Journal of Web Engineering, Vol.3, No.3&4 (2004) 182-199

THE EFFECT OF SPECIALIZED MULTIMEDIA COLLECTIONS ON WEB SEARCHING

BERNARD J. JANSEN
The Pennsylvania State University
jjansen@ist.psu.edu

AMANDA SPINK
University of Pittsburgh
aspink@mail.sis.pitt.edu

JAN PEDERSEN
Overture Web Services Division
jan.pedersen@av.com

Received July 8, 2004
Revised August 16, 2004

Multimedia Web searching is a significant information activity for many people. Major Web search engines are critical resources in people's efforts to locate relevant online multimedia information. It is therefore important that we understand how searchers are utilizing these Web information systems in their quest to retrieve multimedia information to design effective Web systems in support of these information needs. In this paper, we report the results of a research study evaluating the effect of separate multimedia Web collections on individual searching behavior. The AltaVista search engine has an extensive multimedia collection and uses tabs to search specific collections. The motivating questions for this research are: (1) What are the characteristics of multimedia searching on AltaVista? and (2) What are the effects on Web searching of separate multimedia collections? The results of our research show that multimedia searching is complex relative to general Web searching. Searching specific multimedia collections has reduced the complexity of audio searching, but it has not had the same effect for image and video searching. Query length and Boolean usage rates are much higher for image searching, compared to general Web searching. We discuss the implications of the research findings for the design, development and evaluation of Web multimedia retrieval systems.

Key words: Web searching, multimedia searching, audio, image, video, search engines
Communicated by: C Watters

1. Introduction

Web searching is a daily activity for many people, with the Web now the first choice for many seeking online information [1]. Most major search engines support some type of multimedia searching, and there are several multimedia specific search engines, including ImageScape [2], WebSEEK [3] and SingingFish (http://www.singingfish.com/). The estimated number of images on the publicly indexed Web is several hundred million [4], with millions more being added daily. When one includes video and audio files (i.e., songs, movies, and animated computer files), it is clear that the Web is a vast

multimedia repository. As more people and organizations store multimedia objects on the Web, the searching and retrieval of multimedia has become a major challenge for researchers, commercial practitioners, and recreational users alike. The design of multimedia Web systems to support searching of these collections is a complex and fluid task, as is the development of most Web-based systems [5]

The tremendous growth in the quantity of multimedia content is driving the need for more effective methods of storing, searching, and retrieving of this multimedia data. How users search for multimedia on the Web and the design of more effective Web multimedia retrieval systems are growing areas of research. Given the Web's importance, we need to understand how people use and interact with Web search engines in locating multimedia information. Examining Web multimedia searching is an important area of research with the potential to increase our understanding of multimedia searching in general, advance our knowledge of user information needs in this area, and positively impact the design of future online multimedia systems. This understanding will assist in addressing many challenges of multimedia retrieval [6].

We concentrate on these research needs in the present investigation by examining Web searchers using AltaVista (http://www.altavista.com), a major U.S. Web search engine. We examine the effect of multimedia radio buttons, an innovation in multimedia searching, and one that researchers have not previously studied. We analyze multimedia searching characteristics, including session duration, query length, results pages viewed and term usage. In the following sections, we describe our research design and our analysis of data from an AltaVista Web search engine transaction log, along with discussion of results. We then present the key findings and the implications of our research results for Web multimedia system users and system designers. We conclude with directions for future research.

2. What's So Hard About Web Search

One can classify current multimedia retrieval approaches either concept, content based [7], or a mixture of the two [8]. In the concept-based approach, image retrieval research focuses on the retrieval of multimedia objects utilizing indexed collections relying on textual attributes. Practitioners and researchers have created thesauri for visual information, such as the Library of Congress Thesaurus for Graphics Materials, Metadata [9]. Content providers then use these thesauri to index images within a collection. Many Web search engines use textual clues on Web documents to automate this concept-based approach. These Web search engines use text surrounding the multimedia object, along with other clues such as files names of multimedia content. This approach leverages the assumption that these textual clues relate to the object. Although many times valid, this assumption does not always hold. For example, software programs for desktop computers and digital cameras will automatically generate file names for multimedia objects that are random character sequences. However, this approach has proven effective.

The content-based approach focuses on indexing multimedia objects at the pixel level and the implementation of search features using pixel comparison [10]. Content-based systems allow users to search multimedia collections using color, texture, shape, and spatial similarity, among others. Wang's [10] SIMPLIcity multimedia retrieval system (http://ist.psu.edu/research/Index2.cfm?pageID=162) uses this approach. These systems many times also provide text-based search functions for notations and text descriptions embedded within multimedia objects. New technologies have focused much emphasis on content-based retrieval, with commercial systems such as the MediaSite.com

(http://www.mediasite.com) system. On the Web, content image retrieval systems such as WebSEEK [3] and SingingFish (http://www.singingfish.com/) provide a variety of multimedia files to Web searches.

It is not clear how or if the retrieval functionality of either of these approaches systems correlates with the multimedia needs of real users. Concerning the context approach, it has been pointed out that experts are not a good source of terms that are preferred by real users [11], with different searchers using a variety of terms for the same concept or item. Nearby textual clues many times are not related to the multimedia object, as noted with the objects created by the use of digital cameras. Concerning the content approach, users seldom search using content characteristics [12]. Although user studies within specific domains have been conducted [c.f., 13, 14], research [15] shows that Web users differ in their interaction with information retrieval (IR) systems relative to more traditional systems.

Researchers such as Jorgensen [16] and Smeaton [17] review a number of unique systems for image classification. However, little research has examined the relative effectiveness of these various approaches to image indexing or retrieval using Web search engines. There have been studies investigating the automatic assignment of textual attributes using captions from still images, transcripts, closed captioning, and verbal description for the blind accompanying videos [18]. Swain [19] views text cues extracted from Web pages and multimedia document headers, supplemented by off-line analysis, to be the primary sources of information for indexing multimedia Web documents. However, Lawrence and Giles [4] report that the use of Web metadata tags is still not widespread.

