Multimedia Web Searching on a Meta-Search Engine

Dian Tjondronegoro & Amanda Spink

Faculty of Information Technology
Queensland University of Technology
2 George St, Brisbane, QLD 4001

dian@qut.edu.au, ah.spink@qut.edu,au

Bernard J. Jansen

College of Information Science and
Technology
The Pennsylvania State University
State College PA 16802 USA

jjansen@psu.edu

Abstract *This paper provides preliminary results from a major study of multimedia Web searching by Dogpile meta-search engine users, including queries and session characteristics, and changes or differences in image, video and audio searching. The results are compared with multimedia Web searching studies from 1997 to 2002. Image and sexual queries are dominant in multimedia Web searching. The paper provides important implications for the design of multimedia information retrieval systems.*

Keywords

1. Introduction

Tracking the trends in users' interactions with Web search engines has emerged an important area of research within information retrieval (IR). Trends in multimedia Web searching are playing an important role in new designs for multimedia Web retrieval systems. Next generation multimedia search engines need to support user's needs and searching behavior [3, 8]. Research shows that users still find it difficult to develop keywords and queries as most multimedia retrieval needs are often not describable using exact keywords or sample images/sounds [8]. Users' satisfaction on image search is often compromised, as they often cannot formulate better search queries [4].

Many studies have analyzed different aspects of Web query data logs, including multimedia Web search [6, 9]. A study of multimedia Web search behavior using the Excite dataset 1997-2001 [5] has found that multimedia queries are generally longer and mostly performed with multimedia interface buttons that allow users to select the particular type of media to search. Tjondronegoro, et al., (in press) found that: (1) few major Web search engines offer multimedia searching and (2) multimedia Web search functionality is generally limited. Despite the increasing level of interest in multimedia Web search, those few Web search engines offering multimedia Web search, provide limited multimedia search functionality. Keywords are still the only means of multimedia retrieval, while other methods such as "query by example" are offered by less than 1% of Web search engines examined. The study reported in this paper is part of an ongoing research project to investigating Web searching behavior [2, 7, 9, & 11]. Major studies of Web user behavior are significant for the development of more effective multimedia IR systems.

In this paper, we report findings from a study of trends in multimedia Web searching by Dogpile users in 2006, including queries and session characteristics, and changes or differences in image, video and audio searching. To assess the new trends and changes in Web multimedia searching, we compare our 2006 preliminary findings with multimedia Web searching trends from 1997 to 2002 [1, 5]. Our paper provides implications for the development of the next generation of multimedia Web search engines.

2. Research design

Our research goals were to examine (1) multimedia Web searching by Dogpile users in 2006, (2) queries and session characteristics, and changes or differences in image, video and audio searching, and (3) compare our findings previous studies of multimedia Web searching trends from 1997 to 2002 [1, 5].

2.1 Dogpile web query log fields

Dogpile (http://www.dogpile.com/) is a meta-search engine, which combines results from multiple search engines. Searches are based on the exact terms entered by user as a query. The Dogpile transaction log used in this paper consists of 1,228,330 queries that are combined from the multimedia tabbed interfaces, namely image, audio, and video. The data collection date was 15 May 2006 with 128,305 search sessions. Each query log record contains 6 fields:

- *Identification*: anonymous code assigned by Dogpile server to a user machine
- *IP:* user machine's Internet address which can be used to uniquely identify users
- *Cookie*: this number is assigned from the first time

a user is connected to the search engine until the user left a session, therefore is used as sessions identification.

- *Time of Day*: the time (in hours, minutes, and seconds) that a particular query is submitted
- *Query:* the user terms entered into the query box (e.g. "Web search")
- *Organic/Sponsored:* the number of clicks on the organic and sponsored links (indicating that the link is followed up the user and may be deemed as relevant or interesting). Organic links are naturally returned by the search engine (based on the search algorithm), while sponsored links are paid by advertisers.

