

Web Searching on the Vivisimo Search Engine

Sherry Koshman

School of Information Sciences, University of Pittsburgh, 135 N. Bellefield, Pittsburgh, PA 15260.

E-mail: {skoshman}@sis.pitt.edu

Amanda Spink

Queensland University of Technology, Gardens Point Campus, 2 George Street, GPO Box 2434, Brisbane, QLD 4000, Australia. E-mail: ah.spink@qut.edu.au

Bernard J. Jansen

School of Information Sciences and Technology, The Pennsylvania State University, 329F IST Building, University Park, PA 16802. E-mail: jjansen@acm.org

The application of clustering to Web search engine technology is a novel approach that offers structure to the information deluge often faced by Web searchers. Clustering methods have been well studied in research labs; however, real user searching with clustering systems in operational Web environments is not well understood. This article reports on results from a transaction log analysis of Vivisimo.com, which is a Web meta-search engine that dynamically clusters users' search results. A transaction log analysis was conducted on 2-week's worth of data collected from March 28 to April 4 and April 25 to May 2, 2004, representing 100% of site traffic during these periods and 2,029,734 queries overall. The results show that the highest percentage of queries contained two terms. The highest percentage of search sessions contained one query and was less than 1 minute in duration. Almost half of user interactions with clusters consisted of displaying a cluster's result set, and a small percentage of interactions showed cluster tree expansion. Findings show that 11.1% of search sessions were multitasking searches, and there are a broad variety of search topics in multitasking search sessions. Other searching interactions and statistics on repeat users of the search engine are reported. These results provide insights into search characteristics with a cluster-based Web search engine and extend research into Web searching trends.

Introduction

Cumbersome search results lists generated by traditional Web search engines is a well recognized problem in Web information retrieval (IR). Providing the user with a means of viewing groups of similar search results potentially

enhances Web search effectiveness; however, there has been little research into Web searchers' interactions with clustered search engine results. The user assumptions underlying document set clustering include the ability of the clusters to aggregate similar documents based on topic, type, chronology, or other criteria supported by the system. The cluster labels generated are to be indicative of document content, and there is an expectation that the process should be automated and efficient. The notion is to provide an overview of the collection to enable users to make decisions regarding their selections from a document set (Hearst, 1999).

Clustering search output has been incorporated in Web search engines such as Mooter (<http://mooter.com>). Mooter's consumer searching provides clusters using algorithms based on psychological modeling of user information seeking. The clusters are presented in a "starburst" visualization for the user to select. The Web search engine iBoogie (<http://www.iboogie.com/>) is a meta-searching clustering environment that uses the Clusterizer technology (www.clusterizer.com). Each cluster is given a label based on document content. Both the labels and the clusters are based on term-extraction techniques that use linguistic and statistical calculations on text summaries returned by the search. Similar to Vivisimo, Clusterizer technology generates clusters in real time and uses a cluster tree to structure the cluster labels.

As new clustering technologies are introduced on the Web for IR, it is increasingly important to examine how users interact with these systems to evaluate user-system interactions and improve system performance. This investigation is timely since its objective is to better understand the nature of user interaction with Vivisimo, a cluster-based Web search engine. A quantitative and qualitative transaction log analysis was conducted to examine user queries

Received July 16, 2005; accepted November 9, 2005

© 2006 Wiley Periodicals, Inc. • Published online 2 October 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20408

presented to the system to determine characteristics of term distribution, queries, and sessions. Sessions on the Vivisimo Web search engine were analyzed further for topic distribution and to examine the degree of multitasking search by typical Web searchers. The overall goal of this research is to understand user interaction with a clustering Web search engine and further extend the line of user interaction research in Web IR.

Related Studies

Clustering Methods

Clustering methods have been well investigated in the IR environment (Baeza-Yates & Ribeiro-Neto, 1999; Korfhage, 1997). Web-based clustering work in research labs includes adapting algorithms (Wang & Kitsregawa, 2002; Zamir & Etzioni, 1998; Zeng, He, Chen, Ma, & Ma, 2004) and user testing (Chen & Dumais, 2000; Hearst & Pedersen, 1996). This investigation uses transaction log analysis to study Vivisimo user search characteristics. Similarly, a transaction log analysis was used to evaluate user interaction with Grouper, a clustering interface for Web search engine results (Zamir & Etzioni, 1999). The findings from the Grouper study showed that users tended to examine more clusters than hypothesized. The logs also were analyzed to compare Grouper with a traditional text-based interface, Husky-Search, to determine the number of documents clicked on by the users. Their results showed that users followed more multiple documents using the Grouper clustering interface and more single documents using HuskySearch.

Transaction log analysis provides a useful method of determining Web searcher trends. Xie and O'Hallaron (2002) analyzed 1-month's worth of Vivisimo and Excite data for repetition of query terms to determine caching methods. Their findings showed that query repetition is common among users and that caching offers the potential to improve the efficiency of processing queries. Transaction log analyses offer an unobtrusive method to study user interactions with traditional Web search engines (Jansen, Spink, & Saracevic, 2000; Spink & Jansen, 2004; Spink, Jansen, & Saracevic, 2001). These studies have shown that Web users typically enter two terms per query, one query per session, and few use Boolean operators (Wolfram, Spink, Jansen, & Saracevic, 2001).

Multitasking

A user's single session with a Web search engine may often consist of seeking information on single or multiple topics. Recent studies have examined multitasking searching on the Excite and AlltheWeb.com Web search engines. Spink, Ozmutlu, and Ozmutlu (2002) and Spink, Park, Jansen, and Pedersen (2006) showed that IR and Web searches often include multiple topics during a single search session, which we refer to as a *multitasking search*; however, limited knowledge exists on the characteristics and patterns

of multitasking searches. Spink et al. (2006) found that 81% of two-query sessions on the AltaVista search engine included multiple topics, 91.3% of three or more query sessions included multiple topics, a broad variety of topics in multitasking search sessions, and three or more query sessions sometimes contained frequent topic changes. Further research is needed to study the prevalence and characteristics of multitasking by other Web search engine users.

This study extends the large-scale research of Web search queries to a cluster-based search engine. This analysis also provides insight into user interaction with clusters, which currently is unexplored and not well understood.

Research Questions

The research questions addressed by this study include:

- What are the search characteristics of Vivisimo users, including the session length, query length, and use of query operators?
- What is the distribution of terms, query topics, sources, and languages used?
- What is the extent of cluster expansion by users?
- What is the distribution of clusters among post-initial search records?
- What is the occurrence of repeated searching by Vivisimo users?
- What is the occurrence of multitasking sessions by Vivisimo users?

