The current issue and full text archive of this journal is available at
www.emeraldinsight.com/1468-4527.htm

Searching multimedia federated content web collections

Amanda Spink
Faculty of Information Technology, Queensland University of Technology, Brisbane, Australia, and
Bernard J. Jansen
College of Information Sciences and Technology, Pennsylvania State University, Pennsylvania, USA

Refereed article received
5 May 2005
Approved for publication
18 June 2005

Abstract

Purpose – The purpose of this research is to show that federated content collections are important for providing access to multiple content repositories, including image, video, audio and Web sites.

Design/methodology/approach – This paper presents findings from an analysis of differences in users' Web searching patterns as they access various federated content collections. A dataset of 4,056,374 records submitted to the Dogpile.com Web meta-search engine were analysed. An analysis was conducted of search session length, query length, number of results pages viewed, use of systems' assistance and the frequency of repeat queries.

Findings – Overall, users entered two to three terms per query and examined only the first pages of results. However, findings include differences in users' access patterns to various content collections. Web, news and audio queries were longer sessions but shorter queries. More users seeking images and videos sought systems assistance.

Originality/value – This is a large-scale original study using data from a commercial Web search engine. The paper provides a valuable comparison of different types of search – text v. audio, image, etc.

Keywords Multimedia, Content management, Worldwide web, Search engines

Paper type Research paper

Introduction

Web searching is an everyday skill used by many people worldwide. Previous studies show that overall, Web searches are short, and people view few results pages (Spink and Jansen, 2004). Many Web users are now accessing federated content collections via the Web. Federated search, also known as distributed search, is a growing area of information retrieval research (Callan, 2000; Si and Callan, 2005). A federated content collection is a content organising scheme involving multiple repositories of content, instead of a central repository. These individual repositories typically have their own storage, indexing, and retrieval algorithms. Major Web search engines typically offer tabbed interfaces that permit users to search multiple federated content collections, such as Web documents, images, audio, and video files.

Si and Callan (2005) identify three key research problems for federated or distributed Web searching development: first, resource description, or, creating information about the contents of each resource; second, resource selection is the resource set selected for the search; third, results merging is the effective merging of returned results.

The authors thank Infospace, Inc for providing the Web search engine data set.

Online Information Review
Vol. 30 No. 5, 2006
pp. 485-495

1468-4527
DOI 10.1108/14684520610706389

However, few studies have specifically examined users' access to multiple federated search collections via Web search engines. There are several ongoing projects seeking to build federations of learning content and content repositories (EdNA, 2005; Globe, 2005), but there are few user studies. Due to the limited studies in this research area, the examination of how people use federated content collections is an important area of Web research for the future improvement of these systems. Our paper provides results from a large-scale study of user access to federated content collections, via the Dogpile.com Web search engine. The study builds upon previous studies we have conducted, exploring various aspects of Web searching.

The next section of the paper outlines the related studies, followed by the research design and key results from our study.

Related studies

An increasing body of studies is examining various aspects of Web searching (Spink and Jansen, 2004) and federated or distributed Web searching. In the area of federated repositories, Rehak *et al.* (2005) developed a model and infrastructure for federated learning content repositories. Becarevic and Roantree (2004) studied federated multimedia database systems and Martin *et al.* (2002), discuss federated rights management. Recent studies have explored the design and development of federated or distributed Web search (Avrahami *et al.*, 2006; Callan, 2000; Khoossainov and Kushmerick, 2004; Si and Callan, 2005; Xu and Callan, 1998). Powell and Fox (1998) describe a scalable system for searching heterogeneous multilingual collections on the Web.

However, there have been limited studies investigating access to federated collections via Web search. Ozmutlu *et al.* (2003) examined the impact of multimedia interface buttons on the Excite search engine, by investigating multimedia queries in the general query population, prior to, and after the introduction of radio buttons, to search various collections. The researchers reported that the use of radio buttons had decreased the multimedia searches in the general collection. However, the researchers did not examine queries to any of the federated collections.

