Multimedia Web Searching Trends

Seda Ozmutlu
Dept. of Industrial Engineering, Uludag University, Gorukle Kampusu, Bursa, 16059, Turkey. Email:
seda@uludag.edu.tr

Amanda Spink
School of Information Sciences and Technology, The Pennsylvania State University, 004C Thomas
Building, University Park PA 16802.
Email: spink@ist.psu.edu

H. Cenk Ozmutlu
Dept. of Industrial Engineering, Uludag University, Gorukle Kampusu, Bursa, 16059, Turkey. Email:
hco@uludag.edu

In this paper we examine and compare multimedia Web searching by Excite and FAST search engine users in 2001. Findings include: (1) some differences between the Excite and FAST multimedia Web searches, (2) FAST users submit more multimedia queries but fewer audio-related queries, (3) FAST users take more time with queries and sessions compared to Excite users, but spend less time on audio queries, and (4) Excite users submit longer and more complex queries than FAST users. As the Excite search engine is U.S. based and FAST search engine is Norwegian based, these differences suggest some difference in the Web search behavior of U.S. and European Web users.

Introduction

This paper reports findings on Web multimedia searching by Excite and FAST search engines in 2001. The Web query datasets examined during our study were quite large, each with over 1 million queries. Analyzing these query data sets enabled the comparison of multimedia searching behavior. Query and session arrival and duration analyses, and terms analysis are provided. The study also provides insights into the differences between image, video and audio queries as different types of multimedia queries.

Previous studies have analyzed aspects of multimedia searching (Aslandogan, et al., 1997; Brown, et al., 1996; Smith, et al., 1998). Jansen, Goodrum and Spink (2000) studied multimedia Web searching using 1997 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.

Spink, Goodrum and Hurson (2001) examined trends in Excite video and image queries from 1997 to 1999. They found: (1) video and image query formed fewer than 5% of all queries, (2) multimedia query length changed little from 1997 to 1999, and (3) few multimedia queries contained terms such as 'avi', 'mpeg' or 'gif' to delineate multimedia files.

The current study extends previous research by comparing multimedia searching on two different Web search engines.

Research Design

We analyzed query data provided by the Excite Web search engine (htpp://www.excite.com) from May 4,2001and the FAST (htpp://www.FAST.com) Web search engines from February 6, 2001. In both the Excite and FAST query log structure, the entries are given in the order they arrive. New sessions/users are identified through an anonymous user ID and each query is given time stamps in hours, minutes and seconds.

Previous studies (Spink, at al., 1999) show that approximately 84% of Excite users were located in the United States. The majority of the FAST users are believed to be from Europe, mostly from Germany. All the data sets are quite large, each with over one million queries. The specifications of the datasets are in Table 1.

Table 1. Excite and FAST query sets 2001.

Excite 2001	FAST 2001
1,025,910 Date of Query Set Analyzed: May 4, 2001	1,257,891 Date of Query Set Analyzed: February 6, 2001

The multimedia queries were identified in each query logs. Each data set was loaded into a database management application. Both Web search engines interfaces included multimedia search buttons by 2001. Queries in this application were retrieved if they contained multimedia terms. We had no access to identifying the queries run against each multimedia button. Specifically, there were: *Audio queries* – containing 27 audio related terms, *Video queries* – containing 13 video-related terms, and *Image queries* – containing 30 image-related terms. We analyzed: (1) session and query length and structure, (2) mean query and session durations, (3) the number of search terms per query, and (4) terms in multimedia queries.

Results

Multimedia User Session and Query Arrival

The general arrival specifications of Excite and FAST multimedia queries and sessions are shown in Table 2. Multimedia queries form close to 6% of all the 2001 FAST queries. The mean query per session of 2.6 appears to be on the decline for Excite users. The arrival characteristics of the FAST queries seem to conform to those of the Excite queries with 2.2 mean queries per multimedia session. Additional insight is gained by looking separately at the audio, video and image queries. Table 3 compares audio, video and image queries for the Excite and FAST queries.

Table 2: Number and percentage of multimedia queries for the 2001 Excite and FAST.

Variables	Excite 2001	FAST 2001
Number of multimedia queries	18,317	70,697
Percentage of all queries	1.7%	5.6%
Mean queries per multimedia session	2.6	2.2
Mean queries per session for entire dataset	2.3	8.1

Table 3: Comparison of audio and video queries from Excite and FAST datasets.

