Preserving 2008 US Presidential Election Videos

Chirag Shah
School of Information and Library Science
University of North Carolina
Chapel Hill NC 27599, USA
chirag@unc.edu

Gary Marchionini
School of Information and Library Science
University of North Carolina
Chapel Hill NC 27599, USA
march@ils.unc.edu

ABSTRACT

Online digital video hosting and sharing sources are becoming increasingly popular. People are not only uploading and viewing videos at these sites, but are also discussing them. In addition to the comments, visitors also rate and link these videos. Statistics about these and other user behaviors on such sites can tell us a lot about social trends and patterns, which in turn can serve as important context to aid in interpreting archival materials. We are interested in analyzing the correlation between this behavior and an actual social event. We chose the 2008 US Presidential Election as the event and YouTube as the video source. In this paper we report a system that facilitates harvesting presidential election videos from YouTube along with corresponding metadata and contextual information. The system is built to revisit the same video pages periodically and collect the contextual information for every visit. From this data we are hoping to understand various social behaviors that may help us finding some interesting connections with the real-life events and results and inform digital video curatorial policies.

Categories and Subject Descriptors

H.3.6 [**Information Storage and Retrieval**]: Library Automation

General Terms

Design, Human Factors, Management

Keywords

Preservation, Video harvesting, YouTube, Presidential elections, Metadata, Social context

1. INTRODUCTION

Corporations and governments have long recognized the importance of preserving records to serve decision-making and legal requirements. Cultural institutions such as archives, libraries and museums also have long aimed to preserve cultural artifacts and information to document human existence and inform future generations. As more information is produced, distributed, and stored in digital form, the problem of long-term preservation has raised new kinds of challenges for archivists and records managers. An active research and development effort is underway to address these challenges.

Our work in this area focuses on preserving digital video, a medium that is taking on increasing importance as tools to create and distribute it become common. Our work is rooted in two beliefs: First, it is the responsibility of scholars to preserve content that is beyond the massively popular content that mainstream media organizations will preserve and that will be copied by multitudes of people and by various cultural institutions (the LOCKSS hypothesis [3]). Thus we aim to identify and preserve video that includes items below the mainstream, possibly ephemeral in nature. Second, preserving the digital video bits is useless without the metadata that makes finding and viewing possible and the contextual information that makes interpretation possible. Determining what context to include is a kind of collection policy and thus much of our attention has focused on this specific issue.

Our emerging framework for context [5] includes several facets, however, in this paper we focus on temporal and social facets: how context evolves over time, and how the digital medium facilitates and propagates social interactions around video objects. We first describe a system that harvests video, accompanying metadata, and sets of temporal and social contextual information. The system is a component of a generalized video curator?s toolkit that will allow curators to specify sources as well as harvesting policies for video, metadata, and context. The second part of the paper discusses the metadata and context policies and describes the preliminary decisions we have made to date.

Before we proceed, let us try to understand temporal and social context with respect to digital preservation.

Temporal context. Artifacts change over time through physical deterioration and information changes over time through use and interpretation. We view a digital information archive as dynamic, requiring ongoing maintenance and updating. This is in stark contrast to physical archives that grow but depend on maintaining static artifacts.

Social context. Digital access allows many more people to

IWAW'07. June 23, 2007, Vancouver, British Columbia, Canada.

