Web Searcher Interactions with Multiple Federate Content Collections

Amanda Spink
Faculty of IT
Queensland
University of
Technology
QLD 4001 Australia
ah.spink@qut.edu.au

Bernard J. Jansen
School of IST
The Pennsylvania
State University
PA 16802 USA
jjansen@ist.psu.edu

Chris Blakely
Infospace, Inc WA
98004 USA
chris.blakely@Infosp
ace.com

Sherry Koshman
School of Information
Sciences
University of
Pittsburgh
PA 15260 USA
skoshman@sis.pitt.edu

Abstract Federated content collections are important for providing access to multiple content repositories. Our paper provides preliminary results from a large-scale study of user access to federated content collections via the Dogpile.com Web search engine. We examined differences in searching patterns to various federated content collections by analyzing a subset of queries submitted by searchers on Dogpile.com. Findings include differences in content collection searches. Image and audio searches were longer sessions but shorter queries. Most users view few results pages.

Keywords Federated document collection, Web, Dogpile.com

1 Introduction

Federated content collection is a content organizing scheme involving multiple repositories of content instead of a central repository. These individual repositories typically have their own store, indexing, and retrieval algorithms. Major Web search engine typically offer tabbed interfaces that permit users to search multiple federated content collections, such as Web documents, images, audio, and video files. Few studies have 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 [1, 2]. Examination of how people use federated content collections is an important area of document computing research. Our paper provides preliminary results from a large-scale study of user access to federated content collections via the Dogpile.com Web search engine.

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

Proceedings of the 10th Australasia Document Computing Symposium Sydney, Australia, December 12, 2005.

2 Related studies

Rehak, Dodds and Lannom [3] developed a model and infrastructure for federated learning content repositories. Becarevic and Roantree [4] studied federated multimedia database systems. However, there have been limited studies investigated the effect of federated collections on Web search. Ozmutlu, Spink, Ozmutlu [5] examined the impact of multimedia interface buttons on the Excite search engine, investigating multimedia queries in the general query population prior 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, Spink and Pederson [6, 7] 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 searchers were longer than any other type of searching and Boolean usage by image searchers was 28%.

3 Research goals

The major 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:

1. Session length
2. Query length
3. Number of results pages viewed
4. Use of system assistance

5. Repeat queries

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 search topics.

4 Research design

4.1 Dogpile.com

Dogpile.com (http://www.Dogpile.com/) is owned by Infospace, a market leader in the meta-search engine business. 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!). With results from these four Web search engines; Dogpile.com.com leverages one of the most comprehensive content collections on the Web in response to Web searchers' queries. 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 duplicates 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. 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 2005 Dogpile.com had the industry highest visitor-to-searcher conversion rate of 83% (i.e., 83% of the visitors to the Dogpile.com site executed a search).

4.2 Data collection

We recorded the records of searcher – system interactions 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 three fields:

User Identification: an anonymous user code automatically assigned by the Dogpile.com server to identify a particular computer
Cookie: 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.
Query Terms: terms exactly as entered by the given user.
Source: the content collection that the user selects to search (e.g., Web, Images, Audio, or Video) with Web being the default.
Feedback: a binary code denoting whether or not the query was generated by the "Are You Looking for?" query reformulation assistance.

4.3 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 [8, 9, 10]:
Term: series of characters separated by white space or other separator
Query: string of terms submitted by a searcher in a given instance
Repeat query: query submitted more than once during the data collection period, irrespective of the user.
Session: series of queries submitted by a user during one interaction with the Web search engine.
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.

5 Results

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

5.1 Content collections

Table 1 shows the usage of each of the five federated Dogpile.com content collections (Web, Images, Audio, Video and News).

[1] Hitwise, 2005. http://www.clickz.com/stats/sectors/search_tools/article.php/3528456.
[2] comScore, 2005 http://www.comscore.com/press/release.asp?press=325.

Source	Occurrences	Percent
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%

Table 1: Use of the Dogpile.com content collections.

Table 1 shows that the Web was the most popular content collection, with more than 71% of all searches being executed again this content collection. Images were the second most popular content collection, followed by the audio, video and news collection.

5.2 Session length

Table 2 shows the session length (i.e., number of queries) for queries to the diverse federated content collections.