	Excite Audio Queries 2001	FAST Audio Queries 2001	Excite Video Queries2001	FAST Video Queries 2001	Excite Image Queries 2001	FAST Image Queries 2001
Number	9655	7498	4011	29580	4651	33619
Percent of Entire Database	0.9%	0.6%	0.3%	2.4%	0.4%	2.7%
Percent of Multimedia Queries	52.7%	10.6%	21.9%	41.8%	25.4%	47.6%
Mean Queries Per Session	2.6	2.2	2.6	1.8	2.8	2.7

Audio queries dominate video and image queries in the Excite queries. Some 52.7% of Excite multimedia queries were audio queries, 21.9% were video queries and 25.4% were image queries. The distribution of audio, video and image queries for the Excite was not similar to FAST. The FAST audio queries are the least popular type of multimedia queries forming only 10.6% of all multimedia queries, whereas image queries are the dominant type of multimedia queries forming 47.6% of all multimedia queries. Video queries are also quite prevalent within FAST multimedia searches forming 41.8% of multimedia searches.

Multimedia Sessions and Query Duration Analysis
Table 4 shows the mean duration in seconds for the Excite and FAST multimedia sessions and queries. Both FAST sessions and queries tend to be longer than Excite sessions and queries. Although FAST users submit about the same number of queries per session; the query durations are longer, making the session durations longer. The duration

per multimedia session and per query for audio, video and image queries are provided in Table 5. For Excite, both the mean duration of sessions and queries are shorter for video and image search queries than 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. However, FAST users seem to spend less time on audio queries and their sessions than video queries and their session, but more time than image related queries and their sessions.

Terms Per Multimedia Query
The mean terms per query for all multimedia queries, and audio, video and image queries for Excite and FAST datasets are shown in Table 6.

Table 4: Mean duration of multimedia sessions and queries in seconds for Excite and FAST queries.

	Mean Duration	
	Sessions (Seconds)	Queries (Seconds)
Excite 2001	1685.4	469.9
FAST 2001	7940.2	1746.7

Table 5: Mean duration of sessions and queries for audio, video and image Excite and FAST queries.

	Mean Session Duration			Mean Query Duration		
	Audio Queries (Seconds)	Video Queries (Seconds)	Image Queries (Seconds)	Audio Queries (Seconds)	Video Queries (Seconds)	Image Queries (Seconds)
Excite 2001	1885.1	1521.4	1410.7	536.6	428.7	367.5
FAST 2001	5071.9	8283.7	8259.8	1447.8	2270.3	1352.6

Table 6: Total and mean terms per query for all multimedia queries, audio, video and image Excite and FAST queries.

	Excite Multimedia 2001	FAST Multimedia 2001	Excite Audio 2001	FAST Audio 2001	Excite Video 2001	FAST Video 2001	Excite Image 2001	FAST Image 2001
Total terms	80237	207752	39269	25753	17863	82763	23105	99236
Mean terms per query	4.4	2.9	4.4	3.4	4.4	2.8	4.9	3.0

In 2001 the mean terms per query for Excite multimedia searches is 4.4, and the mean terms per query for the entire dataset is 2.6. The mean terms per multimedia query were 50% higher than for queries in the entire dataset. Users who submit multimedia queries do more complicated searches than the general search engine user. This may be a result of entering long music and video file names in multimedia queries. For the FAST queries, we see 2.9 terms per multimedia query, which is lower than for the Excite queries.

In the FAST dataset, the mean number of terms per query is 2.4. The mean terms per multimedia queries in the FAST dataset is slightly different than the mean terms per query of the entire dataset; about 20% higher. In addition, overall FAST users submit shorter queries than Excite users. This could be due to the fact that Excite users submitting more audio queries as mentioned before. Audio queries include many song names, which tend to be long. Therefore fewer audio queries submitted by FAST users might cause a less mean terms per query value.

