# Web Searcher Interaction With the Dogpile.com Metasearch Engine

**Bernard J. Jansen**

*College of Information Sciences and Technology, The Pennsylvania State University, 329F IST Building, University Park, PA 16802. E-mail: jjansen@ist.psu.edu*

**Amanda Spink**

*Faculty of Information Technology, Queensland University of Technology, Gardens Point Campus, 2 George Street, GPO Box 2434, Brisbane QLD 4001, Australia. E-mail: ah.spink@qut.edu.au*

**Sherry Koshman**

*School of Information Sciences, University of Pittsburgh, 610 IS Building, 135 N. Bellefield Avenue, Pittsburgh, PA 15260. E-mail: aspink@sis.pitt.edu*

**Metasearch engines are an intuitive method for improving the performance of Web search by increasing coverage, returning large numbers of results with a focus on relevance, and presenting alternative views of information needs. However, the use of metasearch engines in an operational environment is not well understood. In this study, we investigate the usage of Dogpile.com, a major Web metasearch engine, with the aim of discovering how Web searchers interact with metasearch engines. We report results examining 2,465,145 interactions from 534,507 users of Dogpile.com on May 6, 2005 and compare these results with findings from other Web searching studies. We collect data on geographical location of searchers, use of system feedback, content selection, sessions, queries, and term usage. Findings show that Dogpile.com searchers are mainly from the USA (84% of searchers), use about 3 terms per query (mean = 2.85), implement system feedback moderately (8.4% of users), and generally (56% of users) spend less than one minute interacting with the Web search engine. Overall, metasearchers seem to have higher degrees of interaction than searchers on non-metasearch engines, but their sessions are for a shorter period of time. These aspects of metasearching may be what define the differences from other forms of Web searching. We discuss the implications of our findings in relation to metasearch for Web searchers, search engines, and content providers.**

## Introduction

Metasearch engines have an intuitive appeal as a method of improving the retrieval performance for Web searches.

Unlike single source Web search engines, metasearch engines do not crawl the Internet themselves to build an index of Web documents. Instead, a metasearch engine sends queries simultaneously to multiple other Web search engines, retrieves the results from each, and then combines the results from all into a single results listing, at the same time avoiding redundancy. In effect, Web metasearch engine users are not using just one engine, but many search engines at once to effectively utilize Web searching. The ultimate purpose of a metasearch engine is to diversify the results of the queries by utilizing the innate differences of single source Web search engines and provide Web searchers with the highest ranked search results from the collection of Web search engines. Although one could certainly query multiple search engines, a metasearch engine distills these top results automatically, giving the searcher a comprehensive set of search results within a single listing, all in real time.

We know that there is little overlap among typical search engine result listings (Ding & Marchionini, 1996), and single search engines index a relatively small percentage of the Web (Lawrence & Giles, 1999). Research shows that results retrieved from multiple sources have a higher probability of being relevant to the searcher's information needs (Gauch, Wang, & Gomez, 1996). Finally, a single search engine may have inherent biases that influence what results are returned (Gerhart, 2004; Introna & Nissenbaum, 2000). By combining results from several sources, a metasearch engine addresses all three concerns.

Chignell, Gwizdka, and Bodner (1999) found little overlap in the results returned by various Web search engines. They describe a metasearch engine as useful, since different engines employ different means of matching queries to relevant items,

and also have different indexing coverage. Selberg and Etzioni (1997) further suggested that no single search engine is likely to return more than 45% of the relevant results. Subsequently, the design and performance of metasearch engines have become an ongoing area of study (Buzikashvili, 2002; Chignell, Gwizdka & Bodner, 1999; Dreilinger & Howe, 1997; Meng, Yu, & Lui, 2002; Selberg & Etzioni, 1997; Spink, Lawrence, & Giles, 2000).

However, there has been little investigation into how searchers interact with Web metasearch engines. If metasearch provides an improved Web searching environment, one may expect differences in interactions when compared to Web searching on other search engines. What are the interaction patterns between searchers and a metasearch engine? This question motivates our research.

In the following sections, we review the related studies and list our research questions. We then discuss the Dogpile.com Web metasearch engine and the research design that was used in our study. We then discuss the findings from multiple levels of analysis, concluding with implications for Web metasearching.

## Related Studies

Web research is now a major interdisciplinary area of study, including the modeling of user behavior and Web search engine performance (Spink & Jansen, 2004). Web search engine crawling and retrieving studies have evolved as an important area of Web research since the mid-1990s. Many metasearch tools have been developed and commercially implemented, but little research has investigated the usage and performance of Web metasearch engines. Selberg and Etzioni (1997) developed one of the first metasearch engines, Metacrawler (http://www.metacrawler.com). Largely focusing on the system design, the researchers discuss usage, reporting on 50,878 queries submitted between July 7 and September 30, 1995, with 46.67% (24,253 queries) being unique. The top 10 queries represented 3.37% (1,716) of all queries. The top queries were all one term in length, and commonly occurring natural language terms (e.g., *the*, *of*, *and*, *or*) reported in later Web user studies were not present.

Gauch, Wang, and Gomez (1996) designed the ProFusion metasearch engine and evaluated its performance in a lab setting. The researchers used 12 students who submitted queries and compared ProFusion to the six underlying search engines using the number of relevant documents retrieved, the number of irrelevant documents retrieved, the number of broken links, the number of duplicates, the number of unique relevance documents and precision. How the study participants utilized the metasearch engine was not discussed.

The SavvySearch (Dreilinger & Howe, 1997; Howe & Dreilinger, 1997) is a metasearch engine that selects the most promising search engines automatically. It then sends the user's query to the selected two or three search engines in parallel. The researchers evaluated various implementations of SavvySearch (Dreilinger & Howe, 1997) using system load as the metric of comparison. Searching characteristics were not presented.

Developers of the Mearf metasearch engine (Oztekin, Karypis, & Kumar, 2002) collected transaction logs from November 22, 2000 to November 10, 2001, using click-through as a mechanism for evaluating Mearf performance. They report on the mean documents returned per query, user reranking of results, and the number of documents clicked on by searchers. Approximately 64% of queries included a click on a document, with a mean of 2.02 clicks per query. However, there were a total of 17,055 queries submitted during the one year period, so this may not be a representative sample of metasearch engine users.

Many studies have examined the performance of single Web search engines such as AltaVista, Excite, AlltheWeb (Spink & Jansen, 2004), and NAVER (Park, Bae, & Lee, 2005). Spink, Jansen, Blakely, and Koshman (2006) found little results overlap and uniqueness among major Web search engines. However, limited large-scale studies have examined how searchers interact with Web metasearch engines. An understanding of how searchers utilize these systems is critical for the future refinement of metasearch engine design and the evaluation of Web metasearch engine performance. These are the motivators for our research.

