Selection and Context Scoping for Digital Video Collections: An Investigation of YouTube and Blogs

Robert G. Capra, Christopher A. Lee, Gary Marchionini,
Terrell Russell, Chirag Shah, Fred Stutzman
School of Information and Library Science
University of North Carolina at Chapel Hill
100 Manning Hall
rcapra3@unc.edu, callee@email.unc.edu, march@ils.unc.edu,
unc@terrellrussell.com,chirag@unc.edu, fred@metalab.unc.edu

ABSTRACT

Digital curators are faced with decisions about what part of the ever-growing, ever-evolving space of digital information to collect and preserve. The recent explosion of web video on sites such as YouTube presents curators with an even greater challenge – how to sort through and filter a large amount of information to find, assess and ultimately preserve important, relevant, and interesting video. In this paper, we describe research conducted to help inform digital curation of on-line video. Since May 2007, we have been monitoring the results of 57 queries on YouTube related to the 2008 U.S. presidential election. We report results comparing these data to blogs that point to candidate videos on YouTube and discuss the effects of query-based harvesting as a collection development strategy.

Categories and Subject Descriptors

H.3.7 Information Storage and Retrieval: Digital Libraries, H5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous.

General Terms

Measurement, Human Factors

Keywords

Digital curation, collection management, video, blogs

1. INTRODUCTION

Large amounts of information are born digital and may only be available electronically. Increasing amounts of user-generated multimedia content are available on the web through environments such as Facebook, MySpace, and YouTube. These sites provide popular venues for individuals and organizations to express their views, opinions, and shared lives with the larger community of web users, which in turn may affect thought and behavior in realms such as politics. Environments such as YouTube and the blogosphere are not digital libraries (DLs), however, they present both new challenges and unprecedented opportunities for digital librarians, curators and archivists who want to select resources for their collections. Digital curators are now faced with challenges of how to find and filter large amounts of video data; how to assess and ultimately preserve important, relevant, and interesting videos; and how to create or capture contextual information to increase the long-term understanding and meaningful use of videos and other digital assets in their care. The ease with which videos may be recorded and posted means that videos may respond to events very quickly and this offers new kinds of cultural and historical context for digital collections.

In 2007, two independent data collection efforts related to the 2008 U.S. presidential election were started at the UNC School of Information and Library Science. The first effort was part of the VidArch project [16] to help inform curation of on-line video. Since May 2007, we have been monitoring the results of 57 queries on YouTube related to the 2008 U.S. presidential election. The second data collection effort was part of a project to gain insights into the role and activities of the blogosphere related the election. Since June 2007, this effort has been collecting information about blog pages returned by queries to Google Blog search and Technorati. These two efforts were started independently with different, but overlapping query sets. One of the goals of the VidArch project is to explore sources and methods for collecting Internet videos and their surrounding contextual information. The blogosphere dataset provided a useful point of comparison to the videos and context being found through YouTube alone. In this paper, we compare the procedures and results of both datasets with an eye toward building curator tools and informing collection management.

2. YOUTUBE AND THE BLOGOSPHERE AS SITES OF ONLINE DISCOURSE

YouTube provides a fast method for disseminating video to a large audience. This can provide new opportunities for relatively free and open discourse, while also challenging the control of traditional authorities over the predominant messages associated with particular issues. In the 2006 U.S. elections, for example, the use of YouTube and MySpace limited "the level of control that campaigns have over the candidate's image and message since anybody, both supporters and opponents, can post a video and/or create a page on behalf of the candidates because of the user-driven content of social networking sites." [7, p.8] Based on an analysis of 153 YouTube videos, Keelan et al. found that videos taking a negative position on vaccination and immunization "were more likely to receive a rating… had a higher mean star rating and more views…" and that "45% conveyed messages that contradicted the reference standard." [9, Results section].

JCDL'08, June 16-20, 2008, Pittsburgh, Pennsylvania, USA.
Copyright 2008 ACM 978-1-59593-998-2/08/06...$5.00.

A striking example of how web materials can be used to disseminate content and opinions, and potentially influence the course of history are the YouTube videos related to the U.S. 2008 presidential election. YouTube is playing an increasingly important role in political discourse and may have a significant impact on the political process and voting behaviors [12]. All the major candidates have YouTube channels associated with their campaigns. In conjunction with CNN, YouTube sponsored Republican and Democratic debates that featured video-recorded questions uploaded by YouTube users. Candidates have also used YouTube to "converse" with voters – for example, several candidates posted questions to voters on YouTube asking for them to post their feedback and replies. Perhaps even more importantly, events that would have previously had only a very local impact – such as small public rallies involving only a handful of people or individuals expressing their opinions about particular candidates to their peers – can now attain widespread visibility and impact, because they are posted as videos to YouTube. Thus, curators of collections devoted to politics, political leaders, or history more generally will want to include in their collections some of these videos and the commentary and usage patterns related to them.

An even larger set of web sites, including blogs, provide links to and commentary about this multimedia content. As with YouTube, the political blogosphere is a popular and influential space for discourse around political issues and elections. Like YouTube, the blogosphere is also a relatively unconstrained medium, which provides space for extended discussion, speculation and agenda-pushing that might not happen in traditional media venues. This discursive freedom, combined with the implicit conversational nature of the blogosphere, provides a "third space" for political discussion, and like the YouTube space must be considered by digital curators.

Scholars have examined the political blogosphere extensively, exploring its network structure, linked affiliations, discursive properties, and media effects. Adamic and Glance [1], exploring the political blogosphere surrounding the 2004 U.S. presidential elections, found the blogosphere to be a balkanized place. Studying a sample of 40 A-list blogs (20 left-leaning, 20 right-leaning), the authors found significant in-linking between political affiliation, with less cross-linking across political belief.

The blogosphere at-large has demonstrated scale-free [2] properties, but the emergence of scale-free properties in the political subset of the blogosphere proves interesting. Hindman et. al. [8] explored the network structure of a number of political blogging communities, finding that link patterns follow a power law. The preferential linking patterns drive an A-list or vanguard of political blogs, whose network centrality may provide influence over the information they disseminate.