There has been a limited number of large-scale research examining Web multimedia searching [20, 21] from a variety of search engines. Although the limited studies conducted do provide important insights into multimedia Web searching, further research is needed to validate these results across search engines and over time. This is especially important as Web information systems are continually undergoing incremental and evolutionary changes. Additionally, the multimedia content and indexing is also undergoing changes. Research is therefore needed to evaluate the effect of these changes on system performance and user searching behaviors.

There is a growing body of Web research examining the use of search engines [15, 22, 23]. In a review of the Web searching literature, Jansen and Pooch [15] compare Web searchers with searchers of traditional IR systems and online public access catalogues. The researchers report that Web searchers exhibit different search characteristics than do searchers of other information systems. Spink, Jansen, Wolfram and Saracevic [22] provide a four year analysis of searching on the Excite search engine. They report a shift from entertainment to commercial searching; otherwise, Web searching has remained relatively stable over time. The researchers note that on the Excite search engine Web searching sessions are short as measured by number of queries. Spink and Jansen [24] report that session duration is also short, with the typical Web session being about 15 minutes. Users view a very limited number of results pages, with the majority of Web searchers, approximately 80%, viewing no more than 10 to 20 Web documents. These characteristics have remained fairly constant across multiple studies. Silverstein and fellow researchers [23] analyze a large number of AltaVista queries. They report that sessions consist of few queries and that queries consist of few terms. These studies did not specifically address multimedia searching.

There has been significantly less published research in the area of multimedia Web searching [20, 21, 25]. Using data from the Excite search engine, Goodrum and Spink [20] analyzed image queries. Also from the Excite search engine, Jansen, Goodrum, and Spink [21] analyzed audio, image, and video sessions and queries. Goodrum and Spink [20] 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% of the time, with most terms occurring only once. Jansen, Goodrum, and Spink [21] note that multimedia sessions and queries are generally longer than general Web queries indicating an increased cognitive load for multimedia searching. Audio queries were longer than image or video queries. Both of these studies were on the general database and did not examine the effect of separate multimedia collections on Web searching.

In 2000, Excite introduced three buttons onto the interface of the main Web site for users who specifically wanted to search for image, video or audio content. Ozmutlu, Spink, Ozmutlu [25] examined the impact of multimedia interface buttons on the proportion of multimedia queries in the general query population, and contrasted Web multimedia and non-multimedia search queries. The researchers state that the use of radio buttons had decreased the multimedia searches in the general collection. None of these studies examined queries from separate multimedia collections.

The design of Web information systems is certainly a difficult task [26]. Web search engines are extremely difficult in that these system must respond to potentially millions of diverse queries, indexing perhaps billions of heterogeneous and distributed information objects, combat spam attacks, and do all of these things at marginal costs [27]. To address multimedia searches, most of the general search engines now use some sort of radio buttons, check boxes, or tabs (thereafter, referred to as radio buttons) for multimedia searching. The use of radio buttons attempts to address the semantic gap [28] that occurs between the textual expression of a multimedia information and the actual multimedia content. Most Web search engines now also present retrieved multimedia objects thumbnail images or video key frames to further help address the semantic gap. These approaches attempt to reduce disorientation and cognitive overload [29]. It is not known what the effect of these changes has had on multimedia Web searching.

In this research, we seek to address this issue by investigating Web searching within specific multimedia collections by examining the searching patterns of AltaVista users. We present results from analysis of both general searching and multimedia content collections.

3. Research Questions

More specifically, the research questions driving this study are:

1) What are the characteristics of AltaVista multimedia searching?

2) What are the effects on Web searching of separate multimedia collections?

The broader goal of our study is to gauge the effect of access to separate multimedia collections via radio buttons on Web searching characteristics and thereby influence the design of multimedia retrieval systems.

4. Research Design

In 2002, AltaVista was the 7th most popular search engine [30] in terms of unique visitors and had a content collection of nearly approximately a billion Web pages [31]. AltaVista supported several query operators including AND, OR, NOT, NEAR, MUST APPEAR, MUST NOT APPEAR, and PHRASE operators [32]. AltaVista offers a full range of searching options, has an extremely large content collection, and millions of unique visitors per month. After being an independent company for several years, Overture Services purchased AltaVista in 2003 [33], and Overture Services was subsequently purchased by Yahoo! [34].

4.1 Data Collection

The AltaVista search engine uses a straightforward ontology designed to assist users in searching for multimedia, namely radio buttons that target searches to specific audio, image or video content collections. AltaVista resolves these searches against separate multimedia content collections. Each collection (i.e., audio, image, and video) has different indexing features and relevance ranking functions based on textual attributes of the particular multimedia file. Many have proposed content based indexing [10, 35]; however, these contextual based indexing methods such as the methods used by AltaVista have proven quite effective within the Web searching environment. An example of the radio button and underlying retrieval technology is presented in Figure 1.