Jansen *et al.* (2005) and Jansen and Spink (2005) compare Web searching characteristics among Web, image, audio, and video content collections on the AltaVista search engine. The researchers report that of the four types of searching, image searching was the most multifaceted task, and audio the least complex. The mean terms per query for image searching was notably larger (four terms) than the other categories of searching, which were less than three terms. The session lengths for image searches were longer than any other type of searching, and session lengths for Boolean usage by image searches, was 28 per cent.

Aside from these, there have been few large-scale studies of users' access to federated content collections via Web search engines. One problem that limits the conduct of such large-scale research, is having access to data from Web search engines. Researchers need access to large-scale Web transaction logs, to research more effectively in this area.

Research goals

The research goal of our study was to examine differences in Dogpile.com searching across various federated content collections. Specific goals were to examine search differences in various federated content collections, including:

- session length – number of queries per session;
- query length – number of terms per query;
- number of results pages viewed;
- use of system assistance; and
- repeat queries – if the user entered the same query more than once.

To address these research goals, we examined a subset of queries submitted by searchers on Dogpile.com to gain insight into the nature of their searching behaviour.

Research design

Dogpile.com

Infospace, a market leader in the meta-search engine business, owns Dogpile.com (www.Dogpile.com). Dogpile.com is the only meta-search engine during the study period to incorporate the indices of the four leading Web search indices into its search results (i.e. Ask Jeeves, Google, MSN, and Yahoo!). When a searcher submits a query, Dogpile.com simultaneously submits the query to multiple other Web search engines, collecting the results from each Web search engine, removing duplicate results, and aggregating the remaining results into a combined ranked listing, using a proprietary algorithm.

Dogpile.com has tabbed indexes for searching the Web, Images, Audio, and Video. Dogpile.com also offers query reformulation assistance with query suggestions listed in an "Are You Looking for?" section of the interface. Hitwise (2005) reports that Dogpile.com was the 9th most popular Web search engine in 2005 as measured by number of site visits. comScore Networks (2005) states that in 2005 Dogpile.com had the industry's highest visitor-to-searcher conversion rate of 83 per cent (i.e. 83 per cent of the visitors to the Dogpile.com site executed a search).

Data collection

We recorded the records of searcher–system interactions in a transaction log that represents a portion of the searches executed on Dogpile.com 6 May 2005. The original general transaction log contained 4,056,374 records. Each record contains six fields:

(1) *User identification:* an anonymous user code automatically assigned by the Dogpile.com server to identify a particular computer.

(2) *Cookie:* anonymous cookie automatically assigned by the Dogpile.com server to identify unique users on a particular computer.

(3) *Time of day:* measured in hours, minutes, and seconds as recorded by the Dogpile.com

(4) *Query terms:* terms exactly as entered by the given user.

(5) *Source:* the content collection that the user selects to search (e.g., Web, Images, Audio, or Video) with Web being the default.

(6) *Feedback:* a binary code denoting whether or not the query was generated by the "Are You Looking for?" query reformulation assistance.

Data analysis

We imported into a relational database the original flat ASCII transaction log file of 4,056,374 records. We 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.

We define our terminology similar to that used in other Web transaction log studies (Jansen and Pooch, 2000; Park *et al.*, 2005; Spink and Jansen, 2004):

(1) *Term:* series of characters separated by white space or other separator.
(2) *Query:* string of terms submitted by a searcher in a given instance.
(3) *Repeat query:* query submitted more than once during the data collection period, irrespective of the user.
(4) *Session:* series of queries submitted by a user during one interaction with the Web search engine.
(5) *Session length*: number of queries submitted by a searcher during a defined period of interaction with the search engine.

We also removed all agent and duplicate queries.

Results

This paper provides results from a large-scale study of user access to federated content collections via the Dogpile.com Web search engine.

Federated Dogpile.com content collections

Table I shows the usage of each of the five federated Dogpile.com content collections (Web, Images, Audio, Video and News).

Table I shows that the Web was the most popular content collection, with more than 71 per cent of all searches being executed against this content collection. Images were the second most popular content collection, followed by the audio, video and news collection.

Session length

Table II shows the session length (i.e. number of queries) for queries to the diverse federated content collections.