Drezner and Farrell, in *The Power and Politics of Blogs* [6], explore the distribution of links and resultant power structures in the blogosphere. The authors find a distinctly skewed distribution of links, which essentially funnels traffic to a subset of powerful political blogs. Depending on where one blog exists in the blogosphere's network structure, the information disseminated may have outlying importance effects.

Political blogs have proven to have significant effects in all parts of the discussion surrounding politics, funneling and vetting emergent stories for the media, as well as creating a place for extended research and discourse. Whether it be "Rathergate"[1] from the 2004 U.S. presidential election, the offensive utterance of George Allen[2] (a U.S. senatorial candidate in 2006), or the emerging ecology of the supporter-driven 2008 presidential elections, blogs continue to play an important role as a vehicle for information diffusion and discourse. Thus, digital archivists must consider both YouTube and the blogosphere as sources for both primary and contextualizing content.

3. PREVIOUS WORK ON COLLECTING WEB RESOURCES

Collecting resources from the web is receiving attention in a wide range of for-profit, government, and cultural institutions. Adrian Brown [5] provides a summary of web archiving initiatives, which include the Internet Archive, National Library of Sweden's Kulturarw3, Nordic Web Archive, National Library of Australia's PANDORA (Preserving and Accessing Networked Documentary Resources of Australia), NEDLIB (Networked European Deposit Library), U.S. Library's of Congress's MINERVA (Mapping the Internet Electronic Resources Virtual Archive), DACHS (Digital Archive of Chinese Studies), National Diet Library of Japan's Web Archiving Project (WARP), UK Central Government Web Archive, UK Web Archiving Consortium, European Digital Archive, International Internet Preservation Consortium), and web crawls by the Bibliothèque nationale de France (BnF). Web capture for building digital collections has usually been based on the identification of a set of seed uniform resources locators (URLs) and then recursively following links within a specified set of constraints (e.g. number of hops, specific domains). Several projects have also demonstrated methods for further scoping a topic-based crawl, based on automated analysis of the content of pages [3]. There have also been efforts to automatically populate web entry forms, in order to collect pages that cannot be reached through link-following [10][11][13]. However, there has been relatively little investigation of using term-based queries as the primary mechanism for crawling and selecting resources from existing online use environments, such as YouTube or the blogosphere.

4. VIDARCH PROJECT, GOALS AND RESEARCH QUESTIONS

The VidArch Project focuses on developing strategies and tools for curators to find and incorporate videos and associated contextual information into their collections (http://ils.unc.edu/vidarch). We have harvested videos and a range of metadata from YouTube on a focused set of topics including sustainable energy, elections, natural disasters, diabetes information, and health epidemics. In the work reported here, we focus on videos related to the 2008 U.S. Presidential election.

When YouTube users upload videos, they provide initial metadata about the videos (e.g. title, author, running time). As the video is watched, other members of the YouTube community may rate the

[1] Bloggers challenged the authenticity of documents reported on by news anchor Dan Rather about President George W. Bush's military record.

[2] Allen made a comment at a campaign rally that some people interpreted as a racial slur. The incident and reaction was extensively discussed in the blogosphere and viewed on video sites such as YouTube [12].

video, mark it as a favorite, link to it, leave text comments about it, or post a video response to it. Each time it is launched, the number of views associated with a video is increased. These community-generated metadata elements are one of the primary kinds of contextual information of interest in our work.

Our goal is to help archivists, curators, and digital librarians make effective and efficient decisions about what videos and associated contextual information to add to their collections. Collection development and appraisal decisions have traditionally depended on professional cognitive cycles to apply formal principles, selection aids, peer recommendations, and often direct examination of resources within the universe of potential acquisitions. Collecting decisions have been quantitatively and qualitatively complicated by the onslaught of millions of videos and links to those videos being posted to numerous Internet outposts from countries across the globe on every topic imaginable. Online videos collections support and reflect many new forms of 'behavior' in and around the videos. The information life cycle also has become much more dynamic as digital resources change over time. These changes can be much more diverse than the deterioration of physical objects.

The most pervasive change for online video today is the set of data elements that are added over time, based on either direct user input or as a result of their interaction with a video. Links to a video (in-links) 'say' something about that video just as citations to a paper do. The number of views, comments, and the general 'buzz' surrounding a video also 'say' something about it and become part of its history—a history that may be important to future generations who wish to understand and make meaningful use of the video or better understand the person or event portrayed in the video. Because the numbers and kinds of evidence that accrue to any digital object are so large, curators must leverage technology to help sift the universe of video to find the most pertinent ones to add to their collections; either as primary objects themselves or as sources of contextual information for other objects in their collections. In this paper, we describe our efforts to harvest video files from YouTube, and associated information from both YouTube and the blogosphere. The results will guide our ongoing development of curatorial tools and our investigation of strategies for identifying, selecting, and incorporating contextual information into collections and curatorial decisions.

For a video within YouTube, changes to data associated with the video can strongly influence how the video is perceived, used and understood[3]. A curator who ingests the video into another repository for long-term preservation will, therefore, often find it important to also ingest both metadata created at the time the video was posted to YouTube, as well as metadata that is added or changed through the video's life within YouTube. Metadata changes, annotations, uses, citations and allusions to a video can all serve as valuable contextual information. Other contextual information can also reside in secondary or tertiary environments (e.g., a newspaper article that mentions a YouTube video).

In this paper, we distinguish the following terms:

Primary source – the direct target of preservation (e.g., a video and its metadata)

Secondary source - external resource that discusses, points to, or provides important supplemental information about a primary source (e.g. a blog with a link to a video)

Primary use environment – the environment in which a primary source is directly posted and intended to be used (e.g. YouTube)

Secondary use environment – any use environment outside the primary use environment (e.g. a blog linking to a YouTube video)

Primary distribution – act of posting a digital object directly within the primary use environment

Secondary distribution – act of posting a pointer to a primary source from outside the primary use environment

Primary relevance measure – measure of relevance as determined by the mechanisms built directly into the primary use environment

Secondary relevance measure – any measures of relevance that are not primary relevance measures

In the context of this paper, the **primary sources** are YouTube videos, the **secondary sources** are blog posts, **primary distribution** is the act of posting to YouTube, our main example of **secondary distribution** is the act of posting to a blog, **primary relevance measures** are those generated by YouTube (such as the search rankings on the YouTube site), and **secondary relevance measures** are generated by mechanisms external to YouTube.