## Research Questions

The research questions driving our study are as follows:

1. What are the characteristics of search interactions on the Dogpile.com metasearch engine? To address this research question, we investigated session length, query length, query structure, query formulation, result pages viewed and term usage of these Web searchers.
2. What are the temporal characteristics of metasearching on Dogpile.com? For this research question, we investigated the duration of sessions and the frequency of interactions during these sessions.
3. What are the topical characteristics of searches on the Dogpile.com metasearch engine? To address this research question, we investigated a subset of queries submitted by searchers on Dogpile.com to gain insight into the nature of their search topics using a qualitative analysis.

## Research Design

### Dogpile.com

Dogpile.com (http://www.Dogpile.com/) is owned by Infospace, a market leader in the metasearch engine business. Dogpile.com incorporates into its search result listings the results from other search engines, including results from the four leading Web search indices (i.e., Ask Jeeves, Google, MSN, and Yahoo!). With listings that include results from these four Web search engines, Dogpile.com leverages one of the most comprehensive content collections on the Web in response to searchers' queries.
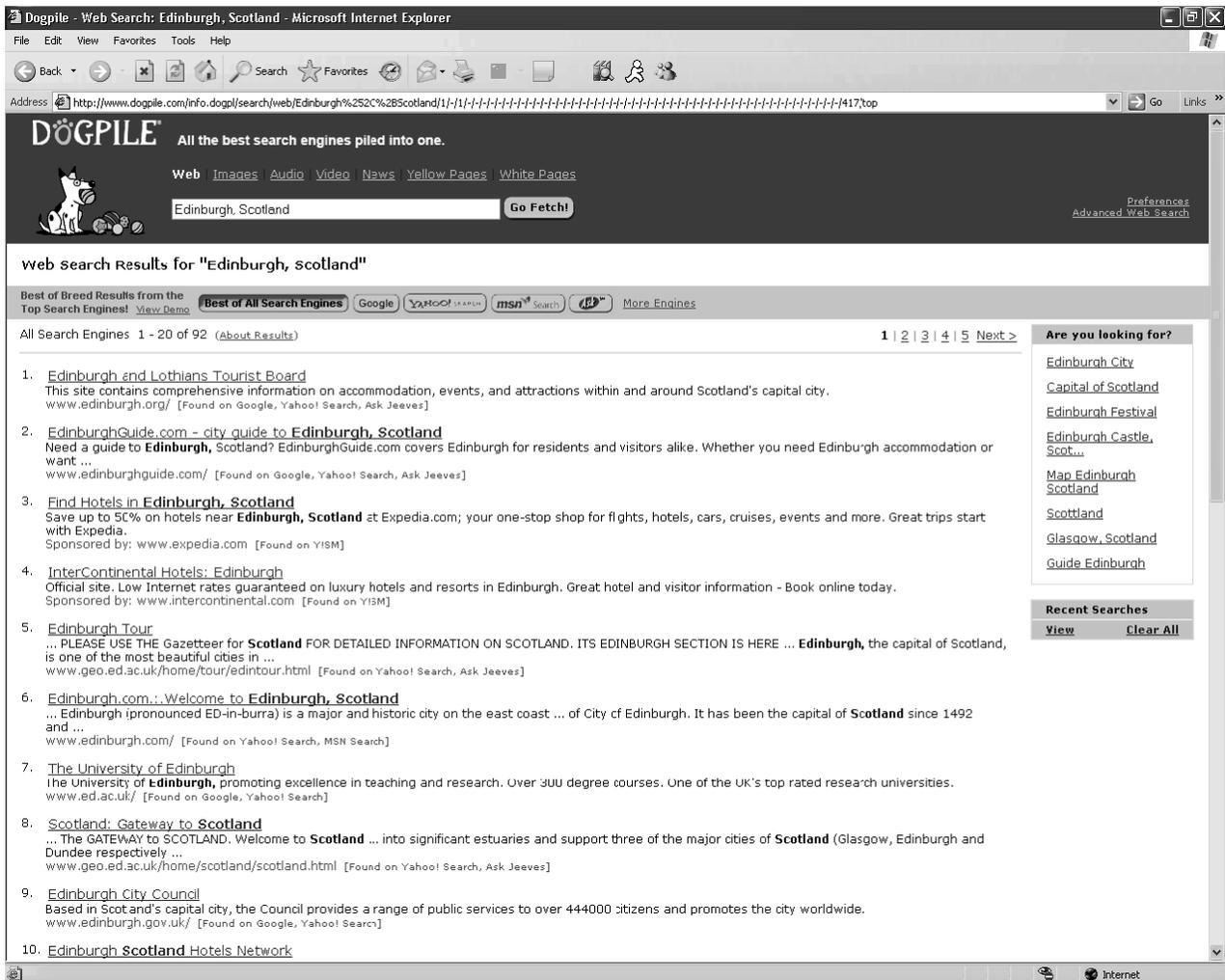
FIG. 1.    Dogpile.com metasearch interface.

When a searcher submits a query, Dogpile.com simultaneously submits the query to multiple other Web search engines, then collects the results from each Web search engine, removes duplicates results, and aggregates the remaining results into a combined ranked listing using a proprietary algorithm. Dogpile.com has tabbed indexes for federated searching of *Web*, *Images*, *Audio*, and *Video* content. Dogpile.com also offers query reformulation assistance with query suggestions listed in an "Are You Looking for?" section of the interface. Figure 1 shows the Dogpile.com interface with query box, tabbed indexes, and "Are You Looking for?" feature.

According to Hit Wise,[1] Dogpile.com was the 9th most popular Web search engine in 2005 as measured by number of site visits. ComScore Networks[2] reports that in 2003 Dogpile.com had the industry's highest visitor-to-searcher conversion rate of 83% (i.e., 83% of the visitors to the Dogpile.com site executed a search).

*Data Collection*

For data collection, we logged the records of searcher-system interactions in a transaction log that represents a portion of the searches executed on Dogpile.com, on May 6, 2005. The original general transaction log contained 4,056,374 records. Each record contains seven fields:

- *User Identification*: a user code automatically assigned by the Web server to identify a particular computer
- *Cookie*: an anonymous cookie automatically assigned by the Dogpile.com server to identify unique users on a particular computer.
- *Time of Day*: measured in hours, minutes, and seconds as recorded by the Dogpile.com server.
- *Query Terms*: terms exactly as entered by the given user.
- *Location*: the geographic location of the user's computer as denoted by the Internet Protocol (IP) address of the searcher's computer.
- *Source*: the content collection that the user selects to search (e.g., *Web*, *Images*, *Audio*, or *Video*) with Web being the default (see Figure 1).
- *Feedback*: a binary code denoting whether or not the query was generated by the "Are You Looking for?" query reformulation assistance provided by Dogpile.com (see Figure 1).

*Data Analysis*

We imported the original flat ASCII transaction log file of 4,056,374 records into a relational database and generated a unique identifier for each record. We used four fields (*Time of Day*, *User Identification*, *Cookie*, and *Query*) to locate the initial query and then recreate the chronological series of actions in a session.

*Data preparation.*  We define our terminology similar to that used in other Web transaction log studies (Jansen & Pooch, 2001; Park et al., 2005).