Most users included between one and three queries in their federated content search sessions. Some 50 per cent of users across the various federated content collections included only one query in their search session. News sessions were shorter and included fewer queries. Audio sessions were longer, but with fewer queries per session.

Table I.
Use of the Dogpile.com content collections

Source	Occurrences	%
Web	1,085,573	71.2
Images	290,571	19.07
Audio	95,118	6.2
Video	48,057	3.1
News	4,474	0.29
Total	1,523,793	100

Session length	Web	%	Images	%	Audio	%	Video	%	News	%
1	258,204	56.8	40,026	51.1	10,404	42.9	6,741	47.4	1,757	69.5
2	77,884	17.1	12,420	15.8	4,004	16.5	2,476	17.4	373	14.7
3	40,793	8.9	6,328	8.08	2,391	9.8	1,357	9.5	179	7.08
4	24,067	5.2	4,087	5.2	1,609	6.6	881	6.2	74	2.9
5	15,341	3.3	2,772	3.5	1,161	4.7	590	4.1	47	1.8
6	10,015	2.2	2,065	2.6	839	3.4	412	2.9	26	1.02
7	6,839	1.5	1,601	2.04	667	2.7	327	2.3	21	0.8
8	4,942	1.08	1,202	1.5	478	1.9	229	1.6	13	0.5
9	3,618	0.7	1,070	1.3	413	1.7	205	1.4	11	0.4
10	2,581	0.5	805	1.02	346	1.4	158	1.1	3	0.1
11	1,873	0.4	718	0.9	251	1.03	115	0.8	7	0.2
12	1,506	0.3	596	0.71	202	0.8	105	0.7	3	0.1
13	1,130	0.2	498	0.6	217	0.8	73	0.5	3	0.1
14	881	0.1	438	0.5	152	0.67	62	0.4	3	0.1
15	729	0.1	358	0.4	118	0.4	61	0.4	1	0.04
16	609	0.1	338	0.4	111	0.4	50	0.3	2	0.07
17	447	0.09	282	0.3	98	0.4	50	0.3	1	0.04
18	368	0.08	246	0.3	80	0.3	35	0.2	0	0.00
19	326	0.07	217	0.2	98	0.4	27	0.1	0	0.00
20	251	0.055	203	0.2	69	0.28	26	0.1	1	0.04

Table II.
Session lengths

Query length

Table III shows the query length (i.e. number of terms) to the diverse federated content collections.

Most queries were between one to three terms per query. Image and audio queries generally included one to two terms. Web, audio and news queries were longer.

Number of results pages viewed

Table IV shows the number of results pages viewed from the diverse federated content collections.

Overall, most users viewed one results page during their Web search session. More image-seeking users also examined second and third page results. Web collection searchers were more likely to view only the first results page.

Use of system assistance

Table V shows the use of system assistance when searching the diverse federated content collections. Dogpile.com offers an alternate query re-formulation feature.

Across the various content collections, most users did not seek systems assistance. Interestingly, more users seeking image and videos sought systems assistance.

Repeat queries

Table VI shows the most common repeat queries to the diverse federated content collections.

Table VI shows the top ten repeat queries from each content collection. There were nine queries that were the in the top queries from more than one source. Most of these queries were for people, places, or things.

Query length (terms)	Web	%	Images	%	Audio	%	Video	%	News	%
1	180,470	16.6	74,054	25.4	15,470	16.2	1,0899	22.6	744	16.6
2	316,338	29.1	122,192	42.05	30,008	31.5	2,0712	43.09	1,752	39.1
3	280,473	25.8	61,043	21.008	20,651	21.7	9,930	20.6	906	20.2
4	153,570	14.1	21,564	7.4	13,854	14.5	4,116	8.5	531	11.8
5	77,820	7.1	7,643	2.6	8,129	8.5	1,516	3.1	226	5.05
6	38,192	3.5	2,577	0.8	3,928	4.1	533	1.1	138	3.08
7	19,192	1.7	906	0.3	1,749	1.8	219	0.4	89	1.9
8	10,185	0.9	364	0.1	829	0.8	76	0.1	46	1.02
9	5,245	0.4	132	0.04	312	0.3	36	0.07	32	0.7
10	2,687	0.2	72	0.02	112	0.1	10	0.02	9	0.2
11	1042	0.09	18	0.006	53	0.05	10	0.02	1	0.022
12	290	0.02	5	0.002	16	0.01	0	0.00	0	0.00
13	55	0.005	0	0.0	6	0.006	0	0.00	0	0.00
14	8	0.001	1	0.0	0	0.00	0	0.00	0	0.00
15	2	0.00	0	0.0	1	0.001	0	0.00	0	0.00
18	1	0.00	0	0.0	0	0.00	0	0.00	0	0.00
24	1	0.00	0	0.0	0	0.00	0	0.00	0	0.00
25	2	0.00	0	0.0	0	0.00	0	0.00	0	0.00
Total	1,085,573	100.0	290,571	100.0	95,118	100.0	48,057	100.0	4,474	100.0