Research Question (RQ) 1: Can information from secondary source environments (including secondary relevance measures) help curators of digital collections to identify videos that are relevant to their collections, which they would not discover solely by consulting information (including primary relevance measures) from the primary use environment?

RQ1 focuses on the potential value of considering the resources identified as relevant within the **union** of primary and secondary relevance measures.

Research Question (RQ) 2: Given a set, S, of digital objects that have been deemed potentially relevant, based on primary relevance measures, can secondary relevance measures act as a useful filter for identifying a subset of S that is particularly important to ingest into a digital collection?

RQ2 focuses on the potential value of paying particularly high attention to resources identified as relevant within the **intersection** of primary and secondary relevance measures.

As described below, we have operationalized both RQ1 and RQ2 by collecting and analyzing data from both YouTube and blog posts that link to YouTube videos.

In this paper we:

- Describe our process and motivations for collecting YouTube and Blog data about the 2008 US presidential election.
- Examine the characteristics of the two data collection methods (YouTube, Blogs).
- Examine the differences and overlaps in the videos that the two methods collected.
- Describe and present a relevance analysis of each collection (YouTube, Blogs, and the overlap) to help understand the outputs of each method.
- Discuss potential implications of our overlap and relevance analysis of the two sources for selection of resources to ingest into digital repositories, based on various collecting objectives.

[3] YouTube does not currently allow changes to video files, but it does support metadata editing. YouTube does remove videos at the request of the provider or if they violate terms of use.

5. DATA COLLECTION

5.1. YouTube Data Set

We chose to analyze video postings and user feedback on YouTube, because it is one of the largest online video sharing sources. As of October 2007, YouTube was reported to be serving 100 million views and 65,000 new video uploads per day [4].

We used the YouTube application program interfaces (APIs) to harvest and monitor videos related to the 2008 U.S. presidential election posted to YouTube, along with associated comments and metadata [14]. The dataset is the result of 57 queries that we have been issuing to YouTube every day (except for days of maintenance), since May 2007. The queries include 50 names of possible candidates obtained from the Wikipedia page "United States presidential election, 2008" [17] on April 27, 2007. We have also included 6 queries that refer to the election in general (e.g. "election 2008"). A complete list of the queries is available on the VidArch project website at http://www.ils.unc.edu/vidarch.

We use the term "crawl" to indicate one instance of executing the following two sets of activities: 1) submitting all 57 queries to YouTube and then collecting data from the top 100 results of each query based on YouTube's search option of sorting by "relevance"; and 2) collecting updated dynamic metadata for each video that has been "discovered" through any instance of step 1. Since our first crawl in May 2007, the queries in each subsequent crawl have generated a combination of (1) references to videos that already appeared in the results of previous queries, and (2) references to videos that did not appear in the results of previous queries. For a given query (e.g. "John Edwards") that is executed as part of a crawl, a video "discovery" is any new reference to a video that has not appeared in the top 100 result set for that query in any previous crawls.

When a video is first discovered within a crawl, we collect static metadata (video ID, title, contributor, date added, description and tags) and dynamic metadata (number of views, ratings, number of honors, and number of times favorited) associated with the video. Then the video is added to a list of "discovered" videos associated with each query. In step 2 of subsequent crawls, the dynamic metadata for each video is collected again. Each time data are captured for a video, a time-stamp is recorded.

Table 1. Statistics Describing the YouTube Data Set

Factor	May 2007	January 2008
Total videos	4,908	15,327
Total views	33,102,746	235,542,200
Total comments	189,507	1,649,547
Total ratings	209,672	1,681,607

At the end of January 2008, we had completed over 200 crawls on YouTube, which generated 16,791 aggregate video discoveries of which 15,327 were unique videos, across all of the 57 queries. Table 1 shows statistics about the videos tracked in our crawls.

YouTube assigns a unique 11-character identifier to each video. We rely on this identifier to determine if a given video has appeared across multiple crawls or across multiple queries. For example, if video ABCDEFGHIJK appeared in the top 100 results for YouTube queries on both "Hillary Clinton" and "Barack Obama," this would count as two total video discoveries but only one unique video.

Note that two videos with different identifiers could still contain the same or similar video content. For example, if two different users both upload exactly the same video of a campaign commercial, there would be two separate video identifiers on YouTube since it was uploaded two different times. YouTube users sometimes take advantage of this ambiguity by re-submitting popular videos, in order to bring attention to the account or channel being used to re-distribute the video.

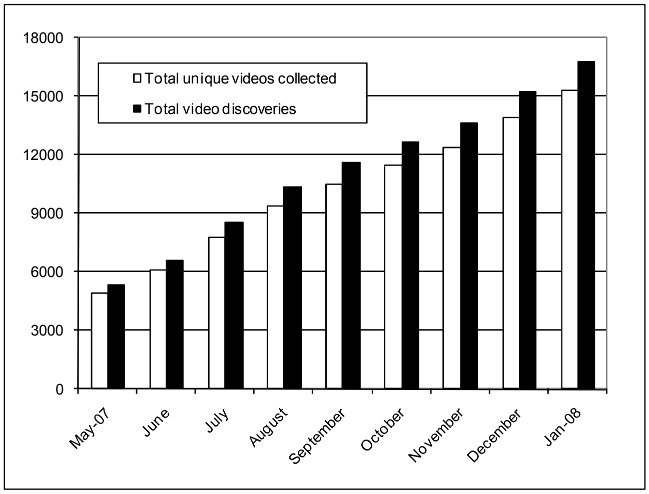

Figure 1. YouTube Data Set Growth, (in terms of number of videos) May 2007 – Jan 2008

Figure 1 shows total number of video discoveries and the unique videos collected from May 2007 until January 2008 and indicates that videos are consistently being added to YouTube. It also shows that there is relatively low overlap among the queries in our crawls (i.e. the difference between the total and unique videos is small). This implies that for this method of collection, the choice of queries is extremely important in determining the scope of the resulting collection.