- Term: a series of characters separated by white space or other separator

  Unique term: a term submitted one or more times in the data set

  Term Pair: two terms that occur within the same query

- Query: string of terms submitted by a searcher in a given instance

  Initial query: first query submitted in a session by a given user

  Identical query: a query within a session that is a copy of a previous query within that session

  Repeat query: a query submitted more than once during the data collection period, irrespective of the user

  Query length: the number of terms in the query (Note: this includes traditional stop words.)

- Session: series of queries submitted by a user during one interaction with the Web search engine

  Session length: the number of queries submitted by a searcher during a defined period of interaction with the search engine

  Session duration: the period from the time of the first interaction to the time of the last interaction for a searcher interacting with a search engine

*Removing agent queries.*  We were only interested in queries submitted by humans, and the transaction log contained queries from both human users and agents. There is no known methodology for accurately distinguishing human from nonhuman searchers in a transaction log. Therefore, researchers interested in human sessions usually use a temporal or interaction cutoff (Montgomery & Faloutsos, 2001; Silverstein, Henzinger, Marais, & Moricz, 1999).

We used an interaction cutoff by separating all sessions with 100 or fewer queries into an individual transaction log to be consistent with the approach taken in previous Web searching studies (Jansen & Spink, 2005; Jansen, Spink, & Pederson, 2005b; Spink & Jansen, 2004). This cutoff is substantially greater than the mean search session (Jansen, Spink, & Saracevic, 2000) for human Web searchers. This increased the probability that we were not excluding any human searches. This cutoff probably introduced some agent

or common user terminal sessions; however, we were satisfied that we had included most of the queries submitted primarily by human searchers.

*Removing duplicate queries.*  Transaction log applications usually record result-pages viewing as separate records with an identical user identification and query, but with a new time stamp (i.e., the time of the second visit). This permits the calculation of results-page viewings. It also introduces duplicate records which skew the queries' calculations. To correct for these duplicate queries, we collapsed the transaction log upon user identification, cookie, and query. We calculated the number of identical queries by user, storing these in a separate field within the transaction log. This collapsed transaction log provided us the records by user for analyzing sessions, queries and terms, and pages of results viewed. The un-collapsed transaction log provided us a means to analyze session duration and the number of interactions within a session.

*Term and term co-occurrence analysis.*  We also incorporated a field for the length of the query, measured in terms. We also generated, from the collapsed data set, a table for term data and a table for co-occurrence data. The term table contains fields for a term, the number of times that term occurs in the complete data set, and the probability of occurrence. The co-occurrence table contains fields for term pairs, the number of times that pairs occur within the data set irrespective of order, and the mutual information statistic.

To calculate the mutual information statistic, we followed the procedure outlined by Wang, Berry, and Yang (2003). The mutual information formula measures term association and does not assume mutual independence of the terms within the pair. We calculate the mutual information statistic for all term pairs within the data set. Many times, a relatively low frequency term pair may be strongly associated (i.e., if the two terms always occur together). The mutual information statistic identifies the strength of this association. The mutual information formula used in this research is

$$I(w_1, w_2) = \ln \frac{P(w_1, w_2)}{P(w_1)\, P(w_2)}$$

where $P(w_1)$, $P(w_2)$ are probabilities estimated by relative frequencies of the two words and $P(w_1, w_2)$ is the relative frequency of the word pair; order is not considered. Relative frequencies are observed frequencies ($F$) normalized by the number of queries:

$$P(w_1) = \frac{F_1}{Q'}; P(w_1) = \frac{F_2}{Q'}; P(w_1, w_2) = \frac{F_{12}}{Q'}$$

The frequency of both term occurrence and of term pairs is defined as the occurrence of the term or term pair within the set of queries. However, since a one-term query cannot

have a term pair, the set of queries for the frequency base differs. The number of queries for the terms is the number of nonduplicate queries in the data set. The number of queries for term pairs is defined as

$$Q' = \sum_{n}^{m} (2n - 3)Q_n$$

where $Q_n$ is the number of queries with $n$ words ($n > 1$), and $m$ is the maximum query length. So, queries of length one have no pairs. Queries of length two have one pair. Queries of length three have three possible pairs. Queries of length four have five possible pairs. This continues up to the queries of maximum length in the data set. The above formula for queries of term pairs ($Q'$) accounts for this term pairing.

*Transaction log structure.* The processed transaction log database now contains four tables (un-collapsed data set for temporal analysis, collapsed data set for session and query analysis, terms, and term co-occurrence). We analyzed the data collected to investigate our first two research questions. We conducted the analysis using a variety of layered queries and Visual Basic for Applications scripts.

*Query topic analysis.* We qualitatively analyzed a random sample of 2,500 queries from the 2005 data set, into 11 non-mutually exclusive general topic categories developed by Spink, Jansen, Wolfram, and Saracevic (2002). Two independent evaluators manually classified each of the queries independently. The evaluators then met and resolved discrepancies. This analysis addressed research question number three.

## Results

*Research Question 1: What Are the Characteristics of Search Interactions on the Dogpile.com Metasearch Engine?*

*Overall results.* We present the aggregate results for the analysis in Table 1 as an overview of the findings. There were 2,465,145 interactions during the data collection period. Of these interactions, there were 1,523,793 queries submitted by 534,507 users (identified by unique IP address and cookie) containing 4,250,656 total terms. There were 298,796 unique terms in the 1,523,793 queries. Most of the users (84%) came from the USA. The mean query length was 2.79 terms and nearly fifty percent of queries contained three or more terms. Session length was also relatively lengthy, with a mean of 2.85 queries per user. More than 46% of users modified their queries and 29.4% of the sessions contained three or more queries.

Nearly 10% of the queries in the data set were repeat queries submitted by 10.8% of the searchers. The 898,393 unique queries represent 58.96% of the 1,523,793 total queries. The remaining 473,987 queries were queries to multiple data sources. In 1,052,554 (69.07%) queries, the

TABLE 1. Aggregate statistics from the Dogpile.com transaction log.

| | | |
|---|---|---|
| Sessions | 534,507 | |
| Queries | 1,523,793 | |
| Terms | | |
| Unique | 298,796 | 7.03% |
| Total | 4,250,656 | |
| Location (USA) | 1,282,691 | 84.1% |
| Mean terms per query | 2.79 sd = 1,54 | |
| Terms per query | | |
| 1 term | 281,639 | 18.5% |
| 2 terms | 491,002 | 32.2% |
| 3+ terms | 751,152 | 49.2% |
| Mean queries per user | 2.85, SD = 4.43 | |
| Users modifying queries | 246,276 | 46.08% |
| Repeat Queries (queries submitted more than once by two or more searchers) | 151,413 (by 57,651 searchers) | 9.9% |
| Unique Queries (queries submitted only once in the entire data set) | 898,393 | 58.9% |
| Queries Generated Via Feedback | 128,126 | 8.4% |
| Session size | | |
| 1 query | 288,231 | 53.9% |
| 2 queries | 88,875 | 16.6% |
| 3+ queries | 157,401 | 29.4% |
| Results Pages Viewed Per Query | | |
| 1 page | 1,052,554 | 69.07% |
| 2 pages | 253,718 | 16.6% |
| 3+ pages | 217,521 | 14.2% |
| Mean Results Pages Viewed Per Query | 1.67, SD = 1.84 | |
| Boolean Queries | 33,403 | 2.1% |
| Other Query Syntax | 116,905 | 7.6% |
| Terms not repeated in data set (172,488 terms; 57.7% of the unique terms) | 172,488 | 4.06% |
| Use of 100 most frequently occurring terms (100 terms; 0.03% of the unique terms) | 752,994 | 17.7% |
| Use of other 126,208 Terms (126,208 terms; 42.24% of the unique terms) | 3,325,174 | 78.2% |
| Unique Term Pairs (occurrences of terms pairs within queries from the entire data set) | 2,209,777 | |

searcher viewed only the first results page. There were a very small percentage of Boolean queries (2.19%) and queries containing advanced query syntax (7.6%), namely syntax for phrase searching. Of the total terms, 4.06% of the terms were used only once in the data set, representing 57.7% of the unique terms. The top 100 most frequently used terms accounted for 17.71% of the total terms. There were 2,209,777 term pairs.