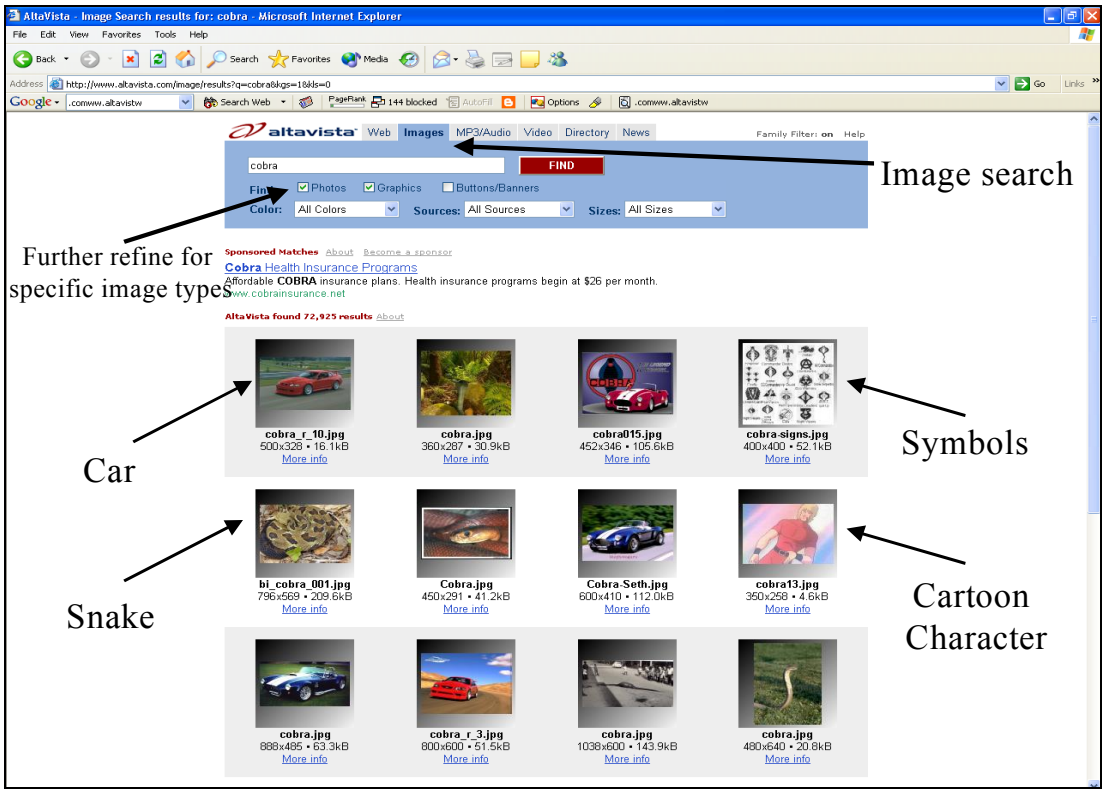

Figure 1 Searching Results from the AltaVista Image Collection for the query *cobra*.

From Figure 1, the query *cobra* has returned a set of results from the AltaVista image collection. If a user clicks on any of the other radio buttons (Web, MP3/Audio, Video), the search engine would execute the query *cobra* against those separate collections. The user can also refine the query further using the radio buttons beneath the text box (i.e., Photos, Graphics, and Buttons/Banners). Like many queries, the query *cobra* can address many different user information needs. Therefore, the search engine offers the user a variety of result types in response, in this case cars, cartoon characters, snakes, and symbols. There are similar searching features for the Web, MP3/Audio, and Video searches.

Like most modern search engines, AltaVista makes uses of many features, primarily textual, in the indexing of multimedia content, including Web page text, link analysis, file names and anchor text. An interesting case is the indexing of video content, since video has both a visual and audio component. What does a user searcher on? AltaVista indexes using textual information, primarily the filename and anchor text, along with text in the page near the anchor. With the anchor text, AltaVista uses a pooling technique where all anchor text from multiple Web pages referring to the same clip multimedia content is pooled together for retrieval purposes. AltaVista uses this pooling technique even if the anchor text is from multiple Web sites.

To address our research questions, we obtained, and quantitatively analyzed, actual queries submitted to the AltaVista main search box in 2002. We also analyzed queries from each of the three multimedia content collections during this same time period. The AltaVista server executed the queries on 8 September 2002, and each transaction logs spans a 24-hour period. We recorded the queries in four transaction logs (*general*, *audio*, *image*, and *video*), representing a portion of the searches executed on the Web search engine on this particular date.

The original general transaction log contains approximately 3,000,000 records. Each record contains three fields:

(1) *Time of Day*: measured in hours, minutes, and seconds from midnight of each day as recorded by the AltaVista server,

(2) *User Identification*: an anonymous user code assigned by the AltaVista server, and

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

Using these three fields, we located the initial query and then recreated the chronological series of actions in a session.

We adopt the terminology outlined by Jansen and Pooch [15] in this research. Specifically,

(1) A *term* is any series of characters separated by white space or other separator.

(2) A *query* is the entire string of terms submitted by a searcher in a given instance, and

(3) A *session* is the entire series of queries submitted by a user during one interaction with the Web search engine.

When a searcher submits a query, then views a document, and returns to the search engine, the AltaVista server logs this second visit with the identical user identification and query, but with a new time (i.e., the time of the second visit). This is beneficial information in determining how many of the retrieved results pages the searcher visited from the search engine, but unfortunately it also introduces duplicate queries in the transaction log.

To address this issue, we collapsed the four data sets by combining all identical queries submitted by the same user to give us the unique queries for analyzing sessions, queries and terms, and pages of

Table 1. Comparison of general, audio, image, and video searching in 2002.

	General	Audio	Image	Video
Sessions	369,350	3,181	26,720	5,789
Queries	1,073,388	7,513	127,614	24,265
Terms				
Unique	297,528 (9.5%)	6,199 (33.4%)	71,873 (14.1%)	8,914 (19.1%)
Total	3,132,106 (100%)	18,544 (100%)	510,807 (100%)	46,708 (100%)
Mean terms per query	2.91 (sd=4.77)	2.47 (sd=1.62)	4.00 (sd=3.21)	1.92 (sd=1.09)
Terms per query				
1 term	218,628 (20%)	2,128 (28%)	27,808 (22%)	9,465 (39%)
2 terms	330,875 (31%)	2,652 (35%)	40,472 (32%)	9,979 (41%)
3+ terms	523,885 (49%)	2,733 (36%)	59,334 (46%)	4,814 (20%)
Mean queries per user	2.91 (sd=4.77)	2.36 (sd=3.85)	4.78 (sd=10.44)	4.19 (sd=6.14)
Users modifying queries	193,468 (52%)	1,496 (47%)	14,838 (56%)	3,350 (58%)
Session size				
1 query	175,882 (48%)	1,685 (53%)	11,882 (44%)	2,439 (42%)
2 queries	75,343 (20%)	657 (21%)	4,759 (18%)	1,016 (18%)
3+ queries	118,125 (32%)	839 (26%)	10,079 (38%)	2,334 (40%)
Results Pages Viewed				
1 page	781,483 (72.8%)	5,551 (73.9%)	80,455 (63.0%)	13,357 (55.0%)
2 pages	139,088 (13.0%)	1,070 (14.2%)	14,498 (11.1%)	3,905 (16.1%)
3+ pages	150,904 (14.1%)	892 (11.9%)	32,661 (25.65)	1,949 (28.9%)
Boolean Queries	61,065 (6%)	210 (3%)	35,955 (28%)	299 (1%)
Terms not repeated in data set	176,196 (6%)	3,720 (20%)	35,955 (7%)	5,292 (11%)
Use of 100 most frequently occurring terms	592,699 (19%)	4,889 (26%)	26,621 (5%)	17,745 (38%)

results viewed [15]. We utilized the complete un-collapsed sessions in order to obtain an accurate measure of the temporal length of sessions and the number of results pages visited.