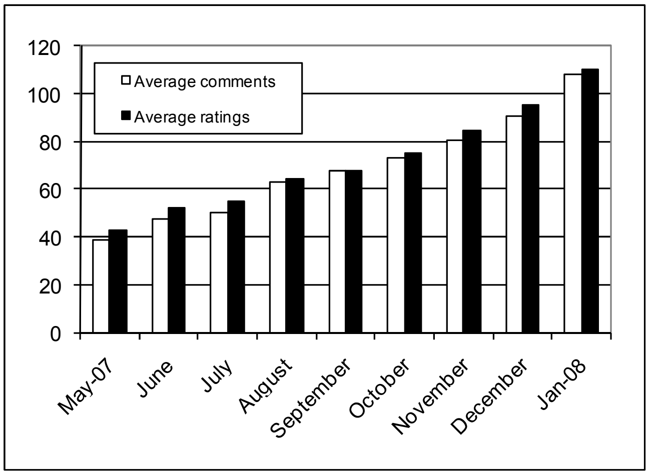

Figure 2. Average Number of Comments and Ratings for YouTube Data Set, May 2007 – Jan 2008

Figure 2 shows the average number of comments (total number of comments / total number of videos) and the average number of ratings over the eight months that we have been crawling. Both averages have increased steadily during our collection, and both have grown much faster than the number of videos. This suggests that videos ranked as highly relevant by YouTube continue to attract new comments and ratings over time. This is a result one might expect to find if many YouTube users were discovering videos by submitting explicit queries to YouTube and then examining some of the videos that appear near the top of the results list.

Our YouTube data set represents videos that were at some point in the collection period ranked in the top 100 results based on relevance to our 57 queries by YouTube's relevance ranking algorithm. Our working assumption is that combining these two criteria (relevance and top 100) can generate a useful first approximation of videos that might be appropriate to ingest into a collection in order to document particular topics, actors or phenomena – in this case, a given candidate or the election more generally.

Primary relevance measures based on query terms are not the only potential crawling or selection criteria for YouTube videos. Any access point (or combination of access points) provided by YouTube could serve as the basis for crawls and selection decisions. We see great potential for repositories to develop complimentary collections of video, based on different collecting missions and crawl criteria. For example, the major candidates in the election all have YouTube channels to which they post official campaign videos. A curator could set up crawls to monitor (or simply ingest without further examination) all videos and associated metadata from a given set of channels. Such a channel-based approach could be very useful if the primary collecting mission were specifically to document only the products of the entity responsible for the channel. However, it would often fail to reflect many highly visible and influential videos that were distributed through other accounts or channels in YouTube (e.g. criticisms, parodies or personal commentaries about presidential candidates; accounts of rapidly emergent events, such as Hurricane Katrina).

5.2. Blogosphere Data Set

Beginning June 6, 2007, one of the authors began a systematic collection of links from blog postings related to the 2008 U.S. presidential election. Queries related to 15 of the presidential candidates[4] were submitted through both Google Blogsearch and Technorati. A complete list of the queries is available on the VidArch project website at http://www.ils.unc.edu/vidarch. For the purpose of this study, a subset of blog postings were captured that either (1) included the name of a presidential candidate in their content or (2) provided one or more links to a candidate's web site.

At the time we conducted the analyses for this paper (late January/early February 2008), the leading candidates based on delegate count were John McCain, Mike Romney, and Mike Huckabee for the Republicans, and Barack Obama, Hillary Clinton, and John Edwards for the Democrats. We often refer to these six candidates as the "leading candidates" in the paper. As the paper was being prepared, John Edwards and Rudolph Giuliani dropped out of the race.

The queries were run using the RSS functions of Google Blogsearch and Technorati three times an hour, every hour of the day, for a total of 72 queries-per-term per day. Each service limits its RSS search to a maximum of ten results per query. Therefore, each query term is limited to a maximum of 720 results/day. Because of this limitation, some potentially relevant blog posts may not be included in the data set. The data set is not intended to be a comprehensive record of all blog posts about all candidates, but rather to provide an illustrative sample of "conversation" around a candidate in the blogosphere. Once the query set was retrieved from the search engine, a web crawler was dispatched to matching pages. This crawler created "profiles" of those pages, collecting the outbound links.

Table 2. Candidate Link Corpus Size

Candidate	# Blog Pages Collected	Out-links	Out-links per Page (Mean)
Clinton	142537	1501864	10.54
Edwards	108629	1192285	10.98
Huckabee	27257	361755	13.27
McCain	71655	745471	10.40
Obama	127526	1361715	10.68
Romney	69081	685446	9.92

The data collected from the blogosphere consists of two main parts. The first part is the set of collected blog pages about a candidate, as determined by the blog search engines. The second part is the set of outbound links generated from the pages gathered in the first part. Table 2 shows the number of pages found and the average number of outbound links on those pages for leading candidates at the time of analysis. Within this data set, an "out-link" is any link from the blog page to a resource outside that page; thus including links to other postings within a blog, navigational links, and links to ads or related postings. The averages provide an estimate of how much outbound content one can expect from a page matching a candidate, thereby establishing a metric for predicting corpus growth.

The average number of outbound links per page for each candidate is roughly the same, suggesting that there is not much difference in the outbound linking behaviors of bloggers writing about one candidate versus another.

6. ANALYSIS AND COMPARISON OF DATA SETS

6.1. Blogosphere Links to YouTube

Table 3 presents statistics about the number of links to YouTube videos in blog pages found for the leading candidates in relation to the total number of out-links. This percentage provides a measure of the relative amount of **video** conversation about a candidate. Note that the numbers in Table 3 are for the total number of video links found rather than for the unique videos.

The analysis shows that the percentages for the top candidates are all in the range of 0.54% to 0.72%. Overall, links to YouTube videos make up a rather small percentage of all the outbound links from this set of blog pages. The low percentages of videos highlights the limitations of our simplistic approach of considering all the out-links on a matching blog page and suggests the need for more robust parsing and contextualization strategies to help identify links and context of interest. However, in terms of raw numbers, these blog pages have thousands of links to YouTube

[4] For the top three democratic candidates, the queries included one or two additional terms such as the URL of the candidate's website and their spouse's name. These query expansions represent differences in how the Blogosphere and YouTube data sets were collected and are a result of the data collection efforts starting independently. In addition, the blog data set did not include any queries for Ron Paul, who is known for his extensive use of the Internet in his campaign.

videos, illustrating important points where the blogosphere and Internet video communities intersect.