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

TABLE 2. Occurrences and percentages of session length in number of queries.

| Session Length in Number of Queries | Occurrences | Percent |
|---|---|---|
| 1 | 288,231 | 53.9% |
| 2 | 88,875 | 16.6% |
| 3 | 47,664 | 8.9% |
| 4 | 29,345 | 5.4% |
| 5 | 19,655 | 3.6% |
| 6 | 13,325 | 2.4% |
| 7 | 9,549 | 1.7% |
| 8 | 7,169 | 1.3% |
| 9 | 5,497 | 1.0% |
| 10 | 4,130 | 0.8% |
| >10 | 21,067 | 3.9% |
| Total | 534,507 | 100.0% |

*Sessions*

**Session length.** Table 2 shows the session length data. More than 79% of the sessions were three or less queries. This finding is similar to other analyses of Web search engines trends. For example, Spink, Jansen, et al. (2002) reported short sessions during Web searches. Jansen and Spink (2005), in their analysis of European searching, noted a similar inclination. Koshman, Spink, and Jansen (2006) found that one in five Vivisimo users entered only two terms during their session. Also, one in ten (10%) Vivisimo users entered only three terms during their session and three in ten (30%) Vivisimo users entered more than three terms in their session.

**Geographical location of users.** Based on the IP address of the user computer, we logged the geographical location of the searcher. The results are presented in Table 3. This table shows that the top four geographical locations for searchers

TABLE 3. Geographical location of searcher based on computer IP address.

| Location | Occurrences | Percent |
|---|---|---|
| USA | 1,282,691 | 84.10% |
| Great Britain | 66,095 | 4.34% |
| Canada | 64,998 | 4.20% |
| Australia | 31,699 | 2.08% |
| Germany | 5,831 | 0.38% |
| India | 4,542 | 0.30% |
| New Zealand | 4,061 | 0.27% |
| South Africa | 3,611 | 0.24% |
| Ireland | 2,850 | 0.19% |
| France | 2,705 | 0.18% |
| Mexico | 2,647 | 0.17% |
| Japan | 2,607 | 0.17% |
| Netherlands | 2,584 | 0.17% |
| Singapore | 2,537 | 0.17% |
| Malaysia | 2,224 | 0.15% |
| Philippines | 2,175 | 0.14% |
| Italy | 2,139 | 0.14% |
| United Arab Emirates | 1,562 | 0.10% |
| Sweden | 1,451 | 0.10% |
| All others | 34,784 | 2.20% |
| Total | 1,523,793 | 100.00% |

TABLE 4. Query lengths.

| Query Length | Occurrences | Percent |
|---|---|---|
| 1 | 281,639 | 18.5% |
| 2 | 491,002 | 32.2% |
| 3 | 373,003 | 24.5% |
| 4 | 193,633 | 12.7% |
| 5 | 95,334 | 6.3% |
| 6 | 45,368 | 3.0% |
| 7 | 22,155 | 1.5% |
| 8 | 11,500 | 0.8% |
| 9 | 5,757 | 0.4% |
| 10 | 2,890 | 0.2% |
| 11 | 1,124 | 0.1% |
| 12 | 311 | 0.0% |
| 13 | 61 | 0.0% |
| 14 | 9 | 0.0% |
| 15 | 3 | 0.0% |
| 18 | 1 | 0.0% |
| 24 | 1 | 0.0% |
| 25 | 2 | 0.0% |
| Total | 1,523,793 | 100.0% |

are predominantly English-speaking countries, representing more than 95% of system users. We could locate no published reports of the geographic locations of Web search engine users. However, the high use of English language queries has been reported in prior research (Jansen & Spink, 2005).

*Queries*

**Query length.** Table 4 presents the length of queries in number of terms. The maximum query length was 25 terms. However, 75% of the queries were three or less terms. After three terms, there is a sharp decline in the frequency of occurrences, dropping to a minimal percentage after five terms per query. The number of one-term queries is notably lower than has been reported elsewhere (Cacheda & Viña, 2001; Spink, Özmutlu, Özmutlu, & Jansen, 2002). Koshman et al. (2006) found that the highest percentage (29.4% and 30%) of Vivisimo queries contained one or two terms, and approximately 72% of the queries contained one to three terms. Searchers on this metasearch engine may be submitting longer queries. However, other published temporal analyses (Jansen & Spink, 2006; Spink, Jansen, et al., 2002) have reported query length moving slowly upwards. Therefore, this difference may be due to the fact that the log files were compiled at a later date.

*Use of system reformulation assistance.* Table 5 displays the number and percentage of queries generated by the searcher when using the system reformulation assistance, which is an area of limited study. Table 5 shows that 8.4% of the queries were generated using the reformation assistance. This is also higher than has been reported elsewhere. Jansen, Spink, and Saracevic (1999) report the use of relevance feedback on the Excite search engine to be approximately 4%. Koshman et al. (2006) report approximately 2% interaction with cluster

TABLE 5. Use of reformulation.

| Reformulation | Occurrences | Percent |
|---|---|---|
| No | 1,395,667 | 91.6% |
| Yes | 128,126 | 8.4% |
| Total | 1,523,793 | 100.0% |

assistance on the Vivisimo search engine. Anick (2003), however, reports a 14% usage of a query reformulation assistance feature (or at least a reformulation that contained an assistance term) on the AltaVista search engine. So, searchers on the Dogpile.com metasearch seem to be generally more receptive to system assistance than the typical Web searcher.

*Use of content collections.* Table 6 displays the most frequent queries in the two data sets according to the use of the various content collections. Table 6 shows that the Web accounts for more than 71% of the query submissions, and next comes images with 19% of queries submitted. The Web was the default. In related research, Özmutlu, Spink, and Özmutlu (2003) examined the impact of multimedia interface buttons on the proportion of multimedia queries in the general query population, and contrasted Web multimedia and nonmultimedia search queries. The researchers state that the use of radio buttons had decreased the multimedia searches in the general collection. Jansen, Spink, and Pedersen (2005a) examined the use of federated content collections on the AltaVista search engine. The researchers report some differences in session and query length based on content collections. Koshman et al. (2006) report that nearly 88% of Vivisimo searchers used the default content collection.