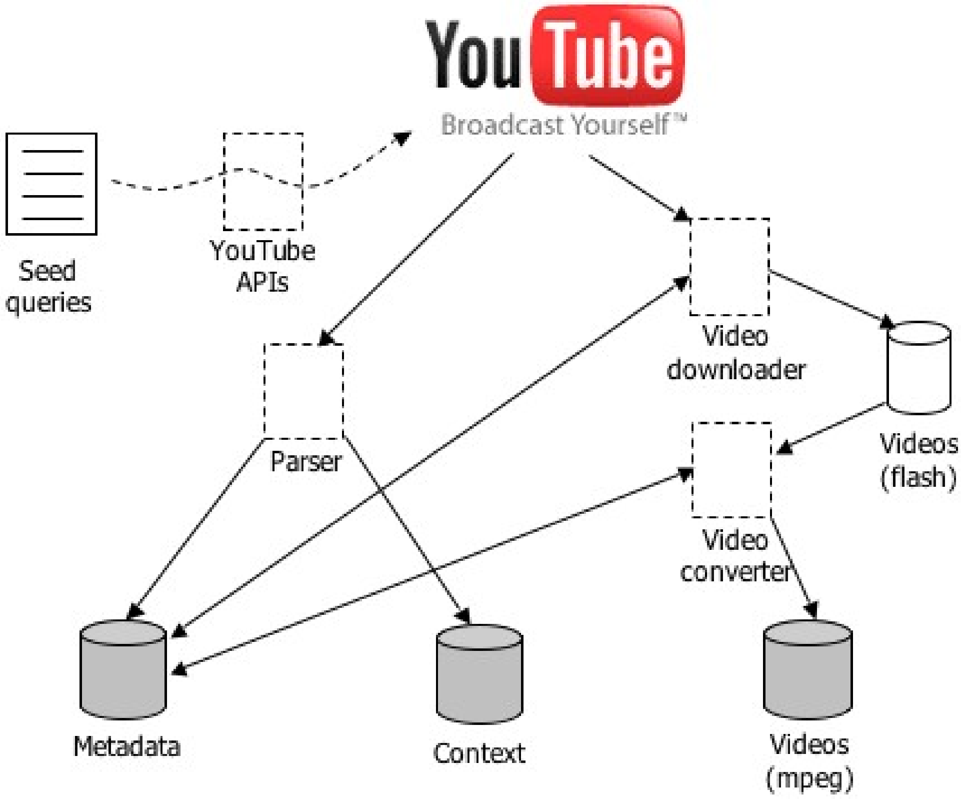


Figure 1: YouTube video harvesting and monitoring

use, interact with, annotate, and share archival materials. With the exponentially increasing participation of web users in blogs, forums, and social networks, studying their online behavior has become an essential part of learning about various social trends and patterns. For many organizations, it has become very important to know what content people are reading or watching, what they are talking about them, and how these activities reflect and affect their lives. For archival materials, these social interactions serve as an important kind of context.

In the work reported here, we are interested in analyzing people's online video watching and their participation for discussions relating to 2008 US Presidential Elections. We chose this event because it is one of the most important political events in the US; a large portion of the US population participates in discussions and consumes a large amount of information related to the elections. We chose to analyze video postings and user feedback on YouTube [7] because it is one of the largest online video sharing sources. As of August 2006 [4], YouTube had hosted more than 6 million videos; and the total time people spent watching these videos since its inception in February 2005 was 9,305 years.

In this paper we describe a system that we have built for harvesting YouTube videos related to 2008 US Presidential Elections. The rest of the paper is organized as follows. In section 2, we provide the overall design schema for the system and some of the implementation details. Specific collection policy issues are discussed in section 3. The paper concludes in section 4 with some pointers to the future work.

2. DESIGN

We ultimately aim to provide a curator?s toolkit that is easy to use and mainly runs automatically. The system described here is a first pass to demonstrate feasibility and gather enough data to investigate curatorial policies for context mining and to inform system design parameters. We implemented the system to take a set of seed queries and then use them to retrieve videos on YouTube. Thus, the curator expresses the collection development policy by specifying queries and the system collects and organizes results. The selection of these seed queries and other policy decisions are described in the next section. The architecture of our system is given in Figure 1. Following is a brief description of its workflow.