Table 3. Outbound Links to YouTube From Blog Pages

Candidate	Total Out-Links (OL)	Out-Links to YouTube Videos (YTOL)	Percentage of Out-Links Pointing to YouTube Videos (YTOL/OL)
Clinton	1501864	9805	0.65%
Edwards	1192285	7794	0.65%
Huckabee	361755	2066	0.57%
McCain	745471	4000	0.54%
Obama	1361715	9777	0.72%
Romney	685446	4361	0.64%

6.2. Overlap and "Blogginess"

Our YouTube data set (YT) relies on the relevance ranking algorithms implemented by YouTube to identify videos of potential interest to archivists. Similarly, our Blog data set (B) relies on the rankings of two blog search engines (Google Blogsearch and Technorati). To better understand the characteristics of each collection method, we conducted an analysis of the overlap (O) between the two sets (YT ∩ B, the intersection of YT and B). Note that the Blog data set did not contain all the candidates that were included in the YT set. Thus, in this analysis, only the candidates that were in both the YT and B sets are included.

Table 4. YT and Blog Overlap

	YT	B	YT∩B	O/YT	O/B
Gravel	404	1677	168	42%	10%
Clinton	484	6794	178	37%	3%
Kucinich	372	2138	135	36%	6%
Edwards	1021	5221	361	35%	7%
Giuliani	525	3348	162	31%	5%
Romney	509	3060	150	29%	5%
Obama	1005	6718	281	28%	4%
Huckabee	521	1574	139	27%	9%
Biden	278	1444	73	26%	5%
Richardson	378	1493	82	22%	5%
McCain	651	3049	127	19%	4%
Tancredo	254	1130	39	15%	3%
Brownback	266	1228	39	15%	3%
Dodd	375	516	14	4%	3%
Vilsack	97	143	3	3%	2%
Thompson	241	2110	2	1%	0%

Analysis of the overlap between the sets is of interest because they were collected using very different methodologies with different goals. However, both sets relied on the rankings of search engine technology – the YT set on the YouTube ranking algorithm, and Blog set on the rankings of Google and Technorati. Related research has shown that there is low overlap among the results returned for queries to different web search engines (e.g. Google, Yahoo!, MSN, and Ask), even in the first page of results [15]. In this analysis, we were interested to see how much overlap there was between our two data sets.

For each candidate in both the YT and Blog data sets, Table 4 shows the: 1) the number of unique videos in our YouTube data set, 2) the number of unique videos in our Blog data set, and 3) the number of videos that appeared in both lists (the overlap). In addition, measures of the overlap divided by the number of YT videos and the overlap divided by the number of Blog videos are shown. The table is sorted by decreasing value of O/YT.

Figure 3 shows a plot of the "blogginess" of each candidate's set of videos as computed by dividing the number of overlapping videos by the total number of YouTube videos (the O/YT percentage in Table 4). This represents what percentage of the YouTube videos for the query were also found in blogs. A higher percentage indicates that videos that YouTube ranks highly for that candidate's query are also being referenced on blogs.

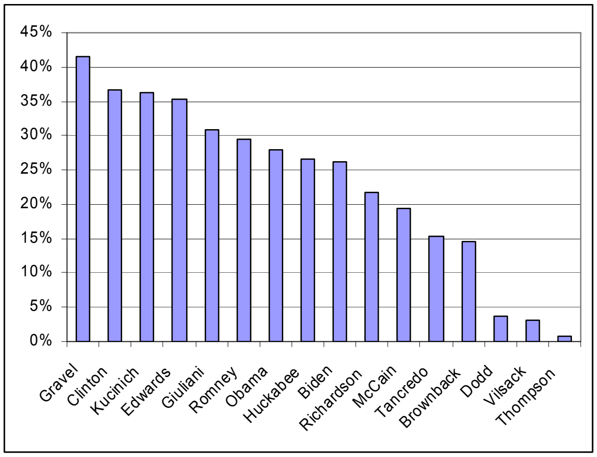

Figure 3. Percentage of YT data set videos also in the Blog data set

This analysis shows a general, but not definitive, trend that the leading candidates have larger percentages of overlap.

6.3. Relevance Analysis

To better understand the characteristics of the data sets, we conducted a relevance analysis. We hypothesized that a higher percentage of videos in the overlap set (YT∩B) would be relevant to a given candidate than those in the YouTube or Blog sets alone.

6.3.1. Method

To conduct the relevance analysis, we selected a subset of all the videos. We selected four candidates: Obama (Democrat) and Romney (Republican) because, at the time of our analysis (January 23, 2008), they were the leading candidates, based on pledged delegate count. We selected Vilsack (Democrat) and Tancredo (Republican) – neither of whom received any delegates, and both of whom have dropped out of the presidential race – because we wanted to see if there was a difference between front-runner and non-front runner candidates. For these four candidates, we selected a random sample of the YouTube, Blog, and overlap (YT∩B) data sets.

Table 5. Relevance Analysis Sample Sizes

	YouTube		Blogs		YT∩B	
	total	sample	total	sample	total	sample
Obama	1005	279	6718	364	281	163
Romney	509	220	3060	342	150	109
Vilsack	97	78	143	105	3	3
Tancredo	254	154	1130	287	39	39

When sampling very large populations, it is common to use the normal distribution to determine the sample size needed. However, for our fixed size data sets, the hypergeometric distribution provides a more accurate estimate of the sample size needed. Thus, we used the hypergeometric distribution to select sample sizes large enough to generate results with 95% confidence intervals and a range of ±5%. Across the three data sets (YouTube, Blog, YT∩B), for the four candidates, a sample of 2,143 videos were selected out of a total of 13,389. Table 5 shows the total number of videos and the number in each sample for each candidate + data set combination.

Five members of the VidArch project team coded videos. We developed an initial set of coding categories and coded a trial set of 75 videos about a different candidate. There was high inter-coder agreement for the trial set, giving us confidence in our coding scheme. Before coding the final data set, all five coders met one time and made slight modifications to the scheme and discussed clarifications. The final coding scheme is shown below.