To obtain human (not agent) sessions, all sessions with more than 101 queries submitted (in a session) were deemed to be conducted by agents. This threshold is used, as it is almost 50 times greater than the reported mean search session for human Web searchers [5].

2.2 Quantitative analysis

Analyses were conducted at multiple levels, using the following metrics:

- *Terms level* - any string of characters bounded by some delimiter such as white space.
- Q*uery* level – a query is a string list of one or more entered terms.

○ *Query length* is measured by counting the number of terms in the query.

○ *Query complexity* examines the query syntax, including the advanced searching techniques such as Boolean and other query operators.

- S*ession* level – a session is the entire sequence of queries entered by a searcher with a given data sampling method.

○ *Session length* is measured by the number of queries per searcher as each searcher is given a unique identifier within the log, namely the IP address.

- *Results pages viewed* level – a results page is the list of results returned by a search engine in response to a query, either organic or sponsored. The result page viewing patterns of Web searchers are analyzed through the number of results pages viewed.
- *Click-through level* – from the results page, a searcher may click on a URL to visit one or more results from the results listing, a method which is often referred as page view analysis. When a link is being followed through by users, we can assume that the result is relevant.

3. Results

After being filtered, we focused our study on 127,613 human Dogpile search sessions.

5.1 Search sessions

Table 1 shows a comparison of sessions and query length in multimedia searching. Image search dominate (i.e. 60% of multimedia sessions), 32% were audio sessions and 18% were video queries. Compared to the 1997 - 2001 Excite Web log trend study [1], this proportion is still consistent; averaged over 3 years, 50% of multimedia search is for image, 28% for video, and 22% for audio. However, video is becoming even less dominant, as it is still the most heavy - band media and probably due to the growing popularity of independent streaming video hosts, such as *YouTube,* which have their own internal search system.

The mean audio's session duration is the highest at 30.3 minutes, while video is 20.7 minutes, and image is 20.3 minutes. The mean queries per session show that users have entered more queries on video search (2.9) followed by audio (2.4) and image (2.1) respectively. Compared to 1997-2001 on 'mean session duration' and 'queries per session' depicted in Table 1, the majority of users in 2001 and 2006 equally are not spending more than 30 minutes per session, while the average session duration has not changed much (i.e. 200s or 16% increase in image, 400s or 22% decrease in audio, and 150s or 13% increase in video).

We also examined the percentage of session duration ranging between less than 5 minutes and more than 5 hours. The findings are presented by Table 2 to show the actual percentages and the overall distribution.

The distribution is almost equivalent between image, audio and video, that is "less than 5 minutes' session duration being the highest (i.e. greater than 58% of the total durations for each type of media), while each of the other session duration length categories being less than 10%.

5. 2 Terms per multimedia query

We studied the mean terms per query for all identified audio, video and image queries. A total of 1,849,410 terms were used in 1,129,041 multimedia queries, the average terms per query for audio search is slightly longer being 3.1, followed by audio and video at 2.3 terms per query equally. This is most likely because audio search queries usually contain terms from songs title, which are generally longer. Query length for multimedia searching generally ranges between 1 to 4 terms. Two terms are most commonly used. Table 3 shows the typical length of multimedia search query and the occurrences based on our Web log data.

Variables	Multimedia by Humans	Image	Audio	Video
Total sessions	127,813 (100%)	76,155 (60%)	41,335 (32%)	23,945 (18%)
Average Duration per Session (in seconds)	1561.6 (26 mins)	1217.1 (20.3 mins)	1820.6 (30.3 mins)	1242.3 (20.7 mins)
Maximum Duration Per Session (in seconds)	86235 (23.9 hours)	86235 (23.9 hours)	85441 (23.73 hours)	86218 (23.9 hours)
Standard Deviation for Duration of Sessions (in seconds)	5673.9 (94.5 mins)	4965.76 (82.8 mins)	6027.39 (100.5 mins)	5244.4 (87.4 mins)
Average Queries Per Session	2.4	2.1	2.4	2.9
Maximum Queries Per Session	86	65	73	86
Standard Deviation for Queries Per Session	2.75	2.03	2.8	3.9

Table 1. Comparison of audio, video and images search sessions.

