



COVER SHEET

Ozmutlu, Seda and Spink, Amanda and Ozmutlu, Huseyin C. (2003) Multimedia web searching trends: 1997-2001. *Information Processing and Management* 39(4):pp. 611-621.

Accessed from <http://eprints.qut.edu.au>

Copyright 2003 Elsevier.

Information Processing and Management

MULTIMEDIA WEB SEARCHING TRENDS: 1997-2001

Seda Ozmutlu
Department of Industrial Engineering
Uludag University
Gorukle Kampusu, Bursa, 16059, Turkey
Tel: (90-224) 442-8176 Fax: (90-224) 442-8021
Email: seda@uludag.edu.tr

Amanda Spink*
School of Information Sciences and Technology
The Pennsylvania State University
004C Thomas Building, University Park PA 16802
Tel: (814) 865-4454 Fax: (814) 865-5604
Email: spink@ist.psu.edu

Huseyin C. Ozmutlu
Department of Industrial Engineering
Uludag University
Gorukle Kampusu, Bursa, 16059, Turkey
Tel: (90-224) 442-8176 Fax: (90-224) 442-8021
Email: hco@uludag.edu.tr

Published as: Ozmutlu, S., Spink, A., & Ozmutlu, H. C. (2003). Trends in multimedia Web searching: 1997-2001. *Information Processing and Management*. 39(4), 611-621.

* To whom all correspondence should be sent.

ABSTRACT

Multimedia is proliferating on Web sites, as the Web continues to enhance the integration of multimedia and textual information. In this paper we examine trends in multimedia Web searching by Excite users from 1997 to 2001. Results from an analysis of 1,025,910 Excite queries from 2001 are compared to similar Excite data sets from 1997 and 1999. Findings include: (1) queries per multimedia session have decreased since 1997 as a proportion of general queries due to the introduction of multimedia buttons near the query box, (2) multimedia queries identified are longer than non-multimedia queries, and (3) audio queries are more prevalent than image or video queries in identified multimedia queries. Overall, we see multimedia Web searching undergoing major changes as Web content and searching evolves.

INTRODUCTION

By 1999 more than 180 million images were present on the Web¹ and the number is growing daily. General studies and statistics on Web searching and search engines appear regularly (Silverman, et al., 1999; Spink, et al., 2002), including Web surfing (Huberman, et al., 1998), Web accessibility (Lawrence & Giles, 1999) and Web searching trends (Montgomery & Faloutsos, 2001). People frequently search the Web for multimedia using search engines that generally require the entry of query terms. A simple image search algorithm locates multimedia files by searching for file extensions and matching the filename to query terms. Web search services are increasingly providing special mechanisms for multimedia searching, e.g., Excite (<http://www.excite.com>), Alta Vista (<http://www.altavista.digital.com>) and Yahoo (<http://www.yahoo.com>). Users can also use file extensions, such as avi, wav or gif, to search for multimedia. As multimedia proliferates on Websites, we need to understand the trends or changes in multimedia Web searching over time. In this paper we compare multimedia Web searching by Excite users from 1997-2001.

Many studies have analyzed different aspects of Web query data logs (Silverman, et al., 1999; Spink, et al., 2002). Research has explored on image retrieval utilizing indexed image collections (Goodrum & Kim, 1998) and the design of multimedia IR systems (Aslandogan, et al., 1997), various aspects of audio and video retrieval (Brown, et al., 1996), and the demand for seeking video when designing a multimedia classroom (Smith, et al., 1998). Jansen, Goodrum and Spink (2000) conducted a major 1997 user study of multimedia searching using large-scale query data from the Excite Web search engine. Image queries were the most common Web multimedia searches with longer sessions than video and audio sessions. Audio queries were longer than image or video queries. Goodrum and Spink (2001) found that Excite image queries in 1997 contained a large number of unique terms. The most frequently occurring image related terms appeared less than 10 percent of the time, with most terms occurring only once.