1) Clearly relevant to the candidate – video focuses on the candidate, video issued by the candidate, or prominently mentions the candidate by name (e.g. Obama girl) where the candidate is the main focus. Does NOT have to be about the candidate in Election 2008, just prominently about the candidate.

2) About Election 2008 but not the candidate specifically – some mention of the candidate or the election 2008, but the candidate is NOT the main focus. If it would be a 1 in any other candidate's search, it is a 2 in this candidate's search. Candidate in a debate with other candidates. If it is primarily or partially about another candidate in Election 2008.

3) Not relevant – everything else, including other political stuff, or an election issue.

Of our set of 2,143 videos, a subset of 188 were randomly selected and included in each coder's set so that we could measure inter-coder agreement (this number was again chosen using the hypergeometric distribution with 95% confidence intervals and a range of ±5%). The subset was randomly chosen, but with the constraint that the proportions for each candidate + data set combination approximate the proportions of the overall set. The remaining videos were split into five groups, again retaining the proportions so that each coder received approximately the same number of videos about each candidate + data set combination. Thus, each coder assessed almost 600 videos (391 unique to them and the 188 in common).

6.3.2. Results

The results of the relevance analysis are shown in Figure 4 (note that the YouTube set is labeled "Y" in the figure). Inter-coder agreement was high (Fleiss' kappa = 89.7%) for the 188 videos coded by all coders. The main findings of the results are summarized below.

YouTube relevance ranking returned results with high precision. Overall, our YouTube data set had high percentages (>85%) of relevant results. The candidate Vilsack was lower and is discussed later in this section. Recall that the YouTube data set was collected by gathering the top 100 results for each query on a daily basis. Furthermore, each query produced a large percentage of results (>=75%) that were explicitly about the candidate of the query (i.e. rated as a "1"). This indicates that YouTube's current relevance algorithm has high precision. Note that we do not know the recall of the YouTube algorithm because we don't know the size of all the relevant videos for each query within YouTube.

Blog search engines returned pages with many non-relevant videos. Our technique of querying Google Blogsearch and Technorati with the candidate's names returned blog pages with large numbers of links to YouTube videos. However, as can be seen in Figure 4, the majority of these videos were not relevant to the candidates or to the election. Queries for Tom Tancredo and Tom Vilsack did return a sizable number of election-related videos. This is discussed in more detail later in the section. We view the blog search collection method that we used to be a straightforward approach to gather a large amount of data. These results suggest that more sophisticated techniques may be needed to mine blog pages to filter non-relevant links (e.g. analysis of text surrounding the links), especially if Blog data is being used as the only data source.

The precision of YouTube alone was not increased by intersecting the Blog data. Intersecting the YouTube and Blog data sets did not increase the precision of the results over YouTube alone. This result was contrary to our expectations. We expected that triangulation of data sources would yield higher percentages of relevant videos than any one source alone. One reason for this result is that YouTube alone had high precision, so there was limited room for improvement. It is interesting that intersecting the Blog data did in many cases reduce the percentage of videos that were not specifically about the candidate, but that were about the election. Depending on the goals of the curator, this could be desirable or undesirable (e.g., if collection decisions are automated, picking higher precision result sets of videos at the sacrifice of missing a few relevant videos).

The intersection of YouTube and Blog data does increase evidence of relevance and provides multiple sources for contextual information. Perhaps the largest advantages to looking at the overlap between two (or more) data sources are that (1) if each method is independent, it gives the curator or archivist additional evidence that a particular video is important, and (2) each independent method provides a "check" against potential changes or biases that may be out of the curator's control. For example, the details of the YouTube relevance ranking algorithm are not known and could be changed without a curator's knowledge. Not only does intersecting the YouTube results with another source such as the Blog search provide additional evidence of importance, it provides a "cushion" against changes in one source. It also gives curators a stronger justification and explanation of why items were selected than just relying on one "black-box" search ranking algorithm that the details of which are not known. For the overlapping sets of videos, not only are a high percentage of them relevant, but additional contextual information is available from the Blog pages beyond what would be found on YouTube alone.