For analysis of the multimedia data sets, we followed the same procedure and compared the results from the analysis of the general AltaVista transaction log to the results from the analyses of the audio, image and video transaction logs.

5. Results

We present the aggregate results for the analysis in Table 1. Preliminary results appear in a poster presentation [36].

In comparing the four types of searching (general, audio, image, and video), we see that the use of unique terms in audio searching (33%) is substantially higher than the other types of Web searching (10% to 19%), indicating that searching for audio medium utilizes a broader jargon.

The mean terms per query for image searching was notably larger (4 terms per query) than the other categories of searching, which were all less than 3 terms. Video searchers also viewed more pages of results than other searchers, with 45% of video searchers viewing more than one results pages. The session lengths for image searchers were longer than any other type of searching, although video sessions were also relatively lengthy. The session lengths of image searches when combined with the longer queries may indicate that image searching is a more difficult cognitive task than other types of searching. Another indicator of the complexity of image searching is Boolean usage, which was 28%. This is more than four times the next highest category of general Web searching. Audio searching had, by far, the highest percentage of terms not repeated in the data set (20%) and the highest percentage usage of most utilized terms (26%).

Table 2. Comparison of Session Length for Audio, Image, and Video Searching on AltaVista 2002.

	General AltaVista		Audio AltaVista		Image AltaVista		Video AltaVista	
Session Length	Freq.	%	Freq.	%	Freq.	%	Freq.	%
1	175,882	47.62%	1,685	53.0%	11,882	44.5%	2,439	42.1%
2	75,343	20.40%	657	20.7%	4,759	17.8%	1,016	17.6%
3	40,445	10.95%	347	10.9%	2,551	9.5%	530	9.2%
4	23,463	6.35%	204	6.4%	1,656	6.2%	342	5.9%
5	14,719	3.99%	105	3.3%	1,135	4.2%	248	4.3%
6	9,726	2.63%	52	1.6%	830	3.1%	182	3.1%
7	6,664	1.80%	39	1.2%	595	2.2%	158	2.7%
8	4,731	1.28%	13	0.4%	433	1.6%	105	1.8%
9	3,481	0.94%	26	0.8%	375	1.4%	102	1.8%
>=10	14,896	4.03%	53	1.7%	2504	9.5%	667	11.5%
	369,350	100%	3,181	100.00%	26,720	100.00%	5,789	100.00%

5.1 Session Level Analysis

Table 2 presents the results of an analysis of session length from the general and three multimedia transaction logs.

Table 2 shows that in all cases of multimedia searching, the predominant session length was 1 query, with single query audio searching sessions being about 10% higher than similar image or video. Given the relatively easy methods of categorizing audio files, especially music files, this seems like a reasonable outcome. Single session audio searching sessions are also about 5% higher than general Web searching on AltaVista. The session lengths for image and video are comparable to that of AltaVista general Web searching.

5.2 Query Level Analysis

We next examined query lengths. Research shows that increased query length measured in number of terms is directly related to retrieval effectiveness [37]. Table 3 presents the query length percentages for each of the four data sets.

Table 3. Query Lengths in Each Data Set.

Query Length	Percentage of All Queries			
	General	Audio	Image	Video
0	0.03%	0.1%	0.04%	0.03%
1	20.4%	28.3%	21.8%	39.0%
2	30.8%	35.3%	31.7%	41.1%
3	22.8%	17.5%	11.9%	13.2%
4	12.0%	9.5%	3.9%	4.1%
5	5.9%	4.7%	2.7%	1.6%
6	2.5%	2.2%	0.5%	0.5%
7	1.2%	1.0%	0.2%	0.2%
8	0.5%	0.5%	0.1%	0.1%
9	3.5%	0.5%	27.0%	0.1%
>= 10	0.4%	0.4%	0.2%	0.1%

From Table 3, we see that generally the percentages are similar at each query length although with some variation. A notable exception is image queries with length of 9 terms at 27%. Naturally, this skewed the average. We cannot positively account for the clustering at this query length. We examined queries of this length in more detail. There were several individual sessions, so a single user did not submit them. There are two possible hypotheses. One, it is an anomaly of the data collection and over a longer collection period, the percentage would be more evenly spread at the longer query lengths for image searchers. Two, search engine optimizers typically use sets of five terms OR'd together as they

retrieve search engine results to check Web page ranking. Therefore, this spike could represent this market segment.

In Table 4 we present the top 10 repeat queries from each of the four data sets, which represent the most frequently occurring queries from each data set across all session, and after we removed the result page requests.

Table 4. Top 10 Repeat Queries from General, Audio, Image, and Video Data Sets.

General	Freq.	%	Audio	Freq.	%	Image	Freq.	%	Video	Freq.	%
google	837	0.09%	folk	94	1.3%	centaur	99	0.08 %	web cam	98	0.40 %
yahoo	727	0.08%	stargate	78	1.0%	celtic Mcdowell knot	98	0.08 %	black sex	94	0.39 %
ebay	720	0.08%	"everybody 's talkin'"	73	1.0%	secret garden	98	0.08 %	mpeg	90	0.37 %
sex	412	0.05%	cops	64	0.9%	"Sylvia Saint"	96	0.08 %	swinger couple	80	0.33 %
yahoo .com	395	0.04%	apollo 1	41	0.5%	eating disorders caused media	96	0.08 %	indian xxx hardcore sex porno	76	0.31 %
dictionar y	374	0.04%	They're coming to take me away	33	0.4%	malcomx commerati ve stamp	94	0.07 %	Lesbians	73	0.30 %
hotmail	336	0.04%	french	31	0.4%	grief	94	0.07 %	teen sex	72	0.30 %
translator	324	0.04%	"rhapsody in blue" - midi	30	0.4%	pino wrist-gear	92	0.07 %	preteen girls	71	0.29 %
hotmail. com	308	0.03%	explosion	27	0.4%	cake	92	0.07 %	gang bang	70	0.29 %
thumbzill a	306	0.03%	color my world	27	0.4%	"Eminem"	91	0.07 %	open mouth	70	0.29 %
	4,739	0.54%		498	6.63%		950	0.74 %		794	3.27 %

From Table 4, we see that in the general searching category that most of the top repeat queries are for other information Web site (e.g., *google*, *yahoo*, *ebay*). Interestingly, the tenth repeat query is *thumbzilla*, which is a well known adult multimedia site. From the audio data set, it appears that many users search on song lyrics or titles. From the image transaction log, there is a variety of searching terms, primarily subject matter. Similarly for video searching, the preferred search mode seems to be subject matter, rather than specific video frames or sound track.