*Top queries.* Table 7 displays the top or most frequent queries. The top queries represent a fairly wide spectrum of possible search topics including celebrities (*lohan pics*, *paris hilton*, *50 cent*), entertainment (*music lyrics*, *american idol*, *playstation 2 cheats*), navigation (*google*, *yahoo*, *mapquest*), current events (*tony blair*), and commerce (*used cars*). This is similar to what was found for users of other Web search engines (Spink, Jansen, et al., 2002). Koshman et al. (2006) found that the most frequently used Vivisimo terms used were *download*, *new*, *software*, *windows*, and *sex*. Beitzel, Jensen, Chowdhury, Grossman, and Frieder (2004) show that there is a variation in topics by hour of the day.

TABLE 6. Use of content collection.

| Source | Occurrences | Percent |
|---|---|---|
| Web | 1,085,573 | 71.2% |
| Images | 290,571 | 19.1% |
| Audio | 95,118 | 6.2% |
| Video | 48,057 | 3.1% |
| News | 4,474 | 0.3% |
| Total | 1,523,793 | 100.0% |

TABLE 7. Top 20 queries with frequency of occurrence and percentage of queries.

| | Query | Occurrences | Percent |
|---|---|---|---|
| 1 | lohan pics | 3153 | 0.2069% |
| 2 | music lyrics | 2464 | 0.1617% |
| 3 | american idol | 1675 | 0.1099% |
| 4 | games | 1278 | 0.0839% |
| 5 | poetry | 1192 | 0.0782% |
| 6 | funny jokes | 1074 | 0.0705% |
| 7 | paris hilton | 1021 | 0.0670% |
| 8 | google | 792 | 0.0520% |
| 9 | yahoo | 710 | 0.0466% |
| 10 | sex | 674 | 0.0442% |
| 11 | ebay | 645 | 0.0423% |
| 12 | tony blair | 642 | 0.0421% |
| 13 | playstation 2 cheats | 639 | 0.0419% |
| 14 | mapquest | 619 | 0.0406% |
| 15 | games cheat | 574 | 0.0377% |
| 16 | food | 560 | 0.0368% |
| 17 | 50 cent | 555 | 0.0364% |
| 18 | iq tests | 541 | 0.0355% |
| 19 | maps | 535 | 0.0351% |
| 20 | used cars | 533 | 0.0350% |
| | | 19,876 | 1.3000% |

There also appears to be a continued use of search engines not to search for information but as a short cut for navigation. Web searchers appear to submit the name of a particular Web site to the search engine and just click on the uniform resource locator (URL) in the results page rather than type the URL in the address box of the browser or locate a bookmark, favorite, or short cut. If the Web page's URL appears in the search engine's first page of results, this method requires less effort than other methods of accessing a particular URL.

*Page results viewed.* Table 8 displays the occurrences and percentage of result listings viewed by query. As noted in prior research, approximately 85% of searchers view only the first or second page of the results listings. Studies also show that searchers typically view only a handful of Web documents

TABLE 8. Results pages viewed.

| Result Pages Viewed | Occurrences | Percent |
|---|---|---|
| 1 | 1,052,554 | 69.1% |
| 2 | 253,718 | 16.6% |
| 3 | 104,233 | 6.8% |
| 4 | 49,458 | 3.2% |
| 5 | 23,483 | 1.5% |
| 6 | 14,609 | 0.9% |
| 7 | 7,519 | 0.4% |
| 8 | 5,194 | 0.3% |
| 9 | 3,288 | 0.2% |
| 10 | 2,402 | 0.1% |
| >10 | 7,335 | 0.4% |
| Total | 1,523,793 | 100.0% |

TABLE 9. Top occurring terms and frequencies.

| | Term | Occurrences | Percent |
|---|---|---|---|
| 1 | of | 60902 | 1.4% |
| 2 | the | 50871 | 1.2% |
| 3 | in | 40197 | 0.9% |
| 4 | and | 34154 | 0.8% |
| 5 | free | 24348 | 0.6% |
| 6 | for | 24161 | 0.6% |
| 7 | a | 23049 | 0.5% |
| 8 | to | 21264 | 0.5% |
| 9 | girls | 13755 | 0.3% |
| 10 | sex | 13418 | 0.3% |
| 11 | on | 12651 | 0.3% |
| 12 | how | 10146 | 0.2% |
| 13 | nude | 9279 | 0.2% |
| 14 | lyrics | 9181 | 0.2% |
| 15 | music | 9067 | 0.2% |
| 16 | new | 9056 | 0.2% |
| 17 | pictures | 8915 | 0.2% |
| 18 | mp3 | 8496 | 0.2% |
| 19 | what | 8460 | 0.2% |
| 20 | is | 7999 | 0.2% |
| 21 | pics | 7439 | 0.2% |
| 22 | school | 7165 | 0.2% |
| 23 | day | 7137 | 0.2% |
| 24 | teen | 6918 | 0.2% |
| 25 | girl | 6836 | 0.2% |
| 26 | i | 6607 | 0.2% |
| 27 | with | 6475 | 0.2% |
| 28 | county | 6455 | 0.2% |
| 29 | black | 6267 | 0.1% |
| 30 | american | 6253 | 0.1% |
| 31 | hot | 6189 | 0.1% |
| 32 | porn | 6154 | 0.1% |
| 33 | you | 6002 | 0.1% |
| 34 | games | 5993 | 0.1% |
| 35 | women | 5901 | 0.1% |
| 36 | my | 5657 | 0.1% |
| 37 | online | 5584 | 0.1% |
| 38 | naked | 5361 | 0.1% |
| 39 | home | 5338 | 0.1% |
| 40 | city | 5126 | 0.1% |
| 41 | video | 5094 | 0.1% |
| 42 | state | 4891 | 0.1% |
| 43 | map | 4806 | 0.1% |
| 44 | history | 4746 | 0.1% |
| 45 | big | 4697 | 0.1% |

(Jansen & Spink, 2003). The percentage of AltaVista searchers viewing only one page of results was 72% in 2002 (Jansen, Spink & Pedersen, 2005a, 2005b). As with users of other Web search engines (Cacheda & Viña, 2001; Hölscher & Strube, 2000; Jansen & Spink, 2005; Spink, Jansen, et al., 2002), Dogpile.com users appear to have a low tolerance for reviewing large numbers of results. However, this may indicate that the Dogpile.com search engine is returning relevant results.

*Terms.* We present a term analysis in Table 9. Table 9 shows that of the top ten terms, there are seven that are common in natural language queries (i.e., *of, the, in, and, for, a, to*). Some patterns present themselves from the term

level of analysis. First, even the most frequently occurring terms represent a small percentage of overall term usage. The most frequently used content term (*free*) accounted only for approximately 0.6% of all term usage. Second, the occurrence of sexual terms (*sex, nude, porn*) was low. Third, there were a significant variety of terms, indicating diverse information needs of Dogpile.com users. This would indicate the user population has similar information interests as users on other Web search engines.