1. The curator provides a set of seed queries to monitor (Figure 2).

2. The system uses these queries to go out and search on YouTube.

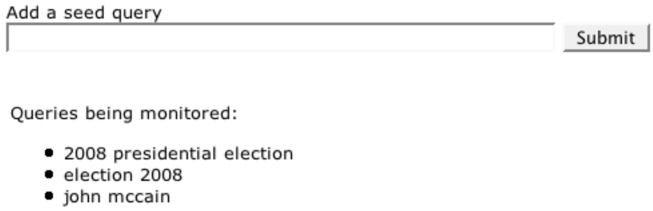
3. A set of metadata is extracted from a subset of the results returned from YouTube. We define metadata to be the information about the given video which are provided by the author of that video, and are usually static in nature. For instance, the genre of the video (called 'channel' on YouTube).

4. The Video downloader component checks the metadata table to see which videos have not been previously downloaded and grabs those videos in Flash format.

5. The video converter component checks which videos are downloaded and not converted, and converts them into MPEG format.

6. The context capturing component goes out to YouTube and captures various contexts about the video items for which the metadata is already collected. Each time context is captured, a time-stamp is recorded. We define social context as the data contributed by visitors to the video page. This would include fields such as ratings and comments. Note that other types of social context in blogs and other sources could also be harvested with different components. The context capturing component runs periodically and updates time-sensitive data such as new comments or video postings, thus capturing temporal context.

Thus, there are four major processes: (1) metadata collection, (2) context collection, (3) video downloader, and (4) video converter. Each of these parts can be run independently and they all will check the overlapping functions with other parts to guarantee consistency and integrity of the whole system.



Queries being monitored:

- 2008 presidential election
- election 2008
- john mccain

Figure 2: Providing seed queries

We implemented the system using PHP 5 and MySQL 5. We also used YouTube APIs [8] to access some of the services from YouTube. The data, in our case, are videos from YouTube. Since these videos are embedded in YouTube web-pages and viewed by streaming, we had to use some mechanism that would allow us to download and store those videos. We used a freely available tool called 'youtube-dl' [2] for downloading the video in Flash format from a YouTube page. In addition to this, we use FFmpeg tool [1] to convert downloaded Flash videos to MPEG format. The system is operational and we are using it to collect a test set of data. As of June 1, 2007, 35 crawls (one each day) were conducted for 56 different queries related to the presidential campaign. These crawls yielded 5129 unique videos. The range of number of viewings per query was enormous, ranging from 123 for one candidate (Rebecca Rotzler) to 29165 for another person who is not currently a candidate (Al Gore). Likewise the number of comments posted varied dramatically. Our next steps will be to systematically investigate patterns over time and the kinds of triggers that change use and social interaction. We postulate two kinds of triggers: predictable (e.g., a debate, a primary election) and unpredictable (e.g., a speech or public appearance snafu, personal event). By examining instances of these two kinds of triggers over time, we hope that we can make recommendations for where curators should put their computational resources and personal analysis time when mining context for their video collections.

3. POLICY DECISIONS

Two main kinds of policies are required in any preservation effort: what to collect (the appraisal process) and how to collect it. In our case, policies must be established for the videos, metadata, and context. In the case of the primary video data, we have identified a small number of classes of video and are in the process of identifying others. Several specific questions related to these two kinds of policy follow.

Specific *what* policy questions:

1. What genre of video?
2. Which metadata to include?
3. Which context to include?

Specific *how* policy questions:

1. Which sources to use to find the video?
2. What queries should be used to find the video, including lexical variants?
3. To what depth in the results lists should video be harvested?
4. How frequently should data be harvested?
5. Should links to other sites be followed and if so, to what depth?
6. How should video responses (videos uploaded by the public as comments on the video) be handled?
7. How should duplication be handled while still acquiring comment and download data?