Our study is also part of a larger ongoing research project to investigating Web searching behavior using Excite data sets (Jansen, Spink & Saracevic, 2000; Spink, Wolfram, Jansen & Spink, 2001; Spink, et al., 2001, 2002; Wolfram, et al., 2001). We also examined: (1) impact of multimedia interface buttons on the proportion of multimedia queries in the general query population, and (2) similarities and differences between Web multimedia and non-multimedia search queries. Major studies of Web user behavior are significant for the development of more effective multimedia IR systems.

RESEARCH GOALS

In this paper we report findings from a major study of trends in multimedia Web searching by Excite users from 1997 to 2001, including changes in queries and session characteristics, and changes or differences in image, video and audio searching. To assess any trends and changes in Web multimedia searching from 1997 to 2001 (Jansen, Spink & Saracevic, 2000), we analyzed a large data set of Excite Web multimedia queries from 2001 (Spink, et al., 2002) and compared the 2001 findings with results from previous studies of Excite multimedia searching from 1997 and 1999 (Jansen, Goodrum and Spink, 2000). In 2000 Excite also introduced three buttons onto the interface for users specifically wanted to find image, video or audio. In this study we examine the impact of these buttons on Excite users' multimedia searching.

RESEARCH DESIGN

Excite Data

Excite (<http://www.excite.com>) searches are based on the exact terms a user enters in a query. Capitalization is disregarded with the exception of logical commands AND, OR, and AND NOT. There is no stemming. The three Excite query logs data sets compared in this paper each consisted of 1,025,908 records (Table 1).

[Place Table 1 Here]

The Excite 2001 log consisted of the general web query log and we were not able to identify any queries entered against the multimedia buttons on the Excite interface. Each Excite query log record contained three fields: *Identification*: anonymous code assigned by Excite server to a user machine. *Time of Day*: in hours, minutes, and seconds. *Query*: user terms as entered. We analyzed user: *Sessions* - entire query sequence by a user; *Queries* - one or more entered terms; *Terms* - any string of characters bounded by white space.

Each data set was loaded into a database management application. Queries in this application were retrieved that contained multimedia terms. Specifically, there were:

- *Audio queries* – containing 27 audio related terms
- *Video queries* – containing 13 video-related terms
- *Image queries* – containing 30 image-related terms

Figure 1 shows the specific terms used in each query. The queries were case insensitive.

[Place Figure 1 Here]

The same multimedia terms were executed against each dataset for 1997, 1999 and 2001. If a user session contained a query not including any of these terms, that query would not appear in the analysis. As the information need it is difficult based on a single term, the result lists were reviewed, and the queries that were obviously not multimedia related were removed. However, when in doubt, the query was not removed from the result lists. We feel confident that majority of the queries in this analysis relate to multimedia searching. The queries were not altered in anyway.

RESULTS

We compare results from our 2001 findings with results from 1997 and 1999 for Excite multimedia searching, including: (1) proportion of multimedia queries, (2) session and query length, and structure, (3) mean query and session durations, (4) terms per query, (5) terms in multimedia queries, and (6) multimedia searching topics.

Multimedia User Sessions and Queries

Table 2 shows the number and percentage of multimedia queries identified in each data set for 1997, 1999 and 2001 data sets.

[Place Table 2 Here]

The percentage of multimedia queries from the Excite data log decreased from 3.7 % in 1997 and 1999 and 1.79 % in 2001. The percentage of multimedia queries as a proportion of the total queries, decreased by nearly 50% from 1997 to 2001. This result suggests that Excite users were conducting fewer multimedia searches on the general search engine in 2001 compared to 1999 and 1997, and were using the multimedia search buttons. This may also be due to the growth of specific multimedia search sites from 1997 to 2001 that cater to multimedia searching.