**Term co-occurrence.** A term co-occurrence (Leydesdorff, 1989) is sometimes more helpful in determining the specific usage of a term intended by a searcher. Table 10 shows the term co-occurrences in a correlation matrix fashion. Of the

TABLE 10. Frequency of term co-occurrence for top 25 terms.

| | Term | Term | Occurrences | Percent | Mutual Information Statistic |
|---|---|---|---|---|---|
| 1 | of | the | 14,753 | 0.00670 | 0.96 |
| 2 | the | in | 7,300 | 0.00331 | 0.68 |
| 3 | how | to | 6,089 | 0.00276 | 2.74 |
| 4 | of | a | 3,818 | 0.00173 | 0.40 |
| 5 | what | is | 3,701 | 0.00168 | 3.41 |
| 6 | to | a | 3,419 | 0.00155 | 1.35 |
| 7 | lohan | pics | 3,389 | 0.00154 | 4.00 |
| 8 | pictures | of | 3,334 | 0.00151 | 1.22 |
| 9 | for | sale | 3,300 | 0.00150 | 2.92 |
| 10 | the | the | 3,163 | 0.00144 | −0.40 |
| 11 | and | the | 3,019 | 0.00137 | −0.04 |
| 12 | of | in | 2,856 | 0.00130 | −0.44 |
| 13 | what | the | 2,778 | 0.00126 | 1.27 |
| 14 | new | york | 2,770 | 0.00126 | 4.06 |
| 15 | the | on | 2,659 | 0.00121 | 0.82 |
| 16 | music | lyrics | 2,652 | 0.00120 | 2.87 |
| 17 | how | a | 2,641 | 0.00120 | 1.83 |
| 18 | the | is | 2,583 | 0.00117 | 1.25 |
| 19 | to | the | 2,452 | 0.00111 | 0.22 |
| 20 | high | school | 2,355 | 0.00107 | 3.80 |
| 21 | and | and | 2,328 | 0.00106 | 0.09 |
| 22 | real | estate | 2,318 | 0.00105 | 4.90 |
| 23 | and | of | 2,234 | 0.00101 | −0.52 |
| 24 | history | of | 2,184 | 0.00099 | 1.43 |
| 25 | paris | hilton | 2,136 | 0.00097 | **5.03** |
| 26 | in | a | 2,121 | 0.00096 | 0.23 |
| 27 | american | idol | 2,036 | 0.00092 | 4.42 |
| 28 | for | the | 1,952 | 0.00089 | −0.13 |
| 29 | mothers | day | 1,950 | 0.00089 | 4.17 |
| 30 | star | wars | 1,897 | 0.00086 | 4.85 |
| 31 | map | of | 1,878 | 0.00085 | 1.26 |
| 32 | mother's | day | 1,638 | 0.00074 | 4.29 |
| 33 | for | in | 1,555 | 0.00071 | −0.13 |
| 34 | is | a | 1,535 | 0.00070 | 1.52 |
| 35 | 50 | cent | 1,410 | 0.00064 | **5.63** |
| 36 | what | a | 1,408 | 0.00064 | 1.38 |
| 37 | what | of | 1,401 | 0.00064 | 0.40 |
| 38 | for | a | 1,391 | 0.00063 | 0.32 |
| 39 | a | the | 1,389 | 0.00063 | −0.43 |
| 40 | free | porn | 1,388 | 0.00063 | 1.63 |
| 41 | how | do | 1,343 | 0.00061 | 2.89 |
| 42 | las | vegas | 1,316 | 0.00060 | **5.77** |
| 43 | of | of | 1,310 | 0.00059 | −1.64 |
| 44 | to | in | 1,267 | 0.00058 | −0.20 |
| 45 | free | sex | 1,219 | 0.00055 | 0.72 |

top 50 term pairs, at least 40 are pairs that one would see in natural language queries. In calculating the mutual information statistics for the data set, there was range of more than 45 between the maximum degree of association (max = 13.2) and the minimum degree of association (min = −31.9). The mean association of term pairs in the data set was −5.3 with a standard deviation of 11.9. Therefore, term pairs with associations of approximately 5 or more would indicate a high degree of association for this data set. Table 10 shows that most term pairs with high degrees of association are proper names. Note also that many of these term pairs have low degrees of association as indicated by the mutual information statistic.

In a study of Excite users, Wolfram (1999) notes high clustering of several term pairs around entertainment, which we do not see in this analysis. Silverstein, Henzinger, Marais, and Moricz (1999) report the co-occurrence of the top 10,000 terms from approximately 313,000,000 million queries. Silverstein et al. (1999) also report highly correlated phrases. Table 11 presents the term pairs that had more than 100 occurrences ranked in descending order by their mutual information statistic. We applied a frequency of 100 since term pairs with low frequency will have high degrees of association but due to their low frequency of occurrence do not have as much system impact as lower associated pairs. Again, there is a high occurrence of proper names and phrases.

TABLE 11. Term pairs with frequency greater than 100 and the highest mutual information statistic.

| | Occurrences | Term | Term | Mutual Information Statistic |
|---|---|---|---|---|
| 1 | 101 | coca | cola | 8.14 |
| 2 | 121 | kung | fu | 7.97 |
| 3 | 131 | sum | 41 | 7.89 |
| 4 | 125 | shania | twain | 7.85 |
| 5 | 116 | vida | guerra | 7.84 |
| 6 | 130 | totally | spies | 7.83 |
| 7 | 151 | mortal | kombat | 7.82 |
| 8 | 130 | foo | fighters | 7.79 |
| 9 | 155 | hong | kong | 7.78 |
| 10 | 203 | elisha | cuthbert | 7.69 |
| 11 | 131 | darth | vader | 7.60 |
| 12 | 101 | peanut | butter | 7.59 |
| 13 | 132 | sporting | goods | 7.57 |
| 14 | 112 | gloria | velez | 7.54 |
| 15 | 131 | disco | inferno | 7.53 |
| 16 | 121 | romeo | juliet | 7.40 |
| 17 | 102 | roller | coaster | 7.37 |
| 18 | 211 | avril | lavigne | 7.35 |
| 19 | 206 | eyed | peas | 7.34 |
| 20 | 334 | blink | 182 | 7.29 |
| 21 | 141 | napoleon | dynamite | 7.29 |
| 22 | 110 | yu | gi | 7.25 |
| 23 | 124 | cameron | diaz | 7.25 |
| 24 | 198 | papa | roach | 7.24 |
| 25 | 280 | trish | stratus | 7.22 |
| 26 | 131 | goo | dolls | 7.20 |
| 27 | 334 | ying | yang | 7.19 |
| 28 | 171 | kenny | chesney | 7.19 |
| 29 | 227 | snoop | dogg | 7.17 |
| 30 | 117 | wells | fargo | 7.16 |

TABLE 12. Occurrences and percentage of session duration.