Figures 2 and 3 show the distribution of terms per query for Excite and FAST datasets for comparison. Although the scales for the graphs are different, it is apparent that the trends in both datasets are quite similar except more single term queries being used in the FAST dataset.
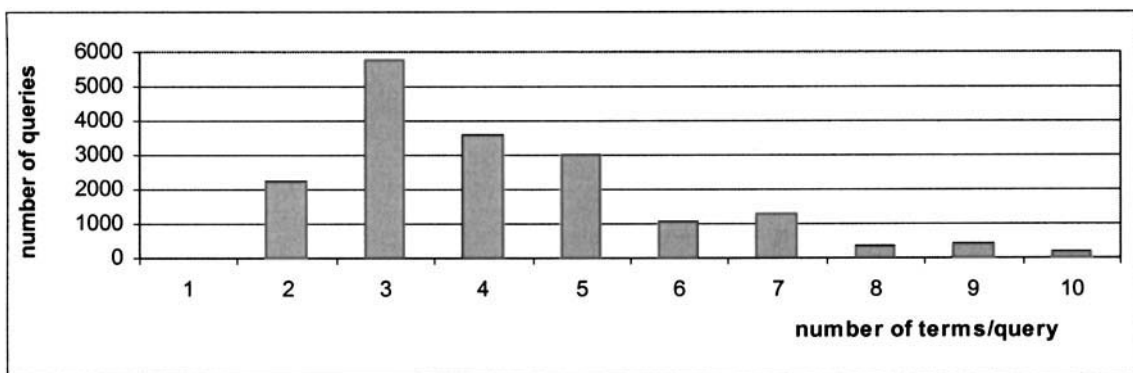
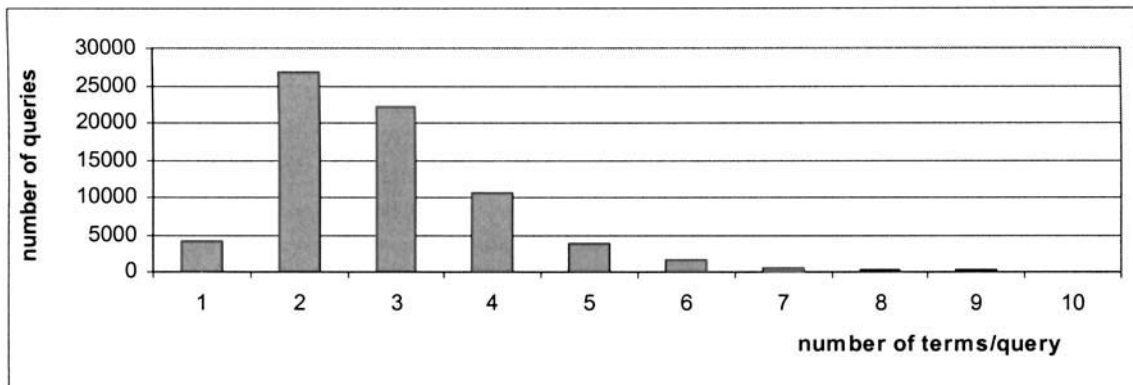Figure 2: Distribution of queries with fewer than or equal to 10 terms/query for the Excite dataset.

Figure 3: Distribution of queries with fewer than or equal to 10 terms/query for the FAST dataset.

In both query sets, many queries had three, four or five terms per query. For the Excite dataset, 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. In the FAST dataset, the number of queries having a single term was higher, 4062. For the FAST dataset, out of 207,752 multimedia queries, 26,600 queries had two search terms, and 22176,10,653 and 3875 queries had three, four or five terms per query, respectively.

Top Multimedia Terms

We analyzed the five most frequently occurring terms in audio, video and image queries for the Excite and FAST datasets (Table 8). The most frequently used terms for all types of multimedia queries in the Excite dataset is "AND". The term "AND" is used to make more complicated searches, therefore cause longer query sizes. 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 spots 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 FAST dataset, "AND" is not one of the top five terms that could be the reason that FAST multimedia searches are shorter than multimedia searches. In addition, in the previous sections we suggested that the audio queries might not be as popular within the FAST users due to the fact that mp3 sharing might not be as prevalent in Europe as it was in the US. This suggestion is supported by the list of top terms. Although mp3 is among the top audio terms in the Excite queries, it is not among the top terms for FAST.

The top multimedia terms for the FAST queries mostly consisted of main elements for multimedia searching like "pictures", "music", "video", etc. Free was also a popular word for FAST multimedia Web searchers as it was for Excite multimedia Web searchers.

Table 8: List of the five most frequently used terms in audio, video and image queries for Excite and FAST datasets.