By 2001, the mean queries per session is 2.6 is slightly higher than the 2.2 mean queries per session for the entire 2001 dataset (Spink, et al., 2002). Identified multimedia sessions tend to have more queries compared to non-multimedia queries that are submitted to the Excite search engine. The mean multimedia queries per session decreased from 3.1 in 1997 and 3.2 in 1999 to 2.6 in 2001. By 2001, (1) the standard deviation for the mean queries per session was 4.6; depicting a wide spread for the queries per multimedia session, and (2) the median queries per session was 1 and the maximum queries per session was 119. This finding also supports the high standard deviation value. The median, maximum and standard deviation values for the entire multimedia queries are not available in the studies on 1997 and 1999 Excite datasets.

Table 3 compares audio, video and image queries for the 1997, 1999 and 2001 Excite data sets.

[Place Table 3 Here]

By 2001, audio queries dominate identified video and image queries, with 52.7% of multimedia queries being audio queries, 21.9% were video queries and 25.4% were image queries. This may be due to the development of mp3 technology and the “napster” software that allowed the free-exchange of audio files. It should be noted that the audio queries were not included in the 1999 dataset analysis.

Terms Per Multimedia Query

The mean terms per query for all identified multimedia queries, and audio, video and image queries separately for 1997, 1999 and 2001 datasets are presented in Table4.

[Place Table 4 Here]

By 2001, a total of 80,237 terms were used in 18,319 multimedia queries, the mean terms per query is 4.3 and the mean terms per query for the entire dataset is 1.4 (Spink, et al., 2002). The mean terms per query for multimedia searches was almost three times higher than for queries in the entire dataset. Multimedia query submitters do more complicated searches than the general search engine user, particularly longer music and video file names. Some data points for 1999 were not previously analyzed.

Table 5 shows the distribution of terms per query for the 2001 Excite dataset.

[Place Table 5 Here]

By 2001, out of 18,317 multimedia queries, only one query had a single search term, 2255 queries had two search terms, and 5772,3562 and 3002 queries had three, four or five terms per query, respectively. It should be noted that the mean terms per query was 1.4 for the entire Excite dataset (Spink, et al., 2002). Many queries had three, four or five terms per query.

Figure 2 shows the distribution of terms per query for less than ten terms per query and Figure 3 shows the distribution for more than ten queries.

[Place Figure 2 Here]

[Place Figure 3 Here]

Session and Query Duration Analysis

Table 6 shows the maximum, median and standard deviation of the mean duration in seconds for 2001 multimedia sessions and queries.

[Place Table 6 Here]

By 2001, the mean duration per session for multimedia queries was 1685.4 seconds and the mean duration per query was 469.9 seconds. The standard deviation for the duration per session was 6276.4 seconds (versus a mean duration per session value of 1685.4 seconds); demonstrating that the duration of sessions varies widely from one session to another. The median and the maximum value for the duration of sessions were 254 seconds and 85668 seconds, respectively. These figures also support the high standard deviation value for duration of sessions. The standard deviation value for duration of queries is even higher compared to the mean duration of queries. The standard deviation of duration per query was 3177.3 seconds versus a mean of 469.9 seconds for mean duration per query, showing a higher fluctuation in duration of queries. This can also be observed by looking at the difference between the median and maximum values for duration of queries. The median for duration of queries is 42 seconds, which is quite low compared to the median for duration of sessions. However, the maximum duration for a query was 85431 seconds, which is quite close to the maximum value for duration of sessions.

The duration per multimedia session and duration per query for audio and video queries are provided in Table 7.

[Place Table 7 Here]

By 2001, the duration per session for audio, video and image queries were 1882.1, 1521.4 and 1410.7 seconds respectively. The duration per query for audio and video queries were 536.6, 428.7 and 367.5 seconds respectively. Both the mean duration of sessions and queries are less for video and image search queries than they are for audio search queries. Web users seem to spend more time on audio queries than on video and image queries. Consequently, the mean duration of sessions and queries are slightly higher for audio queries compared to the mean duration for all the multimedia queries and slightly lower for video queries compared to the mean duration for all the multimedia queries.

Top Multimedia Terms

We analyzed the five most frequently occurring terms in audio, video and image queries for 1997, 1999 and 2001 (Table 8).

