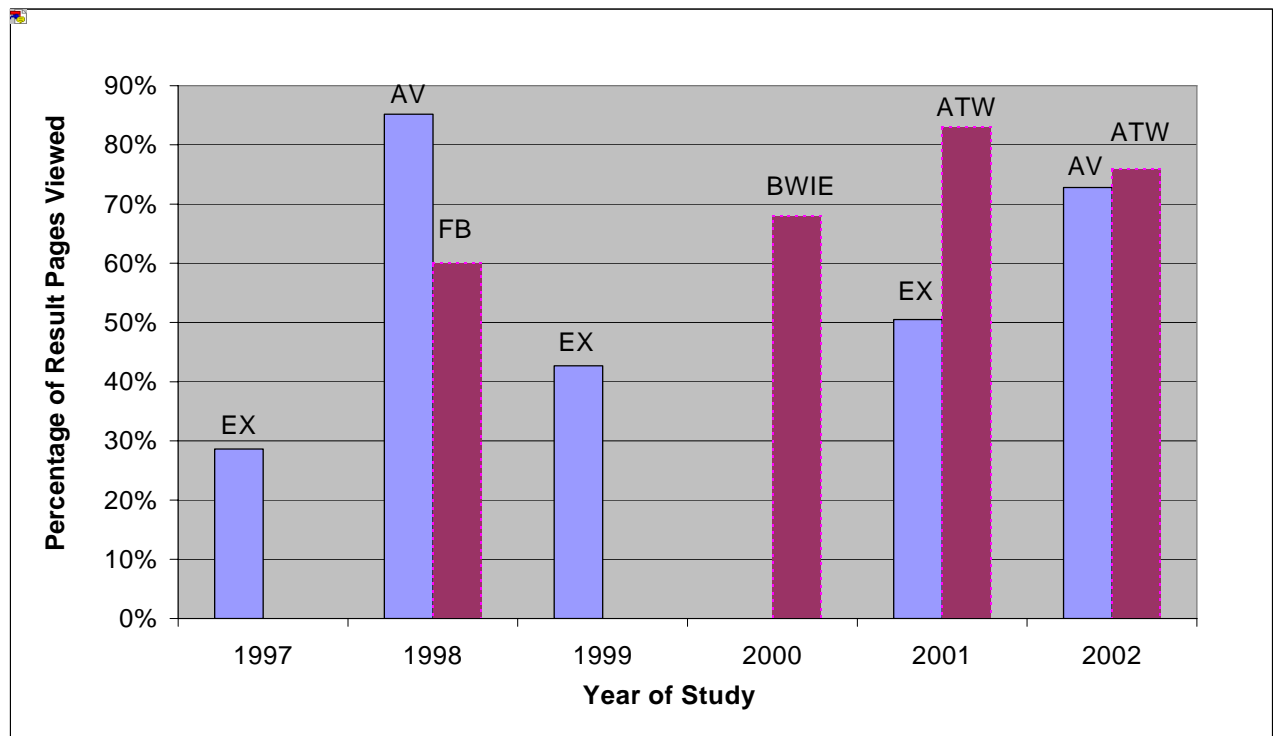


Table 2. Distribution of AlltheWeb.com general topic categories.

	Categories	2001 (2,503 English Queries)	2002 (2,525 English Queries)
1	People, places or things	<b>22.5%</b>	<b>41.5%</b>
2	Computers or Internet	21.8%	16.3%
3	Commerce, travel, employment, or economy	12.3%	12.7%
4	Sex or pornography	10.8%	9.5%
5	Entertainment or recreation	9.1%	4.9%
6	Health or sciences	7.8%	4.5%
7	Society, culture, ethnicity or religion	4.8%	2.6%
8	Performing or fine arts	4.7%	2.5%
9	Education or humanities	2.9%	2.3%
10	Government	2.7%	2.1%
11	Unknown or Other	0.6%	1.1%
		100.0%	100.0%

Note: Bolded percentages indicate the highest ranked topic in a given year.

Table 3. Distribution of Excite and AltaVista general topic categories.

	Categories	1997 Excite (2,414 queries)	1999 Excite (2,539 queries)	2001 Excite (2,453 queries)	2002 AltaVista (2,603 queries)
1	People, places, or things	6.7%	20.3%	19.7%	<b>49.3%</b>
2	Commerce, travel, employment, or economy	13.3%	<b>24.5%</b>	<b>24.7%</b>	12.5%
3	Computers or Internet	12.5%	10.9%	9.7%	12.4%
4	Health or sciences	9.5%	7.8%	7.5%	7.5%
5	Education or humanities	5.6%	5.3%	4.6%	5.0%
6	Entertainment or recreation	<b>19.9%</b>	7.5%	6.7%	4.6%
7	Sex and pornography	16.8%	7.5%	8.6%	3.3%
8	Society, culture, ethnicity, or religion	5.7%	4.2%	3.9%	3.1%
9	Government	3.4%	1.6%	2.0%	1.6%
10	Performing or fine arts	5.4%	1.1%	1.2%	0.7%
11	Non-English or unknown	4.1%	9.3%	11.4%	0.0%
		102.9%	100.0%	100.0%	100.0%

Note: Bolded percentages indicate the highest ranked topic in a given year.