| Session Duration | Occurrences | Percent |
|---|---|---|
| < 1 minute | 302,653 | 56.6% |
| 1 minute to < 5 minutes | 83,236 | 15.5% |
| 5 to < 10 minutes | 36,347 | 6.8% |
| 10 to < 15 minutes | 19,806 | 3.7% |
| 15 to < 30 minutes | 27,210 | 5.1% |
| 30 to < 60 minutes | 18,441 | 3.4% |
| 1 to < 2 hours | 14,236 | 2.6% |
| 2 to < 3 hours | 8,262 | 1.5% |
| 3 to < 4 hours | 5,901 | 1.1% |
| > 4 hours | 18,415 | 3.4% |
| Total | 534,507 | 100.0% |

## Research Question 2: What Are the Temporal Characteristics of Metasearching on Dogpile.com?

*Session duration.* Table 12 presents the session durations for the data set. We measured the session duration from the time the first query was submitted until the user departed the search engine for the last time (i.e., did not return). Silverstein et al. (1999) assigned a temporal cutoff of 5 minutes as the maximum session duration. We defined the sessions as the period from the first interaction with Dogpile.com until the last interaction, as recorded in the transaction log. With this definition of search duration, we measured the total user time on Dogpile.com during this period and the time spent viewing the first and all subsequent results and all subsequent Web documents, except the final document. This final viewing time is not available because the server records the time stamp.

Naturally, the time between visits from the Web document to Dogpile.com may not have been entirely spent viewing the Web document or interacting with the search. However, this may not be a major concern as revealed from the results in Table 12, showing a large percentage of very short session durations. The mean session duration was 26 minutes and 32 seconds, with a standard deviation of 1 hour, 36 minutes and 25 seconds. The maximum session was just under 24 hours (23:57:51), and the minimum session was 0 seconds (i.e., the user submitted one query and performed no other interaction with the search engine).

Jansen, Spink, and Pederson (2005b) found that for AltaVista searches, the mean session duration was 58 minutes and 10 seconds, with a standard deviation of three hours, 34 minutes, and 12 seconds. Fully 81% of the sessions were less than 15 minutes and nearly 72% of the sessions were less than five minutes. This is substantially shorter than earlier reported research on Web session length (Cyber Atlas, 2002; He, Göker, & Harper, 2002), although He et al. (2002) predicted that the mean Web session would be approximately 15 minutes based on analysis of a Reuter transaction log. Hargittai (2002) found similar results in a lab study. The percentage of sessions of five minutes or less was nearly three times that reported for AlltheWeb searchers (26%) (Jansen & Spink, 2005). Surprisingly, the

TABLE 13. Query topic categories.

| Rank | (2,500 Dogpile.com Queries) | Number | Percent |
|---|---|---|---|
| 1 | Commerce, travel, employment, or economy | 761 | 30.4% |
| 2 | People, places, or things | 402 | 16.0% |
| 3 | Unknown or other | 331 | 13.2% |
| 4 | Health or sciences | 224 | 8.9% |
| 5 | Entertainment or recreation | 177 | 7.0% |
| 6 | Computers or Internet | 144 | 5.7% |
| 7 | Education or humanities | 141 | 5.6% |
| 8 | Society, culture, ethnicity, or religion | 119 | 4.7% |
| 9 | Sex or pornography | 97 | 3.8% |
| 10 | Government or legal | 90 | 3.6% |
| 11 | Arts | 14 | 0.5% |
| Total | | 2,500 | 100.0% |

Dogpile.com sessions are even shorter. With the short session durations and the ComScore Networks[3] reports of Dogpile.com's visitor-to-searcher conversion rate of 83%, these factors would seem to indicate that the engine is returning relevant results to the user-submitted queries.

*Research Question 3: What Are the Topical Characteristics of Searches on the Dogpile.com Metasearch Engine?*

*Topical query classification.* We qualitatively analyzed a random sample of 2,500 queries, assigning these queries into one of 11 general topic categories developed by Spink, Jansen, et al. (2002). We chose this number of queries and the categories to be consistent with prior work. Two independent evaluators manually classified each of the queries independently. The independent evaluators were graduate students. The evaluators then met and resolved discrepancies by discussion. Table 13 displays the topic query analysis results. Commerce-related queries were the most frequently occurring (30.4%), followed by people, places and things, and unknown queries (indiscernible or non-English). At 3.8%, sexual and pornographic queries represented a very low proportion of all queries. Recently, Koshman et al. (2006) found that one in five queries submitted to Vivisimo related to commerce, travel, employment, or the economy. In the Vivisimo study, some one in five queries were indiscernible or non-English. This represents a sizable proportion of all queries. Also from the Vivisimo study, one in seven queries was related to people, places, or things. These queries include personal names or the names of locations.

Combined with the evidence from the term and term cooccurrence results, this qualitative analysis extends survey data that the Web is now a major source of information for most people (Cole, Suman, Schramm, Lunn, & Aquino, 2003; Fox, 2002). The move toward the use of the Web as an economic resource and tool (Lawrence & Giles, 1999; Spink, Jansen, et al., 2002) indicates that people use the Web for an increasing variety of information tasks (Fox, 2002; National Telecommunications and Information Administration, 2002).

---

[3] comScore, 2005. http://www.comscore.com/press/release.asp?press=325.

The findings also support the continued drop in sex and pornography as a major topic for search engine users (Jansen & Spink, 2006; Spink, Jansen, et al., 2002).

## Discussion

Overall, the level of user interaction is higher on Dogpile.com than indicated by results from other Web search engines, but Dogpile.com users spend less time on the Web search engine. The highest percentage of queries contained two to three terms and the majority of queries contained one, two, or three terms. The percentage of two-term queries is higher than that found in earlier research, which showed that some 60% of Web searchers used one or two terms (Jansen & Spink, 2006; Spink & Jansen, 2004). Hence, these results differ from results reported in previous studies. When the underlying system is changed (i.e., from a non-metasearch engine to a metasearch engine), we would expect some differences in user behavior. The change may also be related to changes in the period of data collection.

Repeated queries were widely distributed and the top repeated queries represented only one half of one percent of the total number of queries. Web information is heterogeneous and the nature of repeated query entries reflects the span of topic coverage. Term frequency data also showed a wide distribution of information interests. The most frequently used terms were linked to computing, universities, travel, music, and downloading music. The qualitative analysis on the first data set showed that the "Commerce, Travel, Employment or Economy" category contained the highest percentage of queries (30.4%). The topic classification results are similar to the topic distribution found in other U.S.-based Web search engines (Jansen & Spink, 2006; Jansen et al., 2005b; Koshman et al., 2006). These studies show that the most popular topics for Web searches are commerce, travel, employment, or economy related, followed by people, places, or things. For the European Web search engine AlltheWeb.com users, people, places, and things made up the largest category of search topic (Jansen & Spink, 2005).

The highest percentage of sessions contained between two and three queries. However, session duration was generally less than one minute and almost half of the sessions (53%) fell into this category. The search session times were shorter than reported in earlier Web studies, but they were similar to the recent Vivisimo study (Koshman et al., 2006), which analyzed searching on a clustering search engine.