[Place Table 8 Here]

By 2001, the most frequently used terms for all types of multimedia queries in by Excite users was “AND”. Audio query term rankings were not available for 1999. Multimedia queries were more complicated than non-multimedia queries due to the greater use of the Boolean operator “AND”. Another interesting finding is that the main elements for audio and video queries were to follow “AND”, whereas the main search elements for image queries were at lower ranks of the top term listing. The main elements for audio searches, i.e. music, mp3 and sounds have the 2nd through 4th spot on the list of top terms and the main elements for video searches, i.e. video and movie take the 2nd and 3rd spots. However, the main elements for image searches, i.e. pics and photo took 4th and 6th spots on the list of top terms. In addition “free” was a popular term in all three types of queries. In the 1997 and 1999 Excite datasets, “AND” is not one of the top ten terms.

DISCUSSION

How did multimedia Web searching by Excite users change from 1997 to 2001? Multimedia searches decreased as a proportion of general queries as more specialized multimedia search buttons and search engines are becoming available. A major reason for a lower percentage of multimedia queries in 2001 was the introduction of image, audio and video search buttons on the Excite interface, and the increase in multimedia search queries that do not contain multimedia identifying terms. For example, a user searching for a song may enter the name of the song but may not include the term “song” in the query when using the audio search button. In this case, the query will not be entitled as a multimedia query, although in fact it is a multimedia query.

For multimedia searches identified in the general query set, our analysis shows a trend towards less and shorter multimedia searching sessions with less query modification, but longer multimedia queries, more audio queries, a greater use of file names as query terms, and greater use of non-topic terms such as ‘AND’ in multimedia queries.

The increasing number of terms per query signal more complex multimedia searches in 2001. Multimedia queries are almost three times more complex than non-multimedia search queries. The first result derived from this analysis shows no major change in the mean audio, video and image queries compared to all the multimedia queries for the 2001 Excite dataset. The mean terms per query for all types of queries are between 4 and 5 terms per query.

Even though, the total terms for video queries is considerably less compared to the total number of terms for audio and image queries; the mean value is quite close in all three types of multimedia queries. The mean terms per query increased from 1997 and 1999 to 2001. Overall the Web users tend to make more complicated multimedia searches with more terms in a single query.

Users tend to spend more time on audio queries compared to video and image queries. By 2001, the interest in audio queries has increased, whereas the interest on video and image queries has decreased compared to 1997 and 1999. The analysis suggests that audio queries in general are

more prevalent than video and image queries. Other significant changes in the list of top terms is also interesting, especially for audio queries. “mp3” which was not in the list of top terms in 1997 and 1999 was the 3rd mostly used term in 2001, which is a logical result since mp3 is a new technology that has recently been very popular. In 1997 and 1999 “cd” was one of the top terms and not in the top terms in 2001. This shows that Web users do not search for “cd”s on the Web any more, instead they search for “mp3”s. Hence, Web users are more knowledgeable and aware of the jargon for Web services and know that they can reach music over the Internet through “mp3”s. This finding shows the growth in music sharing over the Internet from 1999 to 2001.

CONCLUSION

In conclusion, our study shows trends in multimedia Web searching. The most interesting finding was the impact of multimedia buttons on Excite users searching for multimedia. Despite the presence of the multimedia buttons, many users didn't use the buttons. Maybe they preferred to use the general search box. Multimedia queries identified are still longer than non-multimedia queries, including terms such as “pictures” or “image”. However, users are now searching less for images than in earlier data sets, and more for audio files as more audio and music becomes available over the Web. Overall, multimedia Web searching is undergoing major changes as Web content and searching evolves. To investigate these issues further, we are currently conducting an analysis of large-scale query data from another Web search engine to compare with the Excite findings. Our current analysis also includes an analysis of the click-through data for multimedia searches to determine the Web sites accessed.