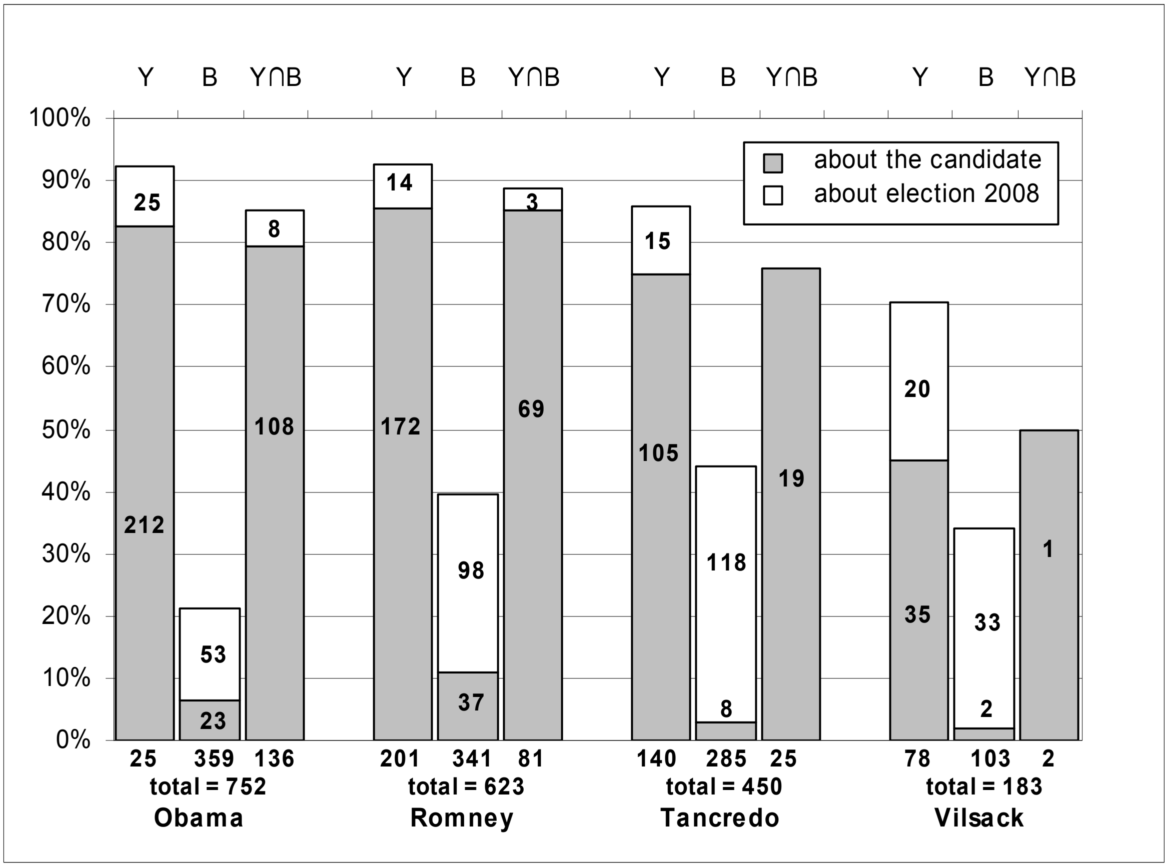

Figure 4. Relevance of videos in each data set (grouped by candidate)

A note about Vilsack percentages. An unusually large portion of the videos from the Vilsack blog data fell into the "related to election but not related to the candidate" category. This appears to be a result of 1) Vilsack having a low number of videos, and 2) one of the blog pages having a large number of video links down the left side (outside the content of individual blog entries), many of which are related to the election but not to Vilsack specifically. Fully 66% of the these links to videos were from this one blog. We suspect that a similar situation occurred for Tancredo, although we have not verified this. These findings suggest that the utility and meaning of blog out-link data can be highly contingent on the collection size and internal structure of blog pages being crawled.

In addition, for the Vilsack query, a lower percentage of videos in the YouTube data set were specifically about Vilsack than was typical for other candidate searches and a larger percentage were about the election in general. We believe that this was due in part to the relatively low number of total videos about Vilsack on YouTube. A search on YouTube for "Tom Vilsack" on February 3, 2008 returned a total of only 91 videos. This means that no filtering was taking place by retrieving only the top 100 videos returned for Vilsack.

7. DISCUSSION

The results of these investigations reflect the complexity of making selection decisions in order to build collections of digital objects from highly inter-linked, dynamic and interactive environments. Analyses of web media behavior may eventually approach the effectiveness of Nielsen ratings for TV viewing or political polls for elections, which could be very helpful in determining how to focus collecting activities. However, clear and well-recognized metrics of societal "attention" are not yet available for environments such as YouTube and the blogosphere. Rather than providing political analysis, the VidArch project is exploring cases such as the documentation of the U.S. presidential elections, in order to test and evaluate approaches for identifying, monitoring, selecting and capturing video content that is relevant to the collecting missions of digital archivists.

Given the generally low overlap in videos in our YouTube and Blogosphere data, we can infer that collections based on crawls of these two environments would tend to look very different. It is not clear whether this is due to the 100 result cut off for each YouTube query, the limitations of the queries themselves (not all variants of candidate names were used), the proprietary ranking algorithms of YouTube, Google Blogsearch, and Technorati, or the nature of the tags, titles, and other metadata that support video mining today. The relevance assessments show that there is some noise in the results from YouTube and the blogosphere, however the high precision tendencies search engines show carry over to YouTube, thus yielding high precision in the result sets it returns for well-formed queries. This suggests that harvesting via query rather than link traversal is a sensible strategy as long as APIs are available.

Our first research question (RQ 1) was whether information from secondary source environments (including secondary relevance measures) help curators of digital collections to identify videos that are relevant to their collections, which they would not discover solely by consulting information (including primary relevance measures) from the primary use environment. Analysis of data from our crawls of YouTube and the blogosphere suggest that the answer to RQ 1 is a qualified "yes." Our crawls of the Blogosphere identified videos that were relevant and that might have important qualitative differences from the videos collected using the YouTube search algorithm, which would warrant the use of both crawling methods. This is something we hope to explore more in future work. If one's collecting mission is to get as much relevant material as possible, even if this means also collecting material that is not directly relevant (i.e. valuing recall higher than precision), then combining techniques could be particularly desirable.

Our study also suggests that further refinement of blog crawling methods could have important pay-offs for building web collections. For example, we discovered many blog out-links that came from parts of the page outside the individual blog entries. Failure to parse out – or at least recognize and accommodate for – such out-links could result in a distorted picture. We saw this particularly with Vilsack, whose small set of video links were significantly skewed by one blog that provided many election-related links in its pages.

Our second research question (RQ 2) was whether, a set, S_1, of digital objects that have been deemed potentially relevant, based on primary relevance measures (in this case, YouTube relevance rankings), secondary relevance measures (in this case, out-links to YouTube videos from crawled blog pages) act as a useful filter for identifying a subset of S_1 that is particularly important to ingest into a digital collection? Once again, our data suggest an answer of qualified "yes."

The answer to RQ 2 in relation to the videos identified as "relevant to the candidate" is less conclusive. Our coding methodology was only based on a 3-point relevance scale. Of the videos that we identified as "relevant to the candidate," we did not then provide any further ranking of relevance or importance. It is quite possible that the subset of relevant videos that appear in both the YouTube and blogosphere data are even more "core" to a collection about that candidate than the YouTube-derived set as a whole.

We do feel confident that the query approach to harvesting is sensible and cost-effective. Rather than a crawl of links, queries via an API were efficient and returned large volumes of relevant videos, including ones that were under discussion in the blogosphere. Thus, we will continue to use query-based approaches in the context mining tools we are building.

8. CONCLUSION AND FUTURE WORK

Potential approaches for identifying and monitoring online resources in order to build digital collections can vary across at least three dimensions: environments crawled (e.g. blogosphere, YouTube), access points from those environments used as crawling or selection criteria (e.g. number of views, primary relevance based on term matching, number of in-links, channel or account from which an item was submitted), and threshold values for scoping capture within given access points (e.g. 100 most relevant query results, at least 5 in-links). The findings in this paper suggest potential implications of choosing or combining two specific crawling approaches. We see great promise and value in further investigation of crawling approaches that integrate environments, access points and thresholds in different ways, in order to determine the combinations that can best meet that needs of those who are building and managing digital collections.