5.3 Term Level Analysis

We next examined the use of query terms. Table 5 presents the term – frequency for the top twenty-five most frequently occurring terms within each data set, after removal of stop words, numbers, and unidentified character strings.

Table 5. Top 25 Most Frequently Occurring Terms in Each Data Set.

General		Audio		Image		Video	
Total Terms	3,132,106	Total Terms	18,544	Total Terms	510,807	Total Terms	46,708
Term	Frequency	Term	Frequency	Term	Frequency	Term	Frequency
free	0.6%	mp3	1.0%	nude	0.6%	sex	3.7%
sex	0.2%	music	0.8%	sex	0.4%	free	1.2%
pictures	0.2%	you	0.6%	girls	0.3%	teen	1.0%
new	0.2%	sounds	0.5%	pictures	0.2%	nude	1.0%
nude	0.2%	free	0.5%	p***y	0.2%	f**k	0.9%
music	0.2%	sex	0.5%	naked	0.2%	porn	0.9%
school	0.2%	john	0.5%	teen	0.2%	girls	0.9%
how	0.2%	me	0.4%	women	0.2%	p***y	0.9%
lyrics	0.2%	song	0.4%	pics	0.2%	f***ing	0.8%
home	0.2%	love	0.4%	free	0.1%	c*m	0.8%
pics	0.2%	sound	0.4%	black	0.1%	gay	0.7%
download	0.2%	by	0.3%	girl	0.1%	lesbian	0.7%
online	0.1%	my	0.3%	porn	0.1%	video	0.7%
american	0.1%	on	0.3%	big	0.1%	black	0.7%
state	0.1%	songs	0.3%	hot	0.1%	anal	0.6%
county	0.1%	download	0.3%	t*ts	0.1%	hardcore	0.6%
university	0.1%	wav	0.3%	young	0.1%	big	0.6%
car	0.1%	world	0.3%	f***ing	0.1%	hentai	0.5%
texas	0.1%	theme	0.2%	flag	0.1%	t*ts	0.5%
real	0.1%	orgasm	0.2%	sexy	0.1%	young	0.5%
games	0.1%	midi	0.2%	gay	0.1%	girl	0.5%
software	0.1%	star	0.2%	c*m	0.1%	videos	0.5%
art	0.1%	your	0.2%	a**	0.1%	movies	0.4%
map	0.1%	down	0.2%	world	0.1%	asian	0.4%
Florida	0.1%	pink	0.2%	map	0.1%	women	0.4%

In other studies of general Web searching trends, researchers have noted a shift away from entertainment to commercial and increased information searching on a variety of topics [22]. These changes have paralleled the increase availability of commercial content on the Web [38]. In other words, as the commercial content represents a larger percentage of available content, searching of commercial content has also increased.

This shift does not appear to be occurring yet with multimedia searching on the Web, as evidenced by the most frequently occurring terms, most of which are entertainment related. Many of these terms are sexual in nature. Although possibly offensive, it is also important for practitioners and researchers in the field of multimedia information retrieval to understand clearly the motivation and information need of many searchers currently seeking multimedia Web content. It remains to be seen whether or not

Web multimedia searching follows a similar trend as general Web searching with a shift to other and broader information domains.

The other notable area of difference is the frequency of term usage across the four searching categories. The term frequencies between general and image searching are nearly identical. Audio and video frequencies are substantially higher. However, one might expect this given the clustering in the entertainment domain.

5.4 Comparison to General Web Searching

We could locate no previously published study or analysis of Web searching using a multimedia ontology such as AltaVista's (i.e., radio buttons on separate multimedia collections). However, there has been previous research on multimedia searching on the Web [20, 21, 39]. These studies report on multimedia searching using the standard textual interface method for general Web searching (i.e., entering the query in the text box).

Table 6 shows a comparison of the means, standard deviation, and significance test results using data from this research and reported values from [21]. If we view increased query length and increased session length as indicators of searching complexity, we see from Table 6 that the use of a relative simple interface (i.e., radio buttons for specific multimedia types) appears to have reduced the complexity of audio and video searching. It does not seem to have had the same effect on image searching; in fact, terms per query and queries per session have actually increased.

Table 6. Comparison of audio, image, and video searching with and without multimedia interface.

	Audio		Image		Video	
	With	Without *	With	Without *	With	Without *
Mean Terms Per Query	2.47 (sd=1.62)	4.11 (sd=2.67)	4.00 (sd=3.21)	3.46 (sd=2.20)	1.92 (sd=1.09)	3.32 (sd=1.96)
t-test	1.68		Not Significant		Not Significant	
	p <= 0.05					
Mean Queries Per User	2.36 (sd=3.85)	2.44 (sd=2.95)	4.78 (sd=10.44)	3.27 (sd=5.49)	4.19 (sd=6.14)	2.91 (sd=3.85)
t-test	Not Significant		Not Significant		Not Significant	

Note: * Data from [21]

We conducted test of significance on all six comparisons of with and without radio buttons (audio terms with and without, audio queries with and without, image terms with and without, image queries with and without, video terms with and without, and video terms with and without) using a two sample independent t-test (critical value of 1.64) at the 0.05 level of significance. The t-test is a parametric evaluation; however, with large sample sizes the t-test is fairly robust to non-normality [40, 41]. From Table 6, of the six comparisons of with and without radio buttons, only the audio was significantly different.