REFERENCES

- Aslandogan, Y., Thier, C., Yu, C., Zou, J., & Rishe, N. (1997). Using semantic contents and WordNet in image retrieval. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Publications, pp. 286-295.
- Brown, M., Foote, J., Jones, G., Sparck Jones, K., & Young, S. (1996). Open-vocabulary speech indexing for voice and video mail retrieval. *Proceedings of the Fourth ACM International Multimedia Conference, ACM Multimedia '96*, pp. 307-316.
- Goodrum, A., & Kim, C. (1998). Visualizing the history of chemistry: Queries to the CHF pictorial collection. *Report to the Chemical Heritage Foundation Pictorial Collection*,
- Goodrum, A., & Spink, A. (2001). Image searching on the Excite Web search engine. *Information Processing and Management*, 37, 295-311.
- Huberman, B. A., Pirolli, P. L. T., Pitkow, J. E., & Lukose, R. M. (1998). Strong regularities in World Wide Web surfing. *Science*, 280 (April 3), 95-97.
- Jansen, B. J., Goodrum, A., & Spink, A. (2000). Searching for multimedia: Analysis of audio, video and image Web queries. *World Wide Web*, 3, 249-254.
- 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.
- Lawrence, S., & Giles, C. L. (1999). Accessibility information on the Web. *Nature*, 400, 107-109.
- Montgomery, A., & Faloutsos, C. (2001). Identifying web browsing trends and patterns. *IEEE Computer* (July), 94-95.
- Silverman, C., Henzinger, M., Marais, H., & Morris, M. (1999). Analysis of a very large Web search engine query log. *ACM SIGIR Forum* 33(3).

Smith, T., Ruocco, A., & Jansen, B. J. (1998). Digital video in education. *Proceedings of the Thirteenth SIGCSE Technical Symposium on Computer Science Education*, pp. 122-126.

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *IEEE Computer*, 35(3), 107-109.

Spink, A., Wolfram, D., Jansen, B. J., & Saracevic, T. (2001). Searching the Web: The public and their queries. *Journal of the American Society for Information Science*, 53(2), 226-234.

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 Sciences and Technology*, 52(12), 1073-1074.

Table 1. Excite data sets: 1997, 1999 and 2001.

1997 Excite Data Set (Jansen, Spink & Saracevic, 2000; Goodrum & Spink, 2000)	1999 Excite Data Set (Jansen, Spink & Saracevic, 2000; Goodrum & Spink, 2000)	2001 Excite Data Set (Spink, Jansen, Wolfram & Saracevic, 2002)
1,025,908 queries Date of Data Set: September 16, 1997	1,025,910 queries Date of Data Set: December 20, 1999	1,025,910 queries Date of Data Set: May 4, 2001

Figure 1. Multimedia terms.

Audio Terms	Image Terms	Video Terms
au	art '	.avi
.au	bitmap	.mjpeg
audio	Bmp	.mov
av	.bitmap	.mov8
.av	.bmp	.mpeg
band	camera	.mpg
cd	cartoon	animated
concerts	gallery	clip
lyrics	gif	clips
mpz	.gif	drivers
multimedia '	image	mjpeg
music	images	mov
noise	jpeg	movie
song	jpg	movies
songs	pcx	mpeg
sonic	.jpeg	mpg
sonics	.jpg	plugins
sound	.pcx	quicktime
sound card	photo	video
sound cards	photographs	viewers
soundblaster	photograph	avi
sounds	photos	
soundwave	pic	
speakers	pics	
track	.pic	
vocals	.pics	
wav	picture	
.wav	pictures	
	png	
	.png	
	tif	
	tiff	
	.tif .tiff	

Table 2: Number and percentage of multimedia queries for 2001, 1999 and 1997 Excite datasets.

Variables	1997	1999	2001
Number of Multimedia Queries in Each Data Set	38,584	39,338	18,317
Percentage of All Queries	3.7%	3.7%	1.7%
Mean Queries Per Multimedia Session	3.1	3.2	2.6
Median Queries Per Multimedia Session	-	-	1
Maximum Queries Per Multimedia Session	-	-	119
Standard Deviation of Queries Per Multimedia Session	-	-	4.6

Table 3: Comparison of audio and video queries from 1997, 1999 and 2001 Excite datasets.