This analysis reveals patterns of searchers using a cluster-based Web search engine.

Research Design

Vivisimo.com

The Vivisimo (<http://vivisimo.com>) interface contains a dialog box for inputting queries and supports Boolean and exact phrase matching. The default search source is the Web, and a drop-down menu provides options for additional source selection (e.g., CBC, CNN, and Wisenut). Searches can be limited by domain or host name, link content, Web page, or Uniform Resource Locator (URL) information. Vivisimo offers an "Advanced" search form containing options for source and language selection, defining the number and display of search results, deciding how links should be opened, and whether the content filter is applied. After a user submits a query, Vivisimo presents the clusters using a tree metaphor, which is similar to that used for viewing folders in Windows Explorer. The clusters appear on the left side of the page, and the results pages are featured on the right of the main search page (Figure 1).

Unlike typical Web search engines, which present lists of search output, Vivisimo's clustering feature creates dynamic postsearch categories in a meta-searching environment. Users can click on cluster labels to retrieve results pages. Clusters can be expanded by clicking on the plus sign to reveal subclusters, and the cluster tree may be elongated by

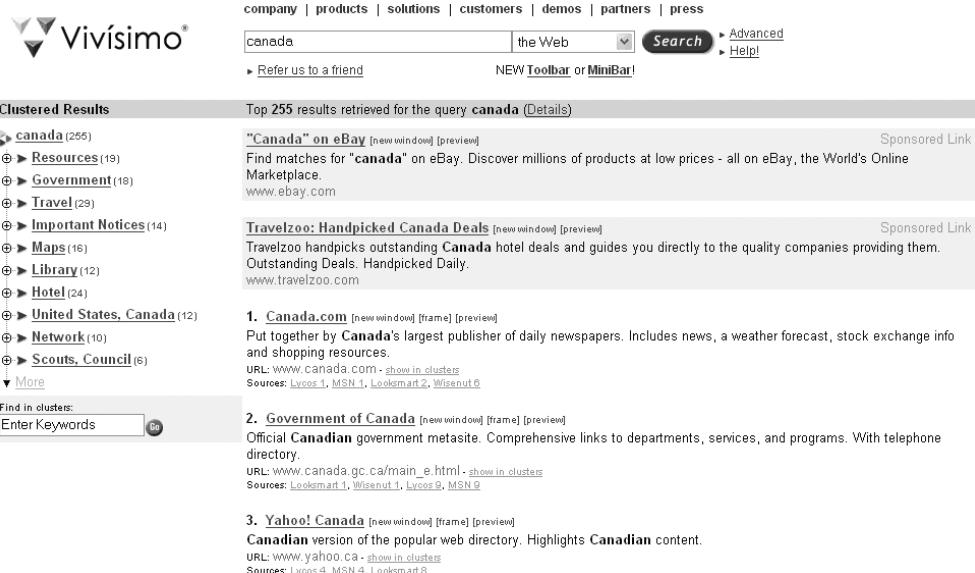


FIG. 1. Vivisimo.com interface.

clicking on the “More” option. Search terms can be entered in the “Find in clusters” search box to search the clusters.

The results pages are initially displayed as a result of the initial search. Results pages are retrieved when the user clicks on the clusters; additional results pages may be selected at the bottom of the window. Hyperlinks may be accessed for individual items, and Web pages may be previewed and then opened in the results frame or in a new window. An item on the results pages may be identified within the clusters by clicking on the “show in clusters” option next to the item. This highlights the clusters on the tree which contain the item. The “Details” feature shows the number of results for the sources searched.

Data Collection

The Vivisimo transaction log data used for this study represents two 1-week periods: (a) March 28 to April 4, 2004, containing 1,082,431 queries and (b) April 25 to May 2, 2004, containing 927,303 queries. The transaction log recorded 100% of the traffic on the Vivisimo Web site during these periods, and in total, 2,029,734 queries and 386,949 unique Internet Protocol (IP) addresses were recorded.

The records in the transaction log files contain a variety of fields, including *User Identification* (an anonymous user code assigned by the Vivisimo server), *Date* (the calendar day as recorded by the Vivisimo server), *Time of Day* (the clock time as recorded by the Vivisimo server), and *Query Terms* (terms exactly as entered by the given user).

Data Analysis

A relational database was created to import the transaction log, which is a flat ASCII file, and each record was given a unique identifier. The initial query was located on the basis of four fields (i.e., user identification, date, time of day, and query terms). A user session was recreated by structuring a chronological series of actions on a given day. A

term is any series of characters separated by white space or other separator. A *query* is the entire string of terms submitted by a searcher in a given instance of interaction. A *session* is the entire series of queries submitted by a user during one interaction with the Web search engine on a given day. An *identical query* is a query that is a copy of a previous query within the same user session. A *repeat query* is a query submitted more than once, irrespective of the user.

Searches from both human users and agents are contained in the transaction log. This analysis focused on queries submitted only by humans, not by an automated process. There is not an established method to accurately identify human from nonhuman searchers, so most researchers utilizing transaction logs must either overlook it (Cacheda & Viña, 2001) or assign some temporal or interaction cutoff (Montgomery & Faloutsos, 2001; Silverstein, Henzinger, Marais, & Moricz, 1999).

A cutoff approach was used, and sessions with 100 or fewer queries were separated into another transaction log. This cutoff was selected because it is almost 50 times greater than the reported mean search session for human Web searchers and because it assured that human searches were not excluded. The assumption was made that the analysis yielded a subset of the transaction log that contained queries submitted primarily by human searchers. The cutoff may have allowed for some agent or common user terminal sessions; however, it is broad enough in minimizing bias introduced by too low of a cutoff threshold.

When a searcher submits a query to Vivisimo, views a document, and returns to the search engine, the Vivisimo server logs this second visit with the identical user identification and query, but with a new time (i.e., the time of the second visit). Vivisimo assigns a unique code to identify a user’s multiple interactions with the system. This is beneficial information in determining how many of the retrieved results pages the searcher visited from the search engine; however, it also introduces duplicate queries.

The transaction log was collapsed by combining all identical queries submitted by the same user to provide unique queries for analyzing sessions, queries and terms, and pages of results viewed. Uncollapsed sessions were used to obtain an accurate measure of the session duration and the number of results pages visited. The number of identical queries by the same user was recorded in a separate field within the remaining records when the sessions were collapsed.