Table III.
Query lengths

Results pages	Web	%	Images	%	Audio	%	Video	%	News	%
1	781,119	71.9	171,869	59.1	64,145	67.4	32,298	67.2	3,123	69.7
2	171,613	15.8	53,875	18.5	17,853	18.7	9,472	19.7	905	20.2
3	56,472	5.2	37,649	12.9	6,730	7.07	3,142	6.53	240	5.3
4	32,295	2.9	12,619	4.3	3,097	3.2	1,337	2.7	110	2.4
5	16,192	1.4	5,316	1.8	1,274	1.3	664	1.3	37	0.8
6	9,551	0.8	3,741	1.2	883	0.9	407	0.8	27	0.6
7	5,200	0.4	1,692	0.5	389	0.4	230	0.4	8	0.1
8	3,621	0.3	1,159	0.3	270	0.2	136	0.2	8	0.1
9	2,338	0.2	727	0.2	138	0.1	80	0.1	5	0.1
10	1,711	0.1	512	0.1	105	0.1	72	0.1	2	0.04
11	1,192	0.1	348	0.1	66	0.06	53	0.1	3	0.06
12	854	0.07	255	0.08	46	0.04	39	0.08	1	0.02
13	668	0.06	172	0.05	29	0.03	15	0.03	1	0.02
14	538	0.05	129	0.04	27	0.02	23	0.04	1	0.02
15	397	0.037	100	0.03	11	0.01	19	0.04	1	0.02
16 +										
Total	1,085,568	100.0	290,569	100.0	95,118	100.0	48,057	100.0	4,475	100.0

Table IV.
Viewing of results pages

System assistance	Source: Web	%	Images	%	Audio	%	Video	%	News	%
Yes	70,049	6.5	44,985	15.5	6,236	6.6	6,401	13.3	455	10.2
No	1,015,524	93.5	245,586	84.5	88,882	93.4	41,656	86.7	4,019	89.8
Total	1,085,573	100.0	290,571	100.0	95,118	100.0	48,057	100.0	4,474	100.0

Table V.
Use of system assistance

Table VI.
Repeat queries

	Query	Web	%	Images	%	Audio	%	Video	%	News	%	Total	%
1	lohan pics	2,586	0.238	555	0.191							3,141	0.206
2	music lyrics	2,436	0.224									2,436	0.160
3	american idol	1,566	0.144							41	0.916	1,607	0.105
4	games	1,240	0.114									1,240	0.081
5	poetry	1,181	0.109									1,181	0.078
6	funny jokes	1,054	0.097									1,054	0.069
7	paris hilton			571	0.197			203	0.422	9	0.201	783	0.051
8	google	694	0.064							5	0.112	699	0.046
9	yahoo					676	0.711					676	0.044
10	ebay	637	0.059									637	0.042
11	playstation 2 cheats	637	0.059									637	0.042
12	sex			311	0.107			201	0.418			512	0.034
13	carmen electra			383	0.132			71	0.148			454	0.030
14	girls			372	0.128			75	0.156			447	0.029
15	p****			353	0.121							353	0.023
16	britney spears			263	0.091							263	0.017
17	eminem					243	0.255					243	0.016
18	pamela anderson			214	0.074							214	0.014
19	green day					209	0.220					209	0.014
20	jennifer lopez			209	0.072							209	0.014
21	candy shop					177	0.186					177	0.012
22	system of a down					174	0.183					174	0.011
23	ludacris					163	0.171					163	0.011
24	porn							135	0.281			135	0.009
25	hollaback girl					133	0.140					133	0.009
26	usher					127	0.134					127	0.008
27	lesbians							86	0.179			86	0.006
28	funny							82	0.171			82	0.005
29	hentai							78	0.162			78	0.005
30	jenna jameson							76	0.158			76	0.005
31	lesbian							67	0.139			67	0.004