Number	Audio Queries			Video Queries			Image Queries		
	1997	1999	2001	1997	1999	2001	1997	1999	2001
	3810		9655	7630	17148	4011	27144	22190	4651
Percentage of Data Set	0.37%	-	0.9%	0.7%	1.6%	0.3%	2.6%	2.1%	0.4%
Percentage of Multimedia Queries	9.8%	-	52.7%	19.8%	43.5%	21.9%	70.4%	56.4%	25.4%
Mean Queries Per Session	2.4	-	2.6	2.9	3	2.6	3.2	3.4	2.8
Median Queries Per Session	2	-	1	2	-	1	2	-	1
Maximum Queries Per Session	51	-	119	70	-	59	267	-	102
Standard Deviation for Queries Per Session	2.9	-	4.8	3.8	-	3.8	5.4	-	5

Table 4: Total and mean number of terms per query for all multimedia queries, audio, video and image queries for the 1997, 1999 and 2001 Excite data sets.

	Multimedia Queries			Audio Queries			Video Queries			Image Queries		
	1997	1999	2001	1997	1999	2001	1997	1999	2001	1997	1999	2001
Total Terms	134022	-	80237	15661	-	39269	24514	-	17863	93847	-	23105
Mean Terms Per Query	3.5	-	4.3	4.1	-	4.4	3.3	3	4.4	3.4	3.5	4.9

Table 5: Distribution terms per query with respect to queries

Terms Per Query	Audio Terms	Video Terms	Image Terms	Total Number of Queries
67	0	0	1	1
57	0	0	1	1
29	1	0	0	1
28	0	0	0	0
27	0	0	0	0
26	0	0	0	0
25	0	0	0	0
24	1	0	0	1
23	0	1	0	1
22	2	0	0	2
21	2	0	0	2
20	0	0	1	1
19	1	2	4	7
18	0	4	1	5
17	5	2	2	9
16	4	5	2	11
15	25	12	10	47
14	6	3	5	14
13	36	10	25	71
12	20	3	23	46
11	75	51	36	162
10	55	18	128	201
9	268	60	95	423
8	150	61	154	365
7	614	346	316	1276
6	391	238	449	1078
5	1324	672	1006	3002
4	1597	853	1112	3562
3	3002	1497	1273	5772
2	2075	173	7	2255
1	1	0	0	1
Total	9655	4011	4651	18317

Figure 2: Distribution of queries with less than ten terms per query.