6. Discussion

This research had the goals of identifying the characteristics of AltaVista Web multimedia searching and investigating the effect of radio buttons on multimedia searching on the AltaVista search engine. From our analysis, multimedia searching appears to require greater interactivity between the user and search engine relative to general Web searching. The increase in the number of query terms, increase session lengths, and the increase in the number of results pages searchers view are indications of this greater interactivity. Overall, the interactions between Web searchers and systems are still relatively simple as evidenced by the low use of query operators. There are indications that the use of query operators has little effect on search engines results [42, 43], which may account for their low usage.

However, the range of information needs appears to be broadening based on the high percentage of unique terms and large number of terms not repeated in the data set. Other Web studies also report a trend toward a broadening of information needs and more interactivity [22, 44]. The increased interactivity could be a useful trend for search engine developers, as it could indicates a move by Web searchers to refine more carefully their information needs, resulting in possible more precise results and more targeted paid results.

There is a sharp decrease in the number of pages viewed, especially between the first and second and the second and third results pages, with very few users viewing more than four or five pages of results. AltaVista users have a low tolerance for reviewing large numbers of results pages. Given that over 70% of Web users utilize search engines to locate other Web sites [45], the implications are rather clear for content providers. Certainly for those publishing multimedia content on the Web or engaged in Web e-commerce in the multimedia area, the need to be ranked within the top 10 or 20 results remains critical in order to direct visitors to one's Web site. This has been a previous well known requirement in the search engine optimizer community, and it appears to be still a valid one.

At the term level of analysis, the most frequently occurring terms represent a small percentage of overall term usage. The most frequently used term (*free*) accounted for approximately 0.6% of all term usage. The use of sexual terms was low, in the general set, and the diversity of terms was quite large. However, sexual searching of multimedia information remains relatively high. Even in the multimedia searching, with more targeted topics, the frequency of top term usage was quite low. Again, this diversity reinforces our initial term analysis findings that these Web users are searching for a variety of information topics.

For the second research question, we examined how multimedia searching on AltaVista compares to general Web searching. Generally, it appears that the use of separate searching interfaces assists users in multimedia searching. Multimedia searching using AltaVista's radio button ontology is less complex in terms of query and session length than searching for multimedia content without such an ontology. However, even with the use of the multimedia radio buttons and specific multimedia content collections, searching for multimedia is more complex than general Web searching. This indicates the need for and the possibility for further system improvements in this area.

Of the four types of searching (general, audio, image and video), image searching appears to be the more multifaceted task and audio the least complex. The mean terms per query for image searching was notably larger (4 terms) than the other categories of searching, which were less than 3 terms. The session lengths for image searchers were longer than any other type of searching, although video

sessions were also relatively lengthy. Boolean usage by image searchers was 28%, over four times the next highest ranked category of general Web searching. Perhaps there is something in the nature of image information that lends itself to Boolean searching. These results certainly indicate the need for more efforts to increase the ease of multimedia searching.

6.1 Strengths

This study contributes to the multimedia searching literature in several important ways. First, the data comes from real users submitting real queries to a real Web information retrieval system. Accordingly, it provides a realistic glimpse into how Web users search for multimedia information, without the self-selection issues or altered behavior that can occur with surveys or lab studies. Second, the sample is quite large, with over 350,000 sessions from general transaction log, and there were 34,000 additional multimedia sessions from the other three transaction logs. Third, we obtained data from a major search engine as measured by both document collection and number of unique visitors to ensure that one could generalize the results. Fourth, it appears to be the first published analysis of Web searching using specific multimedia content collections.

6.2 Limitations

As with any research, there are limitations that one should recognize. The sample data comes from one major Web search engine, introducing the possibility that the queries do not represent the queries submitted by the broader Web searching population. In their review on Web searching, Jansen and Pooch [15] have show that characteristics of Web sessions, queries, and terms are very consistent across search engines.

Other potential limitations are that we do not have information about the demographic characteristics of the users who submitted queries, so we must infer their characteristics from the demographics of Web searchers as a whole. Given the diverse nature of Web retrieval, demographic data can be collected from lab studies [c.f., 46, 47] or surveys [c.f., 48]. In addition, the data was collected on a specific date, introducing the possibility of bias due to these particular dates not being representative. However, a comparison of the collected body of Web transaction log research [22, 44, 49-53] shows a great deal of similarity among Web searchers, indicating particular dates have little effect on session lengths, query lengths, Boolean usage, etc. However, particular dates do have an effect on the usage of popular terms [54-56].

7. Implications

Concerning the effect of radio buttons on multimedia searching, it appears that the use of radio buttons has assisted searchers for certain types of information objects or information needs. The multimedia queries are less complex, and the multimedia sessions are shorter as measured by number of queries. This indicates that radio buttons appear to assist users in their search for multimedia information. Perhaps, a continuation of some type of radio button refinement may assist searchers further, with each radio button corresponding to some type of multimedia attribute or major information need. There are indications that this approach may be worthwhile, with many multimedia objects possessing similar attributes [57]. Additionally, several researchers have utilized Web user behavior to propose or implement system enhancements [c.f., 58, 59].

8. Further Rsearch

There are several avenues for future research. Certainly, we need more analysis in this area on a wider variety of Web search engines, ideally on the most popular search engines such as Microsoft Search, Google, America Online, or Yahoo!. However, access to the transaction log data and the willingness of search engine companies to provide the access to the research community hampers this type of investigation. As Web search engines developers introduce additional searching interfaces changes, researchers should continue the evaluation of these changes to gauge their effect on Web searchers, using either transaction logs analysis or lab studies. Finally, we must continue the analysis of Web searching in order to identify trends, predict future behavior, and identify future user needs.