In addition to the fields for unique identifier and number of identical queries, a field within each record containing the length of the query, measured in terms, was included. Two other tables for the collapsed dataset, one for term data and the other for co-occurrence data, were generated. The term table contains fields for a term and the number of times that a term occurs in the complete dataset. The co-occurrence table contains fields for term pairs and the number of times that a pair occurs within the dataset, irrespective of order.

The database now contains four tables (i.e., uncollapsed dataset, collapsed dataset, terms, and co-occurrence). The data from these four tables were analyzed to investigate our research questions. The analysis was conducted using queries, usually a series of layered queries, Visual Basic for Applications scripts, or a combination of the two. Key fields were extracted from the log file for the clustering analysis of the dataset, and each query was identified by a unique Vivisimo assigned code. A series of

UNIX text-manipulation commands were used to parse and calculate statistics on the clustering data.

Results

Overall Results

Tables 1 and 2 present the aggregate results for both datasets in this analysis.

Terms

Term frequency. The top 30 terms were extracted, removing the terms without context (an, or, de, la, le, etc.), letters, and numbers (Table 3). The term distribution showed 73.3% of the top 30 terms were the same in both datasets. The most frequently used terms used were “download,” “new,” “software,” “windows,” and “sex.” Most of these terms represent a strong user focus on computing issues. Many of the top 30 terms are strongly linked to computing, universities, students, travel, music, and downloading music. Although the term *sex* is a very high-frequency term, sexual or pornographic queries formed less than 5%, or 1 in 20 searches. This is a similar finding to that of users of other Web search engines (Spink, Jansen, Wolfram, & Saracevic, 2002).

TABLE 1. Aggregate analysis for Dataset 1 (3/28–4/04).

Unique IP addresses	193,377				
Total queries	1,082,431				
Terms					
Unique	350,504	10.3%			
Total terms	3,394,989	100%			
Mean terms per query	Mean	Max	Min	Mode	SD
	3.14	51	1	2	3.29
Terms per query	No.	%			
0	8,843	0.9			
1	194,915	20.3			
2	281,995	29.4			
3+	474,869	49.4			
	960,622	100			
	No.	%			
Term pairs Unique	2,072,791	21.9			
Total term pairs	9,479,582				
Users' modifying queries	174,216	62			
Session length	Mean	Max	Min	Mode	SD
	3.87	99	1	1	5.37
	No.	%			
1 query per session	105,550	38			
2 queries per session	53,720	19			
3+ queries per session	120,536	43			
	279,766	100			
Result pages viewed	Mean	Max	Min	Mode	SD
	1	14	1	1	5.37
	No.	%			
Boolean queries	28,590	2.6			
Queries with other operators	217,645	20.1			
Both Boolean and other	246,235	22.7			
Terms not repeated in dataset	190,394	5.6			
100 Most frequently occurring terms	646,161	19			

TABLE 2. Aggregate analysis for Dataset 2 (4/25–5/02).

Unique IP	193,572				
Queries	927,303				
Terms					
Unique	327,440	11.3%			
Total	2,905,592	100%			
	Mean	Max	Min	Mode	SD
Mean terms per query	3.13	85	0	2	3.2
Terms per query	No.	%	Term pairs	No.	%
0	3,251	0.4	Unique	1,907,743	24.3
1	174,338	18.8	Total	7,859,746	
2	278,377	30.0			
3+	471,337	50.8			
	927,303	100			
	No.	%			
Users' modifying queries	160,392	58			
	Mean	Max	Min	Mode	SD
Session size	3.37	99	1	2	4.56
	No.	%			
1 query	115,064	42			
2 queries	54,094	20			
3+ queries	106,298	39			
	275,456	100			
	Mean	Max	Min	Mode	SD
Result pages	1.14	1027	1		1.35
	No.	%			
Boolean queries	23,891	2.6			
	No.	%			
Queries with					
other operators	198,031	21.4			
Both Boolean and other	No.	%			
	221,922	24.0			
	No.	%			
Terms not repeated					
in dataset	183,861	55.9			
	No.	%			
100 Most					
frequently					
occurring terms	547,809	18.9			

Term Co-Occurrence

Although a term analysis is useful, it is sometimes difficult to determine the specific usage of a term intended by a searcher outside the framework of a particular query. In these cases, a term co-occurrence (Leydesdorff, 1989) is more helpful. Tables 4 and 5 show the co-occurrences for the Vivisimo datasets, including the percentage calculated of the top 10 co-occurring term pairs.

Queries

Queries per day. Both datasets showed very similar query distribution patterns during the week (Figure 2). The mean queries per day were 135,304 for Dataset 1 and 115,913 for Dataset 2. Approximately 80% of queries were entered during weekdays, with about 5% fewer queries submitted per day on the weekends.

Query length. Table 6 shows the range of query lengths as defined by the number of terms per query. Zero indicates a null search. In both datasets, the highest percentage (29.4 and

30%) of queries contained two terms. Approximately 72% of the queries in both datasets contained one to three terms. Further analysis of the queries showed that a small percentage (2.6%) in both datasets contained Boolean operators. Approximately 20% of the queries contained other operators.

Most frequent queries. The most frequent queries were calculated for both datasets, and 80% of the top 10 repeated queries are alike for both datasets (Tables 7 and 8). The common terms found in both datasets are shaded in Table 8, which shows that 61.5% of terms were present in both datasets.

Sessions

Sessions per day. Figure 3 shows the percentage of sessions distributed across both datasets. The mean sessions per day for Dataset 1 were 34,971 and 34,432 for Dataset 2.

Session length by query. Table 9 shows that 1 in 3 Vivisimo users entered only one term during their session. Some 1 in 5 Vivisimo users entered only two terms during

TABLE 3. Top 30 term frequency.