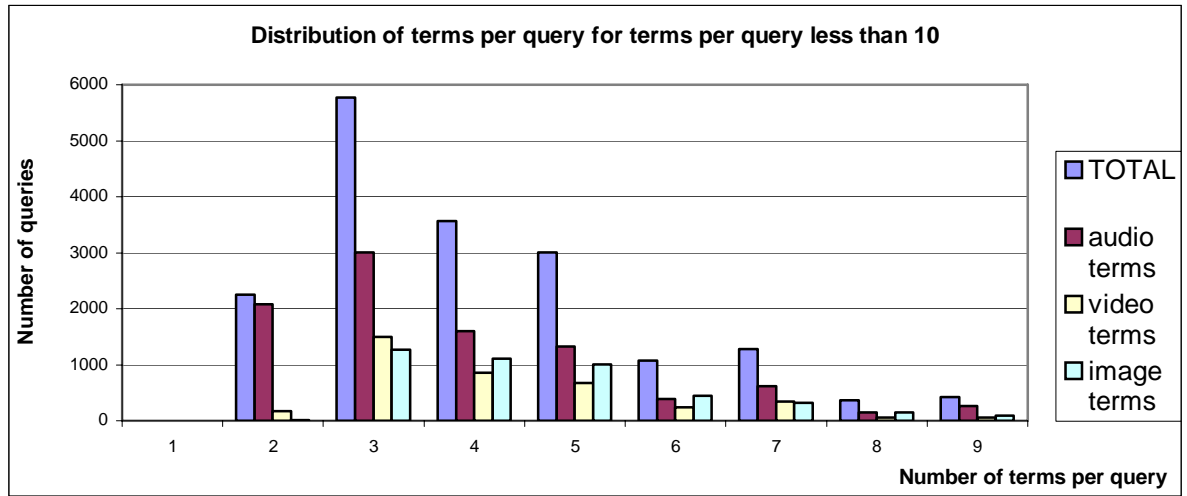


Figure 3: Distribution of queries with more than and equal to ten terms/query

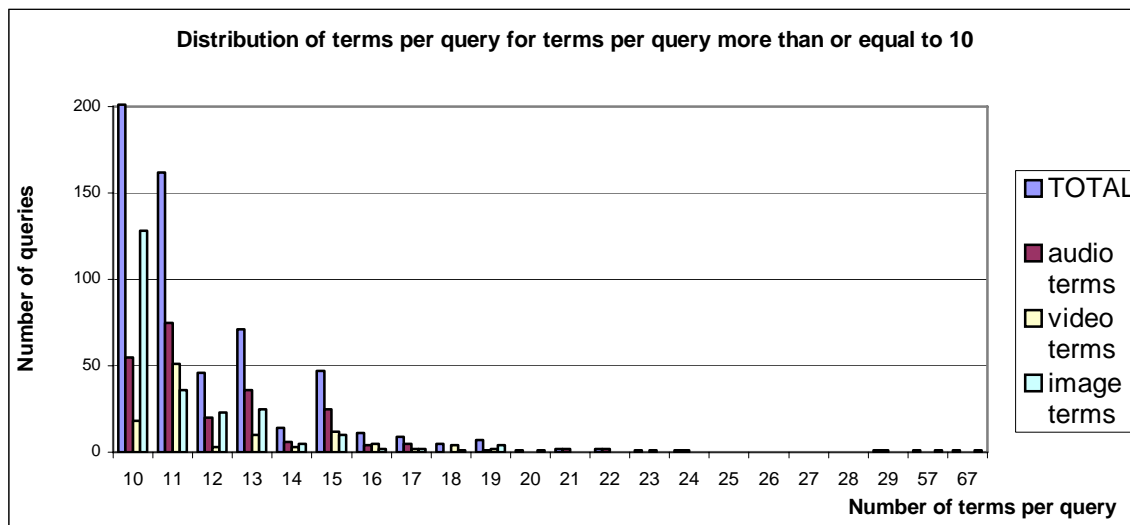


Table 6: Mean duration of multimedia sessions and queries in seconds for the 2001 Excite dataset

	Sessions (Seconds)	Queries (Seconds)
Mean Duration	1685.4	469.9
Median of Durations Per Multimedia Session	254	42
Maximum Duration Per Multimedia Session	85668	85431
Standard Deviation of Duration Per Multimedia Session	6276.4	3177.3

Table 7: Mean duration of sessions and queries for audio and video queries from 2001Excite dataset.

	Audio Queries (Seconds)	Video Queries (Seconds)	Image Queries (Seconds)
Mean Duration of Sessions	1885.1	1521.4	1410.7
Mean Duration of Queries	536.6	428.7	367.5

Table 8: List of the five most frequently used terms in audio, video and image queries for 1997, 1999 and 2001.

1997 Ranking	Audio Query Terms	Video Query Terms	Image Query Terms
1	music	movies	pictures
2	sound	video	photos
3	audio	movie	pictures
4	lyrics	videos	pics
5	cd	clips	photo
1999 Ranking			
1	-	video	pics
2	-	free	free
3	-	movie	of
4	-	clip	art
5	-	movies	nude
2001 Ranking			
1	AND	AND	AND
2	music	video	free
3	Mp3	movie	art
4	sound	free	pics
5	songs	MPEG	of