Session Length	Web	%	Images	%	Audio	%	Video	%	News	%
1	258204	56.843%	40026	51.127%	10404	42.903%	6741	47.462%	1757	69.502%
2	77884	17.146%	12420	15.865%	4004	16.511%	2476	17.433%	373	14.755%
3	40793	8.981%	6328	8.083%	2391	9.860%	1357	9.554%	179	7.081%
4	24067	5.298%	4087	5.220%	1609	6.635%	881	6.203%	74	2.927%
5	15341	3.377%	2772	3.541%	1161	4.788%	590	4.154%	47	1.859%
6	10015	2.205%	2065	2.638%	839	3.460%	412	2.901%	26	1.028%
7	6839	1.506%	1601	2.045%	667	2.751%	327	2.302%	21	0.831%
8	4942	1.088%	1202	1.535%	478	1.971%	229	1.612%	13	0.514%
9	3618	0.797%	1070	1.367%	413	1.703%	205	1.443%	11	0.435%
10	2581	0.568%	805	1.028%	346	1.427%	158	1.112%	3	0.119%
11	1873	0.412%	718	0.917%	251	1.035%	115	0.810%	7	0.277%
12	1506	0.332%	596	0.761%	202	0.833%	105	0.739%	3	0.119%
13	1130	0.249%	498	0.636%	217	0.895%	73	0.514%	3	0.119%
14	881	0.194%	438	0.559%	152	0.627%	62	0.437%	3	0.119%
15	729	0.160%	358	0.457%	118	0.487%	61	0.429%	1	0.040%
16	609	0.134%	338	0.432%	111	0.458%	50	0.352%	2	0.079%
17	447	0.098%	282	0.360%	98	0.404%	50	0.352%	1	0.040%
18	368	0.081%	246	0.314%	80	0.330%	35	0.246%		0.000%
19	326	0.072%	217	0.277%	98	0.404%	27	0.190%		0.000%
20	251	0.055%	203	0.259%	69	0.285%	26	0.183%	1	0.040%

Table 2: Session lengths.

Most users included between one and three queries in their federated content search sessions. Some 50% 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.

5.3 Query length

Table 3 shows the query length (i.e., number of terms) to the diverse federated content collections.

Query Length (Terms)	Web	%	Images	%	Audio	%	Video	%	News	%
1	180470	16.624%	74054	25.486%	15470	16.264%	10899	22.679%	744	16.629%
2	316338	29.140%	122192	42.052%	30008	31.548%	20712	43.099%	1752	39.160%
3	280473	25.836%	61043	21.008%	20651	21.711%	9930	20.663%	906	20.250%
4	153570	14.146%	21564	7.421%	13854	14.565%	4116	8.565%	531	11.869%

Query Length (Terms)	Web	%	Images	%	Audio	%	Video	%	News	%
5	77820	7.169%	7643	2.630%	8129	8.546%	1516	3.155%	226	5.051%
6	38192	3.518%	2577	0.887%	3928	4.130%	533	1.109%	138	3.084%
7	19192	1.768%	906	0.312%	1749	1.839%	219	0.456%	89	1.989%
8	10185	0.938%	364	0.125%	829	0.872%	76	0.158%	46	1.028%
9	5245	0.483%	132	0.045%	312	0.328%	36	0.075%	32	0.715%
10	2687	0.248%	72	0.025%	112	0.118%	10	0.021%	9	0.201%
11	1042	0.096%	18	0.006%	53	0.056%	10	0.021%	1	0.022%
12	290	0.027%	5	0.002%	16	0.017%		0.000%		0.000%
13	55	0.005%		0.000%	6	0.006%		0.000%		0.000%
14	8	0.001%	1	0.000%		0.000%		0.000%		0.000%
15	2	0.000%		0.000%	1	0.001%		0.000%		0.000%
18	1	0.000%		0.000%		0.000%		0.000%		0.000%
24	1	0.000%		0.000%		0.000%		0.000%		0.000%
25	2	0.000%		0.000%		0.000%		0.000%		0.000%
	1085573	100.000%	290571	100.000%	95118	100.000%	48057	100.000%	4474	100.000%

Table 3: Query lengths.

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.

5.4 Number of results pages viewed

Table 4 shows the number of results pages viewed from the diverse federated content collections.