Dataset 1 (3/28–4/04)				Dataset 2 (4/25–5/02)			
Rank	Term	Occurrence	%	Term	Occurrence	%	
1	download	7,212	0.212	free	9,364	0.0032	
2	new	4,965	0.146	download	6,695	0.0023	
3	software	4,233	0.125	new	4,495	0.0015	
4	windows	3,892	0.115	2004	4,374	0.0015	
5	sex	3,767	0.111	pictures	3,989	0.0014	
6	school	3,374	0.099	how	3,937	0.0014	
7	history	3,298	0.097	software	3,808	0.0013	
8	online	3,238	0.095	windows	3,649	0.0013	
9	video	3,129	0.092	history	3,076	0.0011	
10	what	2,970	0.087	sex	3,047	0.0010	
11	music	2,786	0.082	school	2,992	0.0010	
12	world	2,661	0.078	online	2,814	0.0010	
13	lyrics	2,654	0.078	music	2,628	0.0009	
14	hotel	2,587	0.076	video	2,594	0.0009	
15	map	2,506	0.074	world	2,446	0.0008	
16	system	2,406	0.071	hotel	2,377	0.0008	
17	home	2,390	0.070	map	2,352	0.0008	
18	mp3	2,386	0.070	uk	2,334	0.0008	
19	linux	2,361	0.070	what	2,322	0.0008	
20	web	2,337	0.069	lyrics	2,291	0.0008	
21	xp	2,301	0.068	with	2,264	0.0008	
22	john	2,296	0.068	american	2,198	0.0008	
23	Canada	2,293	0.068	linux	2,182	0.0008	
24	University	2,279	0.067	university	2,173	0.0007	
25	Test	2,277	0.067	test	2,103	0.0007	
26	American	2,264	0.067	war	2,082	0.0007	
27	Uk	2,219	0.065	canada	2,080	0.0007	
28	Internet	2,181	0.064	home	2,078	0.0007	
29	Management	2,173	0.064	www.ascocco.it	2,066	0.0007	
30	Business	2,149	0.063	mp3	2,065	0.0007	
		87,584	2.580		92,875	0.0319	

TABLE 4. Dataset 1 top 10 term co-occurrences.

Term	Term	Occurrences	%
And	and	5,973	42.3
Free	download	995	7.0
For	the	986	6.9
For	sale	983	6.9
windows	xp	975	6.9
To	in	894	6.3
Britney	spears	851	6.0
What	the	841	5.9
High	school	813	5.7
For	in	785	5.5
		14,096	100.0

their session. In addition, 1 in 10 Vivisimo users entered only three terms during their session, and 3 in 10 Vivisimo users entered more than three terms in their session.

Session duration and interactions. Table 10 shows the aggregate statistics for session duration. Session duration was measured from the time the first query was submitted until the user departed the search engine for the last time (i.e., does not return) on a given day. This definition allows for the measurement of the total user time on the search engine and

TABLE 5. Dataset 2 top 10 term co-occurrences.

Term	Term	Occurrences	%
New	York	1,484	14.4
What	is	1,346	13.1
history	of	1,144	11.1
Of	pictures	985	9.5
Real	estate	940	9.1
For	sale	890	8.6
download	free	883	8.5
High	school	875	8.5
How	a	867	8.4
windows	xp	860	8.3
		10,274	100.0

the time spent viewing the first and all subsequent Web documents, except the final document. The final viewing time is not available since the Web search engine server records the time stamp. A limitation is that the time between visits from the Web document to the search engine may not have been entirely spent viewing the Web document.

The average session duration is 1½ min, and the most frequently occurring duration value in the dataset is less than 1 min for both datasets. Table 11 shows one in two Vivisimo sessions were less than 1 min in duration, 1 in 10 Vivisimo

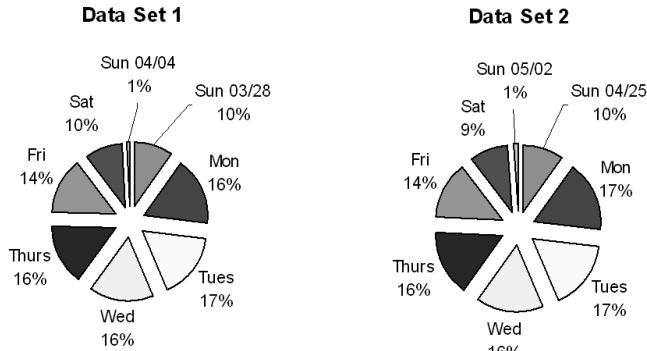


FIG. 2. Queries per day.

TABLE 6. Query length.

Length	Dataset 1 (3/28–4/04)		Dataset 2 (4/25–5/02)	
	Occurrences	%	Occurrences	%
0	8,843	0.9	3,251	0.4
1	194,915	20.3	174,338	18.8
2	281,995	29.4	278,377	30.0
3	216,904	22.6	212,738	22.9
4	120,083	12.5	121,864	13.1
5	61,208	6.4	61,974	6.7
6	29,387	3.1	29,321	3.2
7	14,677	1.5	14,626	1.6
8	7,815	0.8	7,475	0.8
9	5,051	0.5	4,933	0.5
10	2,573	0.3	2,723	0.3
>10	17,171	1.8	15,683	1.7
	960,622	100.0	927,303	100.0

TABLE 7. Dataset 1 top repeat queries.

Rank	Query	Occurrences	%
1	Google	614	0.0542
2	“Mark Twain”	610	0.0303
3	Cloning	596	0.0300
4	Looney Tunes	587	0.0265
5	YAHOO	328	0.0265
6	Sex	325	0.0197
7	Ebay	287	0.0174
8	Science Lesson Plans	287	0.0169
9	“Ralph Nader”	213	0.0165
10	yahoo.com	188	0.0160
11	www.lolitas.com	183	0.0157
12	Test	179	0.0152
13	Chocolate Chips	173	0.0152
14	site de blog fotos de casais de sexo	170	0.0150
15	Dictionary	165	0.0146
16	Games	164	0.0143
17	paris Hilton	162	0.0130
18	Sexo	158	0.0129
19	Backgammon	155	0.0128
20	britney spears pictures	141	0.0128
21	Hotmail	140	0.0126
22	Homenspelados	139	0.0126
23	Porn	139	0.0125
24	britney spears clips	136	0.0124
25	Lyrics	136	0.6138
26	Carolina Dieckman	135	0.0542
27	Vivisimo	134	0.0303
		6,644	0.0300

TABLE 8. Dataset 2 top repeat queries and common terms.