9 Conclusion

The results of our research provide important insights into the current state of multimedia searching and Web information system usage for searchers, search engines developers, and Web sites designers. For searchers, it appears that the use of radio buttons can improve searching in the multimedia domains. Certainly, system design work needs to continue, especially in the area of multimedia searching, and the development of models for the design of Web searching systems [60]. The short session lengths and short queries are challenging issues for designers of Web information systems. This approach does not seem to be a successful strategy to maximize recall or precision, the standard metrics for information retrieval system performance. However, these are the patterns of user interaction, and these patterns, along with the associated information concerns, can be the basis for future effective system development [61]. Given the large number of users that Web search engines continue to attract, it appears that these searchers are finding relevant information with this searching strategy. This may also indicate the need for a wider variety of metrics to evaluate Web information systems. For Web site designers and information architects, knowledge of how Web users are searching will aid in the design of multimedia content for the Web.

Acknowledgements

We thank AltaVista for providing the Web search engine transaction log without which we could not have conducted this research.

References

1. J. I. Cole, M. Suman, P. Schramm, R. Lunn, and J. S. Aquino, "The UCLA Internet Report Surveying the Digital Future Year Three," Accessed on 1 February 2003 on the World Wide Web at http://www.ccp.ucla.edu/pdf/UCLA-Internet-Report-Year-Three.pdf.
2. M. S. Lew, "Next Generation Web Searches for Visual Content," *IEEE Computer*, vol. 33, pp. 46-53, 2000.
3. J. R. Smith, "Webseek at Columbia University," Accessed on 16 July 2003 on the World Wide Web at http://disney.ctr.columbia.edu/webseek/.
4. S. Lawrence and C. L. Giles, "Accessibility of Information on the Web," *Nature*, vol. 400, pp. 107-109, 1999.
5. D. Lowe and J. Eklund, "Client Needs and the Design Process in Web Projects," *Journal of Web Engineering*, vol. 1, pp. 23-36), 2002.

6. M. S. Lew, N. Sebe, and J. P. Eakins, "Challenges of Image and Video Retrieval," in *Lecture Notes in Computer Science*, vol. 2383/2002, M. S. Lew, N. Sebe, and J. P. Eakins, Eds. Heidelberg: Springer-Verlag, 2002, pp. 1-6.
7. E. M. Rasmussen, "Indexing Images," *Annual Review of Information Science and Technology*, vol. 32, pp. 169-196, 1997.
8. J. Li and J. Z. Wang, "Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1075-1088, 2003.
9. J. Greenberg, "Intellectual Control of Visual Archives: A Comparison between the Art and Architecture Thesaurus and the Library of Congress Thesaurus for Graphic and Materials," *Cataloging and Classifications Quarterly*, vol. 16, pp. 85-101, 1993.
10. J. Z. Wang, "SIMPLIcity: A Region-based Image Retrieval System for Picture Libraries and Biomedical Image Databases," in *Proceedings of ACM Multimedia*, Los Angeles, CA, 2000. pp. 483-484.
11. G. W. Furnas, L. T. K., L. M. Gomez, and S. T. Dumais, "The Vocabulary Problem in Human-System Communication," *Communications of the ACM*, vol. 30, pp. 964-971, 1987.
12. H. Chen, "An Analysis of Image Queries of Art History," *Journal of the American Society for Information Science and Technology*, vol. 52, pp. 260-273, 2001.
13. S. Siegfried, M. Bates, and D. Wilde, "A Profile of End-user Searching Behavior by Humanities Scholars: the Getty Online Searching Project Report No. 2," *Journal of the American Society for Information Science*, vol. 44, pp. 273-291, 1993.
14. C. Choo, B. Betlor, and D. Turnbull, "A Behavioral Model of Information Seeking on the Web: Preliminary Results of a Study of how Managers and IT Specialists use the Web," in *Proceedings of the 61st Annual Meeting of the American Society for Information Science*, Pittsburgh, PA, 1998. pp. 290-302.
15. 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*, vol. 52, pp. 235-246, 2001.
16. C. Jörgensen, *Image Retrieval: Theory and Practice*. New York:: Scarecrow Press, 2003.
17. A. F. Smeaton, "Indexing, Browsing, ans Searching of Digital Video," in *Annual Review of Information Sciences and Technology*, vol. 38, B. Cronin, Ed. Medord, Nj, USA: Information Today, 2004, pp. 371-407.
18. J. Turner, "Words and pictures in information systems for moving images," in *Proceedings of the 1998 Association of Moving Image ArchivistsConference*, Miami, Florida, USA, 1998. pp. 309-315.
19. M. J. Swain, "Searching for Multimedia on the World Wide Web," in *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, Florence, Italy, 1999. pp. 9032-9037.
20. A. Goodrum and A. Spink, "Image Searching on the Excite Search Engine," *Information Processing & Management*, vol. 37, pp. 295-311, 2001.
21. B. J. Jansen, A. Goodrum, and A. Spink, "Searching for Multimedia: Video, Audio, and Image Web Queries," *World Wide Web Journal*, vol. 3, pp. 249-254, 2000.
22. A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic, "From E-sex to E-commerce: Web Search Changes," *IEEE Computer*, vol. 35, pp. 107-111, 2002.
23. C. Silverstein, M. Henzinger, H. Marais, and M. Moricz, "Analysis of a Very Large Web Search Engine Query Log," *SIGIR Forum*, vol. 33, pp. 6-12, 1999.
24. A. Spink and B. J. Jansen, *Web Search: Public Searching of the Web*. New York: Kluwer, 2004.
25. C. Ozmutlu, A. Spink, and Z. Ozmutlu, "Multimedia Web Searching Trends: 1997 - 2001," *Information Processing & Management*, vol. 39, pp. 611 - 621, 2003.
26. A. Ginige and S. Murugesan, "Web Engineering: An Introduction," *IEEE Multimedia*, vol. January - March, pp. 14-18, 2001.
27. S. T. Dumais, "Web Experiments and Test Collections," Accessed on 20 April 2003 on the World Wide Web at http://www2002.org/presentations/dumais.pdf.