We have the opportunity to compare data with the Library of Congress (LC) and Internet Archive (IA), who are working together to harvest 2008 Presidential campaign videos. In contrast to the VidArch approaches reported in this paper, LC/IA are starting with the URL of each candidate's official web site as seeds. They are then capturing the web sites themselves, as well as any YouTube video to which there are direct out-links from the candidate sites. The products of this approach will include many clear links between candidates and the videos that they have either produced or are otherwise promoting, which is very valuable contextual (provenance) information. However, it is not likely to capture many of the "viral" videos that play a major role in the election – often more influential than the videos disseminated by the candidate's themselves. Over the next several months, we will be working with LC/IA to identify differences and complementarities between the crawling approaches that we have taken.

Most viable efforts to build digital collections will rely heavily on software to help in scoping, filtering and describing the digital objects of interest. Item-level examination of all potentially relevant information will not be possible. The VidArch project has been developing tools and techniques for the automation of many aspects of the crawling and selection process. As with all other collecting building activities, human evaluation and judgment will be essential, but it must be focused on specific decision points (e.g. identifying the topics and entities to be documented within a collection; picking the most appropriate environments to crawl: setting and revising crawl criteria). The research reported in this paper has demonstrated the value of combining system-generated metrics with human assessments of relevance.

Our future iterations of tools and approaches will be based on further collecting and analysis of both quantitative and qualitative data. For example, one of the hypotheses of the VidArch project is that the collection of blog pages can provide useful contextual information to supplement what is in YouTube about particular videos. Our initial examination of many of the pages in our blog data suggest that they could indeed provide valuable contextual information. However, their utility is likely to be quite variable and may require further scoping and filtering. We plan to analyze the content of blog postings in order to better understand (1) what sort of useful contextual information they might provide, and (2) the reasons for the "false positives" in our data (links to videos that were not actually about the given candidate or the election), so we can investigate techniques for reducing them.

The data and findings in this paper are based specifically on crawls related to the U.S. presidential election. However, we are most interested in the implications for online collection building in general. This study can inform other collecting activities that are based on pre-defined individuals or events. For several months, we have also been running crawls in YouTube based on other topic-related queries. We plan to further investigate the results of these crawls.

Clearly, curators who use systems and techniques such as those discussed here must consider the primary and secondary use environments to include; carefully compose query sets; establish result set thresholds; and determine what link types to include in secondary use environments; all within the constraints of their local selection policies and resources to assess results.

9. ACKNOWLEDGMENTS

This work has been supported by a grant from the NSF/Library of Congress (IIS 0455970). Special thanks to Laura Sheble and Sarah Jorda for coding videos. Thanks are also due to the VidArch team that includes Brenn Hill, Paul Jones, Sarah Jorda, Laura Sheble, Yaxiao Song, and Helen Tibbo.

10. REFERENCES

1. Adamic, L.A. and Glance, N. 2005. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, ACM Press, New York, NY, 36-43.
2. Barabási, A., and Albert, R. 1999. Emergence of scaling in random networks. Science, 286:509-512, October 15, 1999.
3. Bergmark, D. 2006. Collection Synthesis. In *Proceedings of the 2th ACM/IEEE-CS Joint Conference on Digital Libraries*, Portland, OR, *Chapel Hill, NC, July 14-18, 2002*, ACM Press, New York, NY, 253-262.
4. Business Intelligence Lowdown, November, 2007. http://www.businessintelligencelowdown.com/2007/02/top_10_largest_.html
5. Brown, A. 2006. *Archiving Websites: A Practical Guide for Information Management Professionals*. Facet, London.
6. Drezner, D. and Farrell, H. 2004. The Power and Politics of Blogs. In *Annual Meeting of the American Political Science Association*, Chicago, IL.
7. Gueorguieva, V. 2007. Voters, MySpace and YouTube: the Impact of Alternative Communication Channels in the 2006 Election Cycle and Beyond. *Social Science Computer Review*, doi:10.1177/0894439307305636.
8. Hindman, M., Tsioutsiouliklis, K. and Johnson, J. 2005. Measuring Media Diversity Online and Offline: Evidence from Political Websites. The 32nd Research Conference on Communication, Information and Internet Policy.
9. Keelan, J., Pavri-Garcia, V., Tomlinson, G., and Wilson, K. 2007. YouTube as a Source of Information on Immunization: A Content Analysis. *Journal of the American Medical Association*, 298(21): 2482-2484.
10. Masànes, J. 2006. Archiving the Hidden Web. In *Web Archiving*. Springer, New York, 115-29.
11. Ntoulas, A., Zerfos, P. and Cho, J. 2005. Downloading Textual Hidden Web Content through Keyword Queries. In *Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries, Denver, CO, USA, June 7-11, 2005*. ACM Press, New York, 100-109
12. Panagopoulos, C. 2007. Technology and the Transformation of Political Campaign Communications. *Social Science Computer Review 25(4):*423-424.
13. Raghaven, S. and Garcia-Molina, H. 2001. Crawling the Hidden Web. In Proceedings *of the 27th International Conference on Very Large Data Bases*. San Francisco, CA, Morgan Kaufmann, 129-138.
14. Shah, C., and Marchionini, G.. 2007. Preserving 2008 US Presidential Election Videos. In the Proceedings of International Web Archiving Workshop (IWAW) 2007.
15. Spink, A., Jansen, B., Blakely, C., and Koshman, S. 2006. A Study of Results Overlap and Uniqueness Among Major Web Search Engines. Information Processing and Management 42(5):1379-1391.
16. Tibbo, H., Lee, C., Marchionini, M., Howard, D. 2006. VidArch: Preserving Meaning of Digital Video over Time through Creating and Capture of Contextual Documentation. IS&T Archiving 2006.
17. Wikipedia page, “United States presidential election, 2008”, http://en.wikipedia.org/wiki/United_States_presidential_election,_2008#Candidates_and_potential_candidates. Accessed on April 27, 2007.