Rank	Query	Occurrences	%
1	“Mark Twain”	688	0.07
2	Looney Tunes	493	0.05
3	Google	488	0.05
4	Cloning	428	0.05
5	Yahoo	273	0.03
6	Ebay	257	0.03
7	Sex	243	0.03
8	paris Hilton	185	0.02
9	Dictionary	141	0.02
10	yahoo.com	135	0.01
11	hotmail	130	0.01
12	spybot	122	0.01
13	“Ralph Nader”	117	0.01
14	porn	115	0.01
15	mapquest	114	0.01
16	fnord	112	0.01
17	games	101	0.01
18	lyrics	97	0.01
19	search engines	97	0.01
20	Kentucky Derby	96	0.01
21	vivisimo	96	0.01
22	hotmail.com	94	0.01
23	google.com	93	0.01
24	www.google.com	91	0.01
25	routenplaner	85	0.01
26	test	85	0.01
		4,891	0.53

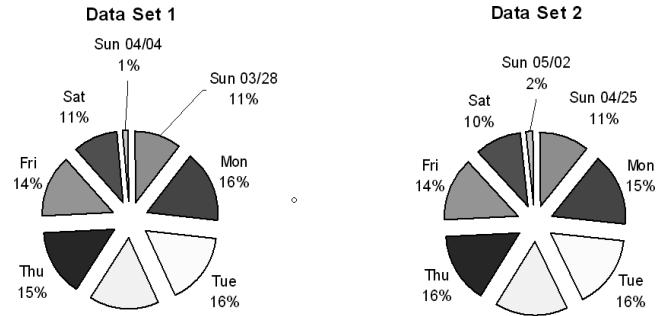


FIG. 3. Sessions per day.

TABLE 9. Session length.

Query Length	Dataset 1 (3/28–4/04)		Dataset 2 (4/25–5/02)	
	Occurrences	%	Occurrences	%
1	105,550	38	115,064	41.8
2	53,720	19	54,094	19.6
3	31,664	11	30,735	11.2
4	21,097	8	19,538	7.1
5	14,782	5	13,322	4.8
6	10,616	4	9,300	3.4
7	7,968	3	6,963	2.5
8	6,069	2	5,180	1.9
9	4,779	2	4,036	1.5
10	3,866	1	3,037	1.1
>10	19,655	7	14,187	5.2
	279,766	100	275,456	100.0

TABLE 10. Aggregate statistics for vivisimo sessions.

	Dataset 1 (3/28–4/04)		Dataset 2 (4/25–5/02)	
	Duration	Interactions	Duration	Interactions
Mean	1:36:35	3.87	1:34:01	3.84
Max	23:59:06	214	23:59:44	1,035
Min	<1s	1	<1s	1
Mode	<1 min	1	<1 min	1
SD	3:48:39	5.4	3:43:49	5.96

TABLE 11. Session duration distribution.

Session duration	Dataset 1 (3/28–4/04)		Dataset 2 (4/25–5/02)	
	Occurrences	%	Occurrences	%
<1 min	125,996	45.04	125,241	45.5
1–5 min	30,001	10.72	30,275	11.0
5–10 min	15,960	5.70	15,592	5.7
10–15 min	9,986	3.57	9,801	3.6
15–30 min	16,965	6.06	16,197	5.9
30–60 min	15,040	5.38	14,412	5.2
1–2 hr	13,424	4.80	13,079	4.7
2–3 hr	8,225	2.94	7,958	2.9
3–4 hr	6,421	2.30	6,345	2.3
>4 hr	37,748	13.49	36,566	13.3
	279,766	100.00	275,466	100.0

sessions were less than 5 min in duration, and 1 in 20 Vivisimo sessions were between 5 and 10 min in duration. Overall, 45% of sessions were less than 1 min in duration for both datasets.

Table 12 shows the results from a random sample of 3,600 sessions containing 4,883 queries that were classified into 11 nonmutually exclusive topic categories developed by Spink, Jansen, et al. (2002). One in five queries submitted to Vivisimo related to commerce, travel, employment, or the economy. This finding relates to the “business” and “management” terms that occur in the top 30 most frequent terms (Table 1). One in five queries were indiscernible or non-English, which represents a sizable proportion of all queries.

TABLE 12. Session topic distribution.

Topics	No. of topics	%
Commerce, Travel, Employment, or Economy	1,025	21
Indiscernible or Non-English	931	19
People, Places, or Things	745	15
Computers or Internet	631	13
Social, Culture, Ethnic, or Religion	458	9
Health or Sciences	275	6
Education or Humanities	221	5
Sex or Pornography	190	4
Performing or Fine Arts	160	3
Government	151	3
Entertainment or Recreation	96	2
	4,883	100

Additionally, one in seven queries were related to people, places, or things. These queries include personal names or the names of locations.

Language and Sources

Table 13 shows that the majority of queries in both datasets did not specify a language (90%); however, the highest foreign-language selection was German (3%). The majority of source requests selected the Web for Vivisimo searches (Table 14). Nine in 10 Vivisimo users did not request a language. The ranking of the German Web source is consistent with the previous language distribution, and may be facilitated by the availability of Vivisimo’s interface in German via a “Deutsch” option on the initial search page.

Clusters

Query Distribution

Clustering data were available for analysis in the second dataset (April 25–May 2, 2004). Figure 4 shows the query distribution with a cutoff of 20 queries. The distribution of queries per IP address shows that the largest percentage of IP addresses exhibited one query occurrence. Small percentages of IP addresses (>.05) contained 21 or more queries. IP addresses that showed above 100 or more query occurrences constituted less than .0015. This query limit was used as a cutoff point to account for the prevalence of network address translation-based firewalls that hide multiple users behind one IP address and the use of dynamic addressing for dial-up connections. A Vivisimo search returns three frame records. The form frame defines the frame set, including tree and list frames. The tree frame represents clusters, and the list frame corresponds to the results pages. The total number of records was 4,219,925. The list records constitute 44% of the total number of records, and the tree records represent 29% of the total.

The higher percentage of list records suggests that more results pages were viewed than were cluster expansions. This finding is consistent with the pattern of cluster usage discussed in the next section. Postquery records were examined to look solely at user interaction with Vivisimo. List records were used approximately three and a half times more than were the tree records. This means that the users clicked on clusters to retrieve results pages. Almost half (48.26%) of the postquery records involved displaying the results pages that come from clicking on a cluster.

Cluster Expansion

Cluster expansion indicates user activity in clicking on the “+” sign next to a cluster label and expanding the tree to reveal subclusters.

Figure 5 shows the distribution and extent to which clusters were expanded. Clusters are most frequently expanded once, representing about 2.3% of the total number of all records (Koshman, Spink, & Jansen, 2005). The maximum number of clusters expanded in a record is 26 (.00014%).

TABLE 13. Distribution of language requests.

	Language	Dataset 1 (3/28–4/04)		Dataset 2 (4/25–5/02)	
		Occurrences	%	Occurrences	%
1	None Specified	977,071	90.27	831,945	89.7
2	All	58,453	5.40	54,672	5.9
3	German	34,417	3.18	33,118	3.6
4	English	2,558	0.24	1,521	0.2
5	Polish	2,135	0.20	1,421	0.2
6	Czech	2,100	0.19	1,208	0.1
7	Spanish	1,945	0.18	754	0.1
8	Italian	1,249	0.12	735	0.1
9	French	669	0.06	608	0.1
10	Dutch	426	0.04	307	0.0
11	Japanese	371	0.03	276	0.0
12	All Others	1,037	0.10	738	0.1
		1,082,431	100.00	927,303	100.0

TABLE 14. Distribution of source requests.