(*continued*)

	Query	Web	%	Images	%	Audio	%	Video	%	News	%	Total	%
32	cdc picaridin									24	0.536	24	0.002
33	cdc picaridin sc johnson									24	0.536	24	0.002
34	copernic									16	0.358	16	0.001
35	kentucky derby									10	0.224	10	0.001
36	griswold iowa fire									6	0.134	6	0.000
37	"debbie fields"									9	0.201	9	0.001
38	50 cent					371	0.390					371	0.024
39	adam long									5	0.112	5	0.000
40	akon					141	0.148					141	0.009
41	akon lonely					119	0.125					119	0.008
	Total	12,031	1.108	3,231	1.112	2,533	2.663	1,074	2.235	149	3.330	19,018	1.248
	Total (of all queries from this source)	1,085,573	100	290,571	100	95,118	100	48,057	100	4,474	100	1,523,793	100.000

Table VI.

Discussion

This paper provides preliminary results from a large-scale study of user access to federated content collections via the Dogpile.com. Across the federated content collections, there were some differences in user access. Overall, searchers of content collections were entering few queries, entered short queries, using limited Web systems assistance and viewing few pages of results. Overall, the Web collection was the most popular, followed by the image collection. Users displayed characteristics found in other studies of public Web searching (Spink and Jansen, 2004; Spink *et al.*, 2002). Most searchers accessed the Web collection, followed by the image and audio collections. Users included between one to three queries in search sessions. Most users across the various federated content collections entered only one query. News sessions were shorter and included fewer queries.

Audio sessions were longer, but with fewer queries per session. Image and audio queries generally included one to two terms. Web, audio and news queries were longer. Most searchers examined only the first results page. However, people seeking images examined further results pages. Across content collections, most users did not seek systems assistance. Interestingly, more users seeking image and videos sought systems assistance. Image searches were longer and used more terms, Web searches were shorter with fewer queries and viewing fewer results pages. Image and audio searches were longer, including more queries. This is similar to findings by Jansen *et al.* (2005).

The nine most frequent queries were for popular people and celebrities, places, or things. This is also similar to recent studies of Web search topics by Jansen *et al.* (in press), and Spink and Jansen (2004). Commerce related queries were the most frequently occurring, followed by people, places and things, and unknown queries (indiscernible or non-English), and sexual and pornographic queries represented a very low proportion of all queries. Recently, Koshman *et al.* (in press) found that one in five queries submitted to Vivisimo related to commerce, travel, employment or the economy. One in five queries were indiscernible or non-English. This represents a sizable proportion of all queries. In addition, one in seven queries was related to people, places or things. These queries included personal names or the names of locations.

Our analysis shows that, similar to Web searching overall, most content collection searches are short, contain few terms, and results pages are viewed, except for image searches. Jansen *et al.* (in press) recently found that Dogpile.com searchers use about three terms per query (mean = 2.85), and generally (56 per cent of users) spend less than one minute interacting with the Web search engine. Overall, the level of user interaction is higher on Dogpile.com than results from other Web search engines, and Dogpile.com users spend less time on the Web search engine. Most users entered between one and three terms per query.

Conclusion and further research

Our findings provide important insights into the current state of federated Web searching and Web usage for users, search engine developers and Web site designers. The study represents a major study of human interaction with a major meta-search Web search engine. We are currently conducting more analysis of further Dogpile.com search data to examine further aspects of multimedia federated search. We are continuing to track Web search trends and characteristics, using either transaction logs analysis or lab studies, in order to assess future behaviour and identify future user needs.

References

Avrahami, T.T., Yau, L., Si, L. and Callan, J. (2006), "The FedLemur project: federated search in the real world", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 3, pp. 347-58.