Session Duration	Occurrences in Image Search (%)	Occurrences in Audio Search (%)	Occurrences in Video Search (%)
Less than 5 minutes	50,819 (66.7%)	24,154 (58.4%)	15,772 (65.9)
5 to 10 minutes	7,323 (9.6%)	3819 (9.2)	2373 (9.9)
10 to 15 minutes	4,171 (5.5%)	2425 (5.9)	1324 (5.5)
15 to 30 minutes	5,802 (7.6%)	3873 (9.4)	1921 (8.0)
30 to 60 minutes	3,413 (4.5%)	2917 (7.1)	1198 (5.0)
1 to 2 hour	1,790 (2.4%)	1525 (3.7)	553 (2.3)
2 to 3 hour	775 (1%)	707 (1.7)	232 (1)
3 to 4 hour	523 (0.7%)	509 (1.2)	146 (0.6)
4 to 5 hour	389 (0.5%)	452 (1.1)	116 (0.5)
More than 5 hours	1,150 (1.5%)	954 (2.3)	360 (1.5)
Total	76,155	41,335	23,945

Table 2. Session duration analysis in multimedia search.

Query Length	Occurrences in Image Search	Occurrences in Image Search (%)	Occurrences in Audio Search	Occurrences in Audio Search (%)	Occurrences in Video Search	Occurrences in video Search
1	31,748	23.9	13,663	15.2	9,575	22.8
2	49,094	36.9	22,601	25.2	15,200	36.1
3	28,424	21.4	18,621	20.7	8,869	21.1
4	13,744	10.3	14,685	16.3	4,713	11.2
5	6,014	4.5	9,531	10.6	2,164	5.1
6	2,390	1.8	5,552	6.2	956	2.3
7	918	0.7	2,857	3.2	397	0.9
8	386	0.3	1,250	1.4	65	0.2
9	180	0.1	613	0.7	58	0.1
>=10	135	0.1	489	0.5	55	0.1
Total	133,033		89,862		42,052	

Table 3: Query length per session statistics in multimedia search.

The distribution statistics of query length in video and image search is almost equivalent. Query length of greater than or equal to 6 is rarely used with percentage of less than 2%. In audio search, however, query length of 5 and 6 are more often used with percentages of 10.6% and 6.2% respectively.

Compared to 1997-2002 study, the query length distribution is almost the same for image, audio and video search. There are two notable differences: 1) in audio search 2002, query length of greater than or equal to 5 happened much less than 2006; 2) in image search 2002, the percentage of query length equal to 9 is 27% (which was deemed as an anomaly of the data collection).

5.3 Click-through analysis

Table 4 show the top 10 terms by which users found many relevant results as indicated by how many times they are entered by users (not necessarily unique) and the number of results viewed (i.e. clicked). All of top 10 terms in image queries are sexual. In audio queries, we found most of them are either song title or artist (or performer).

Image			Audio			Video		
Total Queries	562,380		Total Queries	370890		Total Queries	208115	
Query	Frequency (%)	Organic Links Followed	Query	Frequency (%)	Organic Links Followed	Query	Frequency (%)	Organic Links Followed
pussy	0.08	498	ridin dirty	0.07	427	pussy	0.63	1371
boobs	0.04	251	ridin	0.06	335	boobs	0.30	652
sex	0.04	237	shakira	0.06	318	hentai	0.23	514
hentai	0.03	196	eminem	0.06	343	sex	0.21	468
porn	0.03	181	50 cent	0.06	362	porn	0.17	370
tits	0.02	176	dani california	0.06	266	preteen	0.15	320
paris hilton	0.02	134	temperature	0.04	231	tits	0.15	344
milf	0.02	129	panic at the disco	0.04	191	paris hilton	0.14	310
ass	0.02	122	ms new booty	0.04	197	milf	0.14	315
penis	0.02	113	sean paul	0.04	199	ass	0.12	260

Table 4. Top 25 most occurring and followed up multimedia query terms.