Our study is limited to data from one major commercial Web metasearch engine and possibly does not represent the queries submitted by the broader Web searching population. However, this research was targeted at investigating the behavior of searchers using a metasearch engine. We do not have information about the demographic characteristics of the users who submitted queries, so we must infer their characteristics from the demographics of Web searchers as a whole. Another limitation is that we collected the data on only one day; therefore, the data may not

be representative of overall usage. However, Jansen and Spink (2006) have shown that Web searching is fairly consistent across days and search engines, with the exception of term usage.

## Conclusion and Further Research

Our findings provide important insights into the current state of Web searching and Web usage for developers of search engines, Web sites designers, and e-commerce sites. This represents the first major study of human interaction with a major commercial metasearch Web search engine. There are several avenues for future research. Certainly, we need more analysis of further Dogpile.com search data taken over a longer period. We conducted further analysis of the data set to examine query reformulation, and repeat, successive, and multitasking sessions. Also, in the query classification section, we used categories from prior work (Spink, Jansen, et al., 2002) for consistency. However, a more granular classification may be in order.

Metasearching on the Web is a concept that has been well researched on the system side, but there is little research into how searchers use such systems. The results from this research have the potential to influence positively the future design of information systems and metasearch engines. Researchers should continue to examine and track Web search trends and characteristics. This can be done using either transaction log analysis or lab studies in order to assess future behavior and identify future user needs.

## Acknowledgment

## References

Anick, P. (2003). Using terminological feedback for web search refinement: A log-based study. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 88–95). New York: ACM.

Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., & Frieder, O. (2004). Hourly analysis of a very large topically categorized Web query log. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 321–328). New York: ACM.

Buzikashvili, N. (2002). Metasearch: Properties of common document distributions. In U. Karagiannis & U. Reimer (Eds.), Lecture notes on computer science (Vol. 2569, pp. 226–231). Berlin: Springer.

Cacheda, F., & Viña, Á. (2001). Experiences retrieving information in the World Wide Web. Proceedings of the Sixth IEEE Symposium on Computers and Communications (pp. 72–79). Hammamet, Tunisia: IEEE.

Chignell, M.H., Gwizdka, J., & Bodner, R.C. (1999). Discriminating meta-search: A framework for evaluation. Information Processing and Management, 35(3), 337–362.

Cole, J.I., Suman, M., Schramm, P., Lunn, R., & Aquino, J.S. (2003, February 1). The UCLA Internet report surveying the digital future: Year three. Retrieved February 1, 2003, from http://www.ccp.ucla.edu/pdf/UCLA-Internet-Report-Year-Three.pdf

Cyber Atlas. (2002, January 1). November 2002 Internet usage stats. Retrieved January 1, 2003, from http://cyberatlas.internet.com/big_picture/traffic_patterns/article/0,,5931_1560881,00.html

Ding, W., & Marchionini, G. (1996). A comparative study of Web search service performance. In S. Hardin (Ed.), Proceedings of the 59th Annual Meeting of the American Society for Information Science (pp. 136–142). Baltimore, MD: ASIS.

Dreilinger, D., & Howe, A. E. (1997). Experiences with selecting search engines using metasearch. ACM Transactions on Information Systems, 15(3), 195–222.

Fox, S. (2002, July). Search engines (Pew Internet and American Life Project Report). Retrieved October 15, 2002, from http://www.pewinternet.org/pdfs/PIP_Search_Engine_Data.pdf

Gauch, S., Wang, G., & Gomez, M. (1996). Profusion: Intelligent fusion from multiple, distributed search engines. The Journal of Universal Computer Science, 2(9), 637–649.

Gerhart, S. L. (2004, January 5). Do Web search engines suppress controversy? First Monday, 9(1). Retrieved January 1, 2007, from http://www.firstmonday.org/issues/issue9_1/gerhart/index.html

Hargittai, E. (2002). Beyond logs and surveys: In-depth measures of people's Web use skills. Journal of the American Society for Information Science and Technology, 53(14), 1239–1244.

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.

Howe, A.E., & Dreilinger, D. (1997). SAVVYSEARCH: A metasearch engine that learns which search engines to query. AI Magazine, 18(2), 19–25.

Introna, L., & Nissenbaum, H. (2000). Defining the Web: The politics of search engines. IEEE Computer, 33(1), 54–62.

Jansen, B.J., & Pooch, U. (2001). Web user studies: A review and framework for future work. Journal of the American Society for Information Science and Technology, 52(3), 235–246.

Jansen, B.J., & Spink, A. (2003). An analysis of Web documents retrieved and viewed. In H. Arabnia & Y. Mun (Eds.), Proceedings of the Fourth International Conference on Internet Computing (pp. 65–69). Las Vegas: CSREA.

Jansen, B.J., & Spink, A. (2005). An analysis of Web searching by european alltheweb.com users. Information Processing & Management, 41(2), 361–381.

Jansen, B.J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing & Management, 42(1), 248–263.

Jansen, B.J., Spink, A., & Pedersen, J. (2005a). The effect of specialized multimedia collections on Web searching. Journal of Web Engineering, 3(3/4), 182–199.

Jansen, B.J., Spink, A., & Pederson, J. (2005b). Trend analysis of AltaVista Web searching. Journal of the American Society for Information Science and Technology, 56(6), 559–570.

Jansen, B.J., Spink, A., & Saracevic, T. (1999). The use of relevance feedback on the Web: Implications for Web IR system design. Proceedings of the 1999 World Conference of the World Wide Web and Internet, 550–555.

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 & Management, 36(2), 207–227.

Koshman, S., Spink, A., & Jansen, B.J., (2006). Web searching on the Vivisimo search engine. Journal of the American Society for Information Science and Technology, 57(14), 1875–1887.

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(4), 209–223.

Meng, W., Yu, C., & Lui, K.-L. (2002). Building efficient and effective meta-search engines. ACM Computing Surveys, 34(1), 48–89.

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

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.

Özmutlu, H.C., Spink, A., & Özmutlu, S. (2003). Multimedia Web searching trends: 1997–2001. Information Processing & Management, 39(4), 611–621.

Oztekin, B.U., Karypis, G., & Kumar, V. (2002). Expert agreement and content based reranking in a meta search environment using Mearf. Proceedings of the 11th International Conference on the World Wide Web (pp. 333–345). New York: ACM.

Park, S., Bae, H., & Lee, J. (2005). End user searching: A Web log analysis of NAVER, a Korean Web search engine. Library and Information Science Research, 27(2), 203–221.

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., & Jansen, B.J. (2004). Web search: Public searching of the Web. New York: Kluwer.

Spink, A., Jansen, B.J., Blakely, C., & Koshman, S. (2006). A study of results overlap and uniqueness among major Web search engines. Information Processing & Management, 42(5), 1379–1391.

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., Lawrence, S., & Giles, L. (2000). Inquirus Web meta-search tool: A user evaluation. In Proceedings of WebNet 2000 (pp. 819–820). Chesapeake, VA: AACE.

Spink, A., Özmutlu, S., Özmutlu, H.C., & Jansen, B.J. (2002). U.S. versus European Web searching trends. SIGIR Forum, 32(1), 30–37.

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.

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.