		Dataset 1 (3/28–4/04)		Dataset 2 (4/25–5/02)	
		Occurrences	%	Occurrences	%
1	Web	948,393	87.62	721,925	77.9
2	GermanWeb	73,662	6.81	88,810	9.6
3	MSN, Netscape, Lycos, Looksmart, Overture	8,973	0.83	74,276	8.0
4	English Web	6,547	0.60	6,980	0.8
5	eBay	6,464	0.60	5,765	0.6
6	PubMed	3,997	0.37	3,639	0.4
7	Google, Lycos, Netscape, Looksmart, MSN, Britannica, Dogpile	3,833	0.35	2,926	0.3
8	MSN, Netscape, Lycos, Looksmart, FindWhat	3,609	0.33	2,827	0.3
9	AltaVista, MSN, Netscape, Lycos, Looksmart, FindWhat	2,828	0.26	2,734	0.3
10	AltaVista	2,807	0.26	2,415	0.3
11	MSN	2,304	0.21	1,856	0.2
12	Google, Fast, MSN, BBC	1,889	0.17	1,218	0.1
13	All Other Combinations	17,125	1.58	11,932	1.3
		1,082,431	100	927,303	100.0

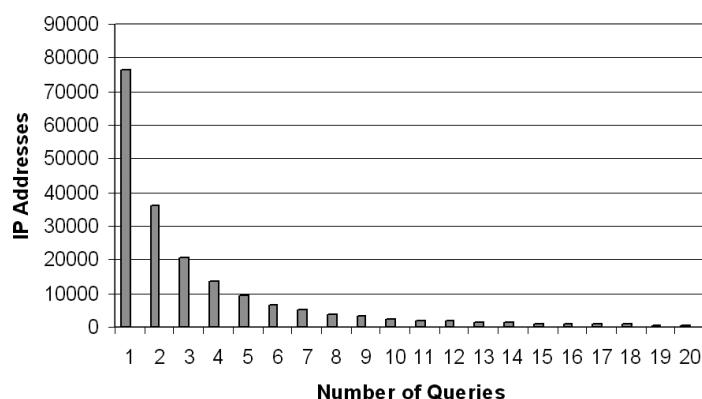


FIG. 4. Query distribution per Internet Protocol (IP) address.

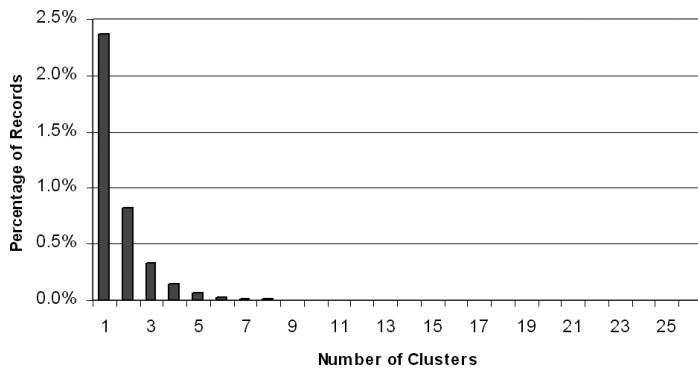


FIG. 5. Distribution of expanded clusters.

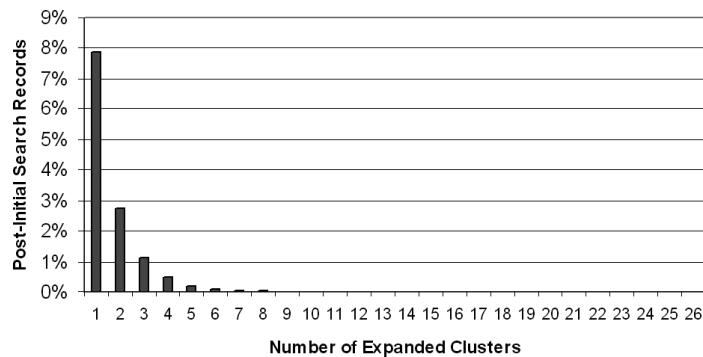


FIG. 6. Postinitial search results cluster distribution.

Cluster Distribution

Clusters were expanded in 12.6% of the records, not including the initial query results in this dataset. This distribution represents user interaction with the system following the presentation of initial search results (Figure 6). This pattern of cluster use may be attributed to users finding what they are searching for via Vivisimo's clustering interface. Clusters appear to be used as anchors or reference points from which users may examine multiple results pages to find information, and represents an efficient use of the clustering feature to organize and retrieve information. Overall, few Vivisimo users manipulated the cluster tree, but almost half of the records showed that postquery activity indicates cluster clicking activity to view the results pages of a cluster.

Repeat Users

The first dataset was analyzed to evaluate the occurrence and characteristics of searchers who return to Vivisimo. Repeat searchers were identified by removing all IP addresses that occurred on only 1 of 8 days. The result was a database of only those users who had visited the search engine on more than 1 day during the 8-day period. Of the 193,377 unique users, the users who made repeat visits to the Vivisimo over the data collection period were identified. There were 40,227 (20.8%) users who visited the Vivisimo search engine two or more times during this period.

Table 15 shows that 68% of the users visited the search engine two or three times, accounting for 50% of the

sessions. Just less than 1% of the users visited the search engine on all 8 days of the collection period; however, the last day of data collection was not a full day (i.e., only slightly more than 4 hr). There were 2.5% of the users who made seven repeated visits to the search engine, so this percentage may be a better indicator of percentage of daily repeat users. A total of 126,616 sessions were generated by 40,227 users. Twenty-one percent of Vivisimo users accounted for 45.2% of the sessions. The mean sessions per day were 15,827, following the overall pattern of the weekdays showing a higher percentage of sessions.

Regarding queries, there were 648,897 queries from 40,227 repeat users. This group of users (21%) accounted for 59.9% of the queries. The mean queries per day were 81,112, with the weekdays again showing a higher percentage of queries.

TABLE 15. Sessions by repeat users of vivisimo.

No. of days user visited search engine	No. of users	%	No. of Sessions during time period	%
2	18,442	45.8	36,884	29.1
3	9,022	22.4	27,066	21.4
4	5,813	14.5	23,252	18.4
5	4,048	10.1	20,240	16.0
6	1,511	3.8	9,066	7.2
7	1,020	2.5	7,140	5.6
8	371	0.9	2,968	2.3
	40,227	100.0	126,616	100.0

TABLE 16. Number of queries by repeat users.