Table 4 suggests that some queries co-occurred in multiple media searches. An analysis of overlapping queries (i.e. unique terms which are used in multiple multimedia types searches) shows that there are 12,730 overlapping queries that are used both in image and video searches. There are 5,742 queries co-occurring in image and audio searches, and 5,754 co-occurring in video and audio. Compared to the 2002 study of frequently occurring terms [1], this study demonstrates that the information need for multimedia Web searching remains consistent. In this new study, we have extended the statistics with the frequencies of followed-up links for each of the popular queries. The numbers have shown that the frequencies of followed up links are proportionally the same with the frequencies of the particular query is entered (i.e. video is much higher than image, and audio is slightly higher than image).

The number of search sessions can be sorted based on the most frequent into: image, audio, and video (highest to lowest). Whereas, the number of links clicked by users (which indicate the impact of the search results) can be sorted into: video, audio, image (highest to lowest). This phenomenon can be interpreted as an interesting search behavior in multimedia domain. It can show that users need to spend more time browsing, clicking and viewing the search results in image, audio, and video (respectively) to find the most relevant file, thereby reducing the number of sessions, which can be seen as new search process.

7. Conclusion

Most of image and video search is for sexual information, while music is the most popular genre for audio. Thus, it will be important for multimedia Web search engines to provide better and real-time support for automated content-based censorship (e.g. skin-based pornographic images identification) to protect children from accessing offensive materials. On the other side of the coin, it is highly likely that sex will remain to be the most popular multimedia genre being sought after by users if search engines would like to maximize the profits advertisements and sponsored links. Our future work aims to conduct a more thorough study using larger scale data and other data analysis techniques to investigate what type of other genres that attract users most (aside from sexual and music or movies).

References

[1] B. J. Jansen, A. Spink and J. Pedersen. An analysis of multimedia searching on AltaVista. In *Proc. SIGMM International Workshop on Multimedia Information Retrieval, Berkeley, CA, 2003.*

[2] B. J. Jansen, A. Spink and T. Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing and Management*, Volume 6, pages 207-227, 2000.

[3] M. L. Kherfi, D. Ziou and A. Bernardi, Image Retrieval from the World Wide Web: Issues, Techniques and Systems. *ACM Computing Surveys (CSUR)*, Volume 36, pages 35-67, 2004.

[4] S. McDonald and J. Tait. Search strategies in content-based image retrieval. In *26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), Amsterdam, Netherlands, 2003*

[5] S. Ozmutlu, A. Spink and H. Ozmutlu, Multimedia web searching trends: 1997- 2001. *Information Processing and Management*, Volume 1 Number 39, pages 611-621, 2003.

[6] C. Silverman, M. Henzinger, H. Marais and M. Morris. Analysis of a very large web search engine query log. *ACM SIGIR Forum*, Volume 33, 1999.

[7] A. Spink and B. J. Jansen. *Web Search: Public Searching of the Web.* Dordrecht: Springer.

[8] A. Spink. A user-centered approach to evaluating human interaction with web search engines: an exploratory study. *Information Processing & Management*, Volume 38, pages 401-426, 2002.

[9] A. Spink, B. J. Jansen, D. Wolfram and T. Saracevic. From e-sex to e-commerce: web search changes. *IEEE Computer*, Volume 35, pages 107-109, 2002.

[10] D. Tjondronegoro and A. Spink. Search engine Multimedia functionality. *Information Processing and Management* (In Press).

[11] D. Wolfram, A. Spink, B. J. Jansen and T. Saracevic. Vox populi: the public searching of the web. *Journal of the American Society for Information Sciences and Technology*, Volume 52, Number 12, pages 1073-1074, 2001.