Results Pages	Web	%	Images	%	Audio	%	Video	%	News	%
1	781119	71.955%	171869	59.149%	64145	67.437%	32298	67.208%	3123	69.788%
2	171613	15.809%	53875	18.541%	17853	18.769%	9472	19.710%	905	20.223%
3	56472	5.202%	37649	12.957%	6730	7.075%	3142	6.538%	240	5.363%
4	32295	2.975%	12619	4.343%	3097	3.256%	1337	2.782%	110	2.458%
5	16192	1.492%	5316	1.830%	1274	1.339%	664	1.382%	37	0.827%
6	9551	0.880%	3741	1.287%	883	0.928%	407	0.847%	27	0.603%
7	5200	0.479%	1692	0.582%	389	0.409%	230	0.479%	8	0.179%
8	3621	0.334%	1159	0.399%	270	0.284%	136	0.283%	8	0.179%
9	2338	0.215%	727	0.250%	138	0.145%	80	0.166%	5	0.112%
10	1711	0.158%	512	0.176%	105	0.110%	72	0.150%	2	0.045%
11	1192	0.110%	348	0.120%	66	0.069%	53	0.110%	3	0.067%
12	854	0.079%	255	0.088%	46	0.048%	39	0.081%	1	0.022%
13	668	0.062%	172	0.059%	29	0.030%	15	0.031%	1	0.022%
14	538	0.050%	129	0.044%	27	0.028%	23	0.048%	1	0.022%
15	397	0.037%	100	0.034%	11	0.012%	19	0.040%	1	0.022%
16+										
	1085568	100.0%	290569	100.0%	95118	100.0%	48057	100.0%	4475	100.0%

Table 4: Viewing of results pages.

Overall, most users' viewed one results page during their 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.

5.5 Use of system assistance

Table 5 shows the use of system assistance when searching the diverse federated content collections. Dogpile.com offers an alternate query re-formulation feature.

System Assistance	Source									
	Web	%	Images	%	Audio	%	Video	%	News	%
Yes	70049	6.5%	44985	15.5%	6236	6.6%	6401	13.3%	455	10.2%
No	1015524	93.5%	245586	84.5%	88882	93.4%	41656	86.7%	4019	89.8%
	1085573	100.0%	290571	100.0%	95118	100.0%	48057	100.0%	4474	100.0%

Table 5: Use of system assistance.

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

5.6 Repeat queries

Table 6 shows the most common repeat queries to the diverse federated content collections.

	Query	Web	%	Images	%	Audio	%	Video	%	News	%	Total	%
1	lohan pics	**2586**	**0.238%**	555	0.191%							3141	0.206%
2	music lyrics	2436	0.224%									2436	0.160%
3	american idol	1566	0.144%							**41**	**0.916%**	1607	0.105%
4	games	1240	0.114%									1240	0.081%
5	poetry	1181	0.109%									1181	0.078%
6	funny jokes	1054	0.097%									1054	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%

	Query	Web	%	Images	%	Audio	%	Video	%	News	%	Total	%
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%
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	12031	1.108%	3231	1.112%	2533	2.663%	1074	2.235%	149	3.330%	19018	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 6. Repeat queries.

Table 6 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 popular people, places, or things.

6 Discussion

Our 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 users' access. 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.

Most users' entered between one and three terms per query. 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, similar to findings by Jansen, Spink and Pedersen [6], and Spink and Jansen [10].

The nine most frequent queries were for popular people and celebraties, places, or things.

7 Conclusion and further research

Our preliminary analysis shows that users' differ in their access to the various content collections. Similar to Web searching overall, most content collection searches are short, and contain few terms and results pages are viewed, except for image searches. We are currently conducting further analysis of the Dogpile.com users and their search processes.

Acknowledgment We thank Infospace, Inc for providing the Web search engine data set.

References

[1] EdNA Online: Education Network Australia http://www.edna.edu.au

[2] Globe (Globe Learning Object Brokered Exchange) http://taste.merlot.org/initiatives/globe.htm

[3] D. R. Rehak, P. Dodds and L. Lannom. A model and infrastructure for federated learning content repositories. In *WWW 2005: International World Wide Web Conference, May 10-14, Chiba, Japan*.

[4] D. Becarevic and M. Roantree. A metadata approach to multimedia database federations. *Information and Software Technology*, Volume 46, Number 3, pages 195-207, 2004.

[5] C. Ozmutlu, Spink and S. Ozmutlu. Multimedia web searching trends: 1997-2001. *Information Processing & Management*, Volume 39, Number 4, pages 611-621, 2003.

[6] B. J. Jansen, A. Spink and J. Pederson. Trend analysis of AltaVista web searching. *Journal of the American Society for Information Science and Technology*, Volume 56, Number 6, pages 559-570, 2005.

[7] B. J. Jansen and A. Spink. An analysis of web searching by European Alltheweb.com users. *Information Processing and Management,* Volume 41, Number 2, pages 361-381, 2005.

[8] B. J. Jansen and U. Pooch. Web user studies: A review and framework for future work. *Journal of the American Society of Information Science and Technology*. Volume 52, Number 3, pages 235-246, 2000.

[9] S. Park, H. Bae and J. Lee. End user searching: A web log analysis of NAVER, a Korean Web search engine. *Library & Information Science Research*, Volume 27, Number 2, pages 203-221, 2005.

[10] A. Spink and B. J. Jansen. *Web Search: Public Searching of the Web*. New York: Kluwer, 2004.

.