Becarevic, D. and Roantree, M. (2004), "A metadata approach to multimedia database federations", *Information and Software Technology*, Vol. 46 No. 3, pp. 195-207.

Callan, J. (2000), "Distributed information retrieval", in Croft, W.B. (Ed.), *Advances in Information Retrieval*, Kluwer Academic Publishers, New York, NY, pp. 127-50.

comScore Networks (2005), available at: www.comscore.com/press/release.asp?press=325

EdNA (2005), EdNA Online: Education Network Australia, available at: www.edna.edu.au

Globe (2005), Globe Learning Object Brokered Exchange, available at: http://taste.merlot.org/initiatives/globe.htm

Hitwise (2005), available at: www.clickz.com/stats/sectors/search_tools/article.php/3528456

Jansen, B.J. and Pooch, U. (2000), "Web user studies: a review and framework for future work", *Journal of the American Society of Information Science and Technology*, Vol. 52 No. 3, pp. 235-46.

Jansen, B.J. and Spink, A. (2005), "An analysis of Web searching by European Alltheweb.com users", *Information Processing and Management*, Vol. 41 No. 2, pp. 361-81.

Jansen, B.J., Spink, A. and Pederson, J. (2005), "Trend analysis of AltaVista Web searching", *Journal of the American Society for Information Science and Technology*, Vol. 56 No. 6, pp. 559-70.

Jansen, B.J., Spink, A., Blakely, C. and Koshman, S. (in press), "Web searcher interaction with the Dogpile Web search engine", *Journal of the American Society for Information Science and Technology*.

Khoossainov, R. and Kushmerick, N. (2004), "Specialization dynamics in federated Web search", *WIDM 2004: 6th ACM Workshop on Web Information and Data Management*, Washington, DC, 12-13 November, pp. 112-9.

Koshman, S., Spink, A. and Jansen, B.J. (in press), "Web searching on the Vivisimo search engine", *Journal of the American Society for Information Science and Technology*.

Ozmutlu, C., Spink, A. and Ozmutlu, S. (2003), "Multimedia web searching trends, 1997-2001", *Information Processing & Management*, Vol. 39 No. 4, pp. 611-21.

Park, S., Bae, H. and Lee, J. (2005), "End user searching: a Web log analysis of NAVER, a Korean Web search engine", *Library and Information Science Research*, Vol. 27 No. 2, pp. 203-21.

Powell, J. and Fox, E. (1998), "Multilingual federated searching across heterogeneous collections", *D-Lib Magazine*, September.

Rehak, D.R., Dodds, P. and Lannom, L. (2005), "A model and infrastructure for federated learning content repositories", *Proceedings of WWW 2005: International World Wide Web Conference, May 10-14, Chiba, Japan.*

Si, L. and Callan, J. (2005), "Modeling search engines' effectiveness for federated search", *SIGIR 2005, Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in information Retrieval, August 15-19, Salvador, Brazil.*

Spink, A. and Jansen, B.J. (2004), *Web Search: Public Searching of the Web*, Springer, Berlin.

Spink, A., Jansen, B.J., Wolfram, D. and Saracevic, T. (2002), "From e-sex to e-commerce: Web search changes", *IEEE Computer*, Vol. 35 No. 3, pp. 133-5.

Xu, J. and Callan, J. (1998), *"Effective retrieval with distributed collections", ACM SIGIR 1998: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in information Retrieval, August 15-19, Salvador, Brazil*, pp. 112-20.

About the authors

Amanda Spink is a Professor of Information Technology at the Queensland University of Technology. Her research into human information behaviour and information retrieval/Web studies includes more than 240 publications and the recent books, *Web Search: Public Searching of the Web: New Directions in Cognitive Information Retrieval,* and *New Directions in Human Information Behavior,* for Springer. She is the corresponding author and can be contacted at: ah.spink@qut.edu.au

Bernard J. Jansen is Assistant Professor of Information Sciences and Technology at the Pennsylvania State University. His research focuses on improving access to information in various domain settings, including sponsored search, Web searching, and complex information spaces. He has numerous publications in a variety of outlets.