No. of queries	No. of users	% of repeat users
2	3,046	7.6
3	3,424	8.5
4	3,226	8.0
5	2,726	6.8
6	2,632	6.5
7	2,348	5.8
8	2,036	5.1
9	1,728	4.3
10	1,507	3.7
11–20	8,883	22.1
21–30	3,594	8.9
31–50	2,796	7.0
51–100	1,777	4.4
100+	504	1.3
	40,227	100.0
Mean	162	
SD	511	
Max	503	

Table 16 shows a query analysis of how many queries users submitted over the data-collection period. The majority of users (56%) submitted 10 or less queries over the 8-day period. Although the mean was 162 queries for that time period, a small percentage of users (1.3%) skewed the average and the *SD*.

Multitasking

The Vivisimo Web session dataset used in the qualitative analysis is shown in Table 17. For this dataset, of the 2,960 search sessions qualitatively analyzed, we identified a total of 329 multitasking Web search sessions. Overall, 11.1% of Vivisimo users seemed to be conducting multitasking searches. Spink et al. (2006) examined Alta Vista Web searching and found that 81% of two-query sessions included multiple topics and that 91.3% of three or more query sessions included multiple topics.

Table 18 shows the data relating to topic changes during the Vivisimo two-query dataset. For this dataset, there were 728 topic changes in 329 multitasking sessions, yielding a

TABLE 17. Vivisimo session data.

Vivisimo 2004	
Total sessions	2,960
Multitasking sessions	329
% of sessions	11.1%
Total queries	4,000
Queries in multitasking sessions	1,368

TABLE 18. Topic changes in multitasking sessions.

Multitasking Sessions	Vivisimo 2004
Total topic changes	728 (in 329)
Mean topic changes per session	2.2
Means topics per session	2.7
Mean queries per topic	2.0

TABLE 19. Number and percentage of queries in each topic category.

Topic category	No. of queries	%
People/Places/Things	383	27.9
Education	159	11.6
Entertainment	157	11.4
Social	134	9.7
Computing	117	8.5
Shopping	103	7.5
Business	88	6.4
Medical	54	3.9
Science	53	3.8
Sex	50	3.6
Arts/Humanities	39	2.8
Travel	14	1.0
Job	10	0.7
Government	7	0.5
Total	1,368	100

mean of 2.2 topic changes per session. Vivisimo users who engaged in multiple-topic sessions investigated more than one topic per search session. For the Excite users, there were 246 topic changes in the 114 user query sessions, with a mean of 2.2 topic changes per session that also can be interpreted as a mean of 3.2 topics per session.

Spink et al. (2006) examined multitasking during Excite and AlltheWeb.com searching. They showed that the mean queries per multitasking search session were 14.9 for Excite and 14.3 for Alltheweb.com users. The mean queries per session for the entire Excite sample was 10, making Excite multitasking sessions about 50% longer than regular search sessions. The same statistics for the AlltheWeb.com dataset show that the mean queries were 10.3 for the entire sample and 14.3 for multitasking sessions.

The queries in multitasking sessions were categorized with respect to the topics provided in Spink, Ozmutlu, and Ozmultu (2002). The categories used in the study and the number of queries falling into each category for the Vivisimo dataset is provided in Table 19.

Vivisimo multitasking topics were most commonly related to seeking information on people, places, and things, followed by education and entertainment, respectively. These categories form about 40% of the queries in multitasking sessions. Spink et al. (2006) showed that Excite users preferred the categories of hobbies, shopping, and business that form about 47% of all queries in multitasking sessions. The subject categories might reflect topic at the same time. Spink et al. (2006) found that multitasking search sessions included more than three topics per search session. They also found that Alta Vista users multitasking sessions were seeking information on general information, computers, and shopping.

Discussion

Vivisimo Queries and Sessions

The two Vivisimo datasets displayed similar analytical patterns. The findings indicate that higher numbers of queries were presented to the system during weekdays, with

an average of over 100,000 queries per day. The queries were generally brief. The highest percentage of queries (30%) contained two terms, and the majority of queries (72%) contained one, two, or three terms. This is a slightly higher percentage than that found in earlier research showing that 60% of Web searchers used one or two terms (Jansen, Spink, & Pedersen, 2005). Hence, this result is not unique to a clustering environment.

Repeated queries were widely distributed, with the top repeated queries representing only one half of 1% 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. Interestingly, 60% of the top 10 term co-occurrence pairs (including one inverted term pair) was found in previous term co-occurrence analyses of the search engine datasets. Some elements of search strings are common over long periods of time.

The terms most frequently used were linked to computing, universities, travel, music, and downloading music. The qualitative analysis on the first dataset showed that the "Commerce, Travel, Employment or Economy" category contained the highest percentage of queries (21%). The topic classification results are similar to the topic distribution found in other U.S.-based Web search engines (Spink, Jansen, et al., 2002). At the time of their study, Spink, Jansen, et al. (2002) showed that the most popular topics for Web searches were 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 was the largest category of search topic. For Vivisimo users, the percentage of indiscernible or non-English queries was 19%.

Higher percentages of search sessions occurred on weekdays. The highest percentage of the sessions (38–41.8%) contained one query. The session duration mode value was less than 1 min; almost half of the sessions (45%) fell into this category. The search session times were less than those shown in previous research. In an AltaVista study, 72% of sessions were less than 5 min, and 82% were less than 15 min (Jansen et al., 2005). Over three fourths of the searches used the "Web" as the information source. The majority of searches (90%) did not specify a language preference.

Clustering Sessions

Regarding user interaction with clusters, the higher percentage of list records shows that more results pages were viewed than were cluster expansions. This means that the users clicked on clusters to retrieve results pages or that they clicked on the "more" option to retrieve more results pages. The record analysis shows that almost half of the postsearch user interactions involve clicking on Vivisimo clusters; however, expanding the cluster tree is infrequently used.

Repeat searchers are an important component of the Vivisimo user population. This group is a small, but noticeable, percentage of Vivisimo users since they account for a

sizable percentage of all sessions and queries. These repeat users account for nearly a majority of all sessions and account for a majority of all queries. The repeat users mirror that of the overall population of Vivisimo users in terms of patterns of submissions and short sessions.