Excite 2001 Ranking	Audio Terms	Video Terms	Image Terms
1	AND	AND	AND
2	Music	Video	free
3	Mp3	Movie	art
4	Sound	Free	pics
5	Songs	MPEG	of
FAST 2001 ranking			
1	Music	Movie	pictures
2	Sound	Movies	jpg
3	Wav	Free	pics
4	Songs	Video	photos
5	Lyrics	Mpeg	free

Discussion

The characteristics of Excite and FAST query logs showed various similarities and differences. Overall, our analysis shows that Excite multimedia queries tend to get larger over time with more search terms and tend to be larger than FAST multimedia search queries. Excite searches are also more complex than FAST searches. However, FAST search engine users submit more multimedia queries than Excite users, and spend more time on queries. Although the Excite users seem to lose interest in multimedia Web material, the

FAST users seem to maintain a significant level of interest, even higher than the level of interest by Excite users in 1997 and 1999. This result may be due to the growth of specific multimedia search sites in USA from 1997 to 2001 to satisfy multimedia needs. Such sites might not be as widely used in Europe as in the US. This finding may also be due to widely known and used multimedia exchange Web sites such as Napster.

FAST queries provide similar results to the Excite queries, with 2.2 as the mean number of queries per session. One interesting difference between the Excite and FAST queries

is the comparison of mean queries for multimedia queries and the entire datasets. For the Excite queries, the mean multimedia queries per session have always been higher than that of entire query logs. However, this is not true for the FAST data set. The mean queries per session for the multimedia queries are much fewer than the mean number of queries for the entire data log.

The difference in the mean queries between regular sessions and multimedia sessions of FAST can be interpreted in different ways. For example, the difference may be due to high number of Web sites with multimedia content, or the multimedia topics can be found in popular Web sites that are highly rated by Web search engines. Another reason can simply be that regular FAST users are more patient in looking to the results compared to the FAST users searching for multimedia topic. However, the correct reason can be found only by interviewing the users.

Audio queries dominate video and image queries in the Excite query log. In 2001, 52.7% of multimedia queries were audio queries, the rest image and video queries. This may be due to the recent development of mp3 technology and the "napster", and other similar software that allowed the free-exchange of audio files. The number of Web sites providing audio files increased from 2000 to 2001. It should be noted that the audio queries were not included in the 1999 dataset analysis. The distribution of audio, video and image queries are also quite different for the Excite and FAST search engines. For the 2001 FAST dataset, the audio queries are the least popular type of multimedia queries (10.6% of multimedia queries). Image queries (47.6% of multimedia queries) and video queries (41.8% of multimedia queries) are submitted much more than audio queries. This result might indicate that mp3 file sharing might not be as popular in Europe as it is in the US or the knowledge of mp3 sharing software is not widely known for FAST users. From the findings, we can conclude that visual multimedia Web searching is more prevalent in Europe, and audio multimedia Web searching is more dominant in the US.

FAST multimedia sessions and queries were longer then Excite multimedia sessions and queries. The longer duration per query implies that FAST users may spend more time to evaluate the results of their queries compared to the time spent by the Excite users. Excite users may spend more time on audio queries than on video and image queries compared to FAST users. FAST users seem to spend more time on image queries. Recall that in the FAST 2001 query log, audio queries formed a small part of entire multimedia queries. For FAST users, the interest in audio queries is really diminutive both in terms of number of queries submitted and the duration of queries and sessions. FAST users submit fewer audio queries and when they do, they spend less time on them. Although FAST users submit a significant number of image queries, these queries do not take much time.

FAST multimedia searches have fewer terms per query than Excite searches, meaning that the Excite searches are longer. The distribution of the terms per query for both search engine users is similar, with two, three, four or five term queries dominating the dataset. Users submitted more single query terms than Excite users. The finding that the Excite searches are longer is supported by the frequent use of the term "AND" in the Excite searches.

Conclusion

Excite multimedia searches were longer and more complex, and FAST users submitted more multimedia queries taking a longer time. These results suggest some potential differences in the multimedia Web searching of U.S. and European users. Our further research is exploring changes in Web multimedia searching from 1997 to 2001, and investigating regional differences.

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*, 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.

Henzinger, M. R. (2001). Hyperlink analysis for the Web. *IEEE Internet Computing*, 45-50.

Jansen, B. J., Goodrum, A., & Spink, A. (2000). Searching for multimedia: Analysis of audio, video and image Web queries. *World Wide Web Journal*, 3, 249-254.

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., Bateman, J., & Jansen, B. J. (1999). Searching the Web: Survey of Excite users. Internet Research: Electronic Networking Applications and Policy, 9(2), 117-128.

Spink, A., Goodrum, A., & Hurson, A. R. (2001). Multimedia Web queries: Implications for design. Proceedings of the 2001 *International Conference on Information Technology: Coding and Computing, April 2-4, Las Vegas, NV.*