28. V. V. Gudivada and V. V. Raghavan, "Content-based Image Retrieval Systems," *IEEE Computer*, vol. 28, pp. 18-22, 1995.
29. M. Villanova-Oliver, J. Gensel, and H. Martin, "A Progressive Access Approach for Web-based Information Systems," *Journal of Web Engineering*, vol. 1, pp. 27-57, 2004.
30. D. Sullivan, "Nielsen / NetRatings Search Engine Ratings," Accessed on 6 January 2002 on the World Wide Web at http://www.searchenginewatch.com/reports/netratings.html.
31. D. Sullivan, "Search Engine Sizes," Accessed on 30 August 2000 on the World Wide Web at http://searchenginewatch.com/reports/sizes.html.
32. AltaVista, "Special search terms," Accessed on 16 May 2003 on the World Wide Web at http://www.altavista.com/help/adv_search/syntax.
33. B. Morrissey, "Overture to Buy AltaVista," Accessed on 16 May 2003 on the World Wide Web at http://www.internetnews.com/IAR/article.php/1587171.
34. B. Berkowitz, "Yahoo! to Buy Overture in $1.63 Bln Deal," Presented at Los Angeles, CA, USA.2003.
35. C. Djeraba, "When Image Indexing Meets Knowledge Discovery," in *Proceedings of Sixth ACM SIGKDD International Conference on Know edge Discovery and Data Mining*, Boston, MA, USA, 2000. pp. 73-81.
36. B. J. Jansen, A. Spink, and J. Pederson, "An Analysis of Multimedia Searching on Alta Vista," in *Proceedings of the 5th ACM SIG Multimedia International Workshop on Multimedia Information Retrieval.*, Berkeley, California, 2003. pp. 186 - 192.
37. N. Belkin, C. Cool, W. B. Croft, and J. Callan, "The Effect of Multiple Query Representations On Information Retrieval Systems," in *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1993. pp. 339 - 346.
38. S. Lawrence and C. L. Giles, "Searching the World Wide Web," *Science*, vol. 280, pp. 98-100, 1998.
39. C. Ozmultu, A. Spink, and Z. Ozmultu, "Multimedia Web Searching Trends: 1997 - 2001," *Information Processing & Management*, vol. 39, pp. 611 - 621, 2003.
40. D. Hull, "Using statistical testing in the evaluation of retrieval experiments," in *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, Pittsburgh, Pennsylvania, 1993. pp. 329 - 338.
41. R. R. Newton and K. E. Rudestam, *Your Statistical Consultant: Answers to your Data Analysis Questions*. Thousand Oaks, CA, USA: Sage Publications, 1999.
42. B. J. Jansen and C. M. Eastman, "The Effects of Search Engines and Query Operators on Top Ranked Results," in *Proceedings of the IEEE 4th International Conference on Information Technology*, Las Vegas, NV, 2003. pp. 135-139.
43. C. M. Eastman, "30,000 Hits May be Better than 300: Precision Anomalies in Internet Searches," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 879-882, 2002.
44. B. J. Jansen and A. Spink, "An Analysis of Web Searching By European Alltheweb.com Users," *Information Processing and Management*, vol. 41, pp. 361-381, 2005.
45. Alexa Insider, "Alexa insider's page," Accessed on 30 March 2000 on the World Wide Web at http://insider.alexa.com/insider?cli=10.
46. E. G. Toms, C. Dufour, and S. Hesemeier, "Evaluation: Measuring the user's experience with digital libraries," in *Proceedings of the 2004 Joint ACM/IEEE conference on Digital libraries*, Tuscon, AZ, USA, 2004. pp. 51 - 52.
47. E. Hargittai, "Beyond logs and surveys: In-depth measures of people's web use skills," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 1239-1244, 2002.
48. A. Spink, J. Bateman, and B. J. Jansen, "Searching the Web: A Survey of Excite Users," *Journal of Internet Research: Electronic Networking Applications and Policy*, vol. 9, pp. 117-128, 1999.

49. G. Abdulla, B. Liu, and E. Fox, "Searching the World-Wide Web: Implications from Studying Different User Behavior," in *Proceedings of the World Conference of the World Wide Web, Internet, and Intranet*, Orlando, FL, 1998. pp. 1 - 8.
50. F. Cacheda and Á. Viña, "Experiences retrieving information in the World Wide Web," in *Proceedings of the 6th IEEE Symposium on Computers and Communications*, Hammamet, Tunisia, 2001. pp. 72-79.
51. C. Hölscher and G. Strube, "Web Search Behavior of Internet Experts and Newbies," *International Journal of Computer and Telecommunications Networking*, vol. 33, pp. 337-346, 2000.
52. A. Montgomery and C. Faloutsos, "Identifying web browsing trends and patterns," *IEEE Computer*, vol. 34, pp. 94-95, 2001.
53. E. Selberg and O. Etzioni, "The MetaCrawler architecture for resource aggregation on the Web," *IEEE Expert*, vol. 12, pp. 11 - 14, 1997.
54. D. Wolfram, "Term Co-occurrence in Internet Search Engine Queries: An Analysis of the Excite Data Set," *Canadian Journal of Information and Library Science*, vol. 24, pp. 12-33, 1999.
55. D. Wolfram, A. Spink, B. J. Jansen, and T. Saracevic, "Vox Populi: The Public Searching of the Web," *Journal of the American Society of Information Science and Technology*, vol. 52, pp. 1073-1074, 2001.
56. Searchwords.com, "Top Search Words," Accessed on 30 May 2003 on the World Wide Web at http://www.searchwords.com.
57. B. J. Jansen, "A Preliminary Mapping of Web Queries Using Existing Image Query Schemes," in *Proceedings of the E-Learning 2002 Conference (Web Track)*, Montreal, Canada, 2002. pp. 1-5.
58. H.-M. Chen and M. D. Cooper, "Stochastic modeling of usage patterns in a web-based information system," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 536-548, 2002.
59. B. J. Jansen and U. Pooch, "Assisting the Searcher: Utilizing Software Agents for Web Search Systems," *Internet Research - Electronic Networking Applications and Policy*, vol. 14, pp. 19-33, 2004.
60. M. D. Jacyntho, D. Schwabe, and G. Rossi, "A Software Architecture for Structuring Complex Web Applications," *Journal of Web Engineering.*, vol. 1, pp. 37-60.
61. D. Schwabe, R. Guimarães, and G. Rossi, "Cohesive Design of Personalized Web Applications," *IEEE Internet Computing*, vol. 6, pp. 34 - 43, 2002.