Multitasking Sessions

Not as many Vivisimo users, as compared to Alta Vista users, engaged in multitasking sessions. The prevalence of multitasking varied across different Web search engines; however, we have found multitasking users in every Web query set examined. Therefore, multitasking over various topics is a fairly common behavior among many people who are using Web search engines to find information, but the level of multitasking Web search is evolving.

Many users' Web searching behaviors, including typical Web search sessions of two or three queries, often include a need to seek and search on more than one topic concurrently due to the complex nature of work or living tasks. People also multitask in work environments. For example, they use the microprocessor while they talk on the telephone. People have many tasks at hand at the same time, including information-seeking tasks. In these cases, a person may pool topics together and interact with the Web on more than one related or unrelated topic.

Implications

From a search engine design perspective, the general use of clusters in synthesizing Web search results may represent a more efficient method to the display and rendering of Web search results. The brevity of search session times with Vivisimo implies that using clusters offers a more direct approach to finding the information that users are seeking. Interaction with the clusters point toward a similar pattern. Cluster clicking activity was well represented in the data, and the initial cluster display may have been sufficient to resolve the information need, thus reducing the user's need to extend the cluster tree to find more cluster labels. This supposition needs to be tested further in usability studies. A portion of users repeatedly return to Vivisimo as an information source, which indicates that a cluster-based approach supports successive searching.

From an interface design perspective, the direct manipulation of clusters works well in the handling of search results. Clicking on cluster labels is better utilized than elongating the cluster tree. Cluster label selection may be more intuitive since it is analogous to clicking on hyperlinks or file folders whereas the cluster tree expansion option is not immediately visible to the end user. However, the Vivisimo's search entry dialogue box, which is similar to traditional search engine technologies, did not impact the nature of query construction. The general search characteristics of user searches were similar to those of nonclustering or traditional search engines. Further, the large percentage of default language and source selections indicates that options

immediately viewable in the drop-down menu were predominantly used for searching.

Earlier cluster-related user studies support these findings. In the Scatter/Gather study, Hearst and Pedersen (1996) showed that documents similar to each other are more relevant than are dissimilar ones. Their users interacted well with clusters, and initial results showed that they selected clusters containing the most relevant documents. Similarly, Chen and Dumais (2000) found that the user interface which categorized Web search results measured better with users than did the traditional list interface.

Conclusions and Further Research

The transaction log analysis showed that while clustering is a new feature in search engine technology that is actively used by Vivisimo searchers, search characteristics such as query length, session length, and topic distribution remain relatively stable in comparison to earlier Web search research.

Future research will include the examination of cluster usage on a per-query basis and investigation of user interaction with Vivisimo to specifically determine patterns in cluster label selection, depth of clusters selected, and use of "find in clusters" feature. Usability studies need to be conducted to determine the patterns of real-time usage of a cluster-based interface.

Acknowledgments

We thank Vivisimo for providing the data for this analysis, without which we could not have conducted this research. We also acknowledge graduate students Minsoo Park and Christopher Field, University of Pittsburgh, for their assistance with the qualitative analysis.

References

- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval. New York: ACM Press.
- Cacheda, F., & Viña, Á. (2001). Experiences retrieving information in the World Wide Web. Paper presented at the 6th IEEE Symposium on Computers and Communications, Hammamet, Tunisia (pp. 72–79).
- Chen, H., & Dumais, S. (2000). Bringing order to the Web: Automatically categorizing search results. Paper presented at the SIGCHI conference on Human Factors in Computing, Hague, The Netherlands (pp. 145–152). New York: ACM Press.
- Hearst, M. (1999). User interfaces and visualization. In R. Baeza-Yates & B. Ribeiro-Neto (Eds.), *Modern information retrieval* (pp. 257–323). New York: ACM Press.
- Hearst, M., & Pedersen, J. (1996). Reexamining the cluster hypothesis: Scatter/gather on retrieval results. Paper presented at the 19th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland (pp. 76–84). New York: ACM Press.
- Jansen, B.J., Spink, A., & Pedersen, J. (2005). 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. (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.
- Korfhage, R.R. (1997). *Information storage and retrieval*. New York: Wiley.
- Koshman, S., Spink, A., & Jansen, B.J. (2005). Using clusters on the Vivisimo Web search engine. Paper presented at the HCI 2005 HCI International Conference, Las Vegas, NV. Mahwah, NJ: Erlbaum.
- Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy*, 18, 209–223.
- Montgomery, A., & Faloutsos, C. (2001). Identifying Web browsing trends and patterns. *IEEE Computer*, 34(7), 94–95.
- Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large Web search engine query log. *ACM SIGIR Forum*, 33(1), 6–12.
- Spink, A., & Jansen, B.J. (2004). Web search: Public searching of the Web. Dordrecht, The Netherlands: Kluwer Academic.
- Spink, A., Jansen, B.J., & Saracevic, T. (2001). Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226–234.
- Spink, A., Jansen, B.J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *IEEE Computer*, 35(3), 107–111.
- Spink, A., Ozmutlu, H.C., & Ozmutlu, S. (2002). Multitasking information seeking and searching processes. *Journal of the American Society for Information Science and Technology*, 53(8), 639–652.
- Spink, A., Park, M., Jansen, B.J., & Pedersen, J. (2006). Multitasking during Web search sessions. *Information Processing and Management*, 42(1), 264–275.
- Wang, Y., & Kitsregawa, M. (2002). Evaluating contents-link coupled Web page clustering for Web search results. Paper presented at the 11th International Conference on Information and Knowledge Management, McLean, VA (pp. 499–506). New York: ACM Press.
- Wolfram, D., Spink, A., Jansen, B.J., & Saracevic, T. (2001). Vox Populi: The public searching of the Web. *Journal of the American Society for Information Science and Technology*, 52(12), 1073–1074.
- Xie, Y., & O'Hallaron, D. (2002). Locality in search engine queries and its implications for caching. Paper presented at the IEEE Infocom Conference, New York.
- Zamir, O., & Etzioni, O. (1998). Web document clustering: A feasibility demonstration. Paper presented at the 21st annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia (pp. 46–54). New York: ACM Press.
- Zamir, O., & Etzioni, O. (1999). Grouper: A dynamic clustering interface for Web search results. Retrieved January 10, 2004, from <http://www.cs.washington.edu/research/projects/WebWare1/etzioni/www/papers/www8.pdf>
- Zeng, H., He, Q., Chen, Z., Ma, W., & Ma, J. (2004). Learning to cluster Web results. Paper presented at the SIGIR Conference on Research and Development in Information Retrieval, Sheffield, England (pp. 210–216). New York: ACM Press.