For this paper, we consider the 2008 US Presidential campaign videos hosted on YouTube as answers to the genre and source questions. In order to search for these videos, we decided to run a set of queries on YouTube. To determine queries, we decided to use all the present presidential candidates' names available from WikiPedia [6] as queries. In addition, we also chose the following six general queries: *election 2008, US election 2008, United States election 2008, presidential election 2008, campaign 2008, decision 2008.* This yielded a total of 56 queries. To simplify the process in this early state, the queries are entered as free form text (no special delimiters such as quotes) and candidate names are exactly as expressed in the Wikipedia listing rather than all variants.

The question of what metadata was answered as follows. We collect the following metadata: title, description, categories, time-stamp, the query for which a given video was retrieved, username of the user who posted that video, date when the video was uploaded, time when it was uploaded, URL of the video page, URL of the thumbnail image, tags, recording location and country. A snapshot of collected metadata for various videos is given in Figure 3.

The question of what context is complex at the heart of our research effort. For these preliminary investigations, we collect the following temporal and social contextual information: rank of the video for a given query, number of views, number of ratings, average rating, number of comments, all the comments along with the user names, time-stamp, number of times the video was favorited, number of links coming to that video, number of clicks from the sites that link to the video, number of times the video received some honor (e.g., best in comedy, video of the week, etc.), and time when it was last updated. A snapshot of collected context for various videos is given in Figure 4.

There is no experiential or literature to guide a decision on how many results to harvest for each query. For convenience, we could simply get the top 20 results for each query. However, we do not know how the results are ordered and what the rank of a result signifies. We choose to run a pilot run for a month collecting top 100 results for each query to see how different results we get every time we crawl. Although it is too soon to tell what kinds of ?churn? occurs in the top 100 ranks over time, our preliminary estimates are that the top 100 give suitable coverage and do not severely burden the system resources.

The frequency decision is also tricky. Some preliminary run data suggested that there are no significant differences between data collection at one day's time interval when only top 20 results are considered. However, no significant political event happened during those crawls. Once again, in order to understand the effect of different sampling period, we decided to crawl once every day. Given that the load has not been overly burdensome, we will continue to crawl on a daily basis.

4. CONCLUSIONS

This paper outlines the first steps in an ongoing effort to capture and preserve context for digital video. Both temporal and social contexts are taken to be crucial to future understanding and interpretation of the video, requiring an ongoing data collection effort. Automated systems guided by human-generated and controlled collection policies are the only workable approach to such dynamic data collection. This paper presents such a system and some of the policy decisions behind its operation.

At present we have completed a pilot run for a month and offer the following observations: Harvesting videos from systems such as YouTube that publish an API is quite feasible and does not require enormous system resources (good bandwidth and a few terabytes of disk). Conducting daily crawls to the depth of 100 returns per query is highly feasible. New kinds of analytical tools for analyzing the collected data are required, especially with respect to changes over time. Textual comments in this particular domain yield much chaff?ranging from the profane to the arcane. Query variations make some difference in results and deserve further investigation. Context mining policies will likely be distinct for two kinds of event ?triggers? that cause large changes in ranks of results and interactions with specific videos. Examination of the ?churn,? the ebb and flow of temporal and social context each over time are especially interesting for this genre of heavily contemporary video. Comparing this to video that is less event-driven will be important in our sister studies.

In the coming months we will continue collecting the data, metadata, and contexts from YouTube. After this trial phase, we shall be able to make more informed policy decisions about some of the questions that we discussed in the previous section. We plan to keep harvesting and preserving such data, metadata, and context at least until the presidential inauguration in January of 2009. We have also begun similar crawls for several other topics that will provide us with data that will allow more generalizable suggestions for curatorial policies. We believe our collection will serve as a valuable resource for understanding some of the social trends and patterns relating to the elections during this time period and inform the continued development of tools for digital video curators.

5. ACKNOWLEDGMENTS

The authors wish to thank Helen Tibbo, Cal Lee, Paul Jones, Terrell Russell, and Yaxio Song for their valuable inputs in this project. This work is supported by a grant from the National Science Foundation and the Library of Congress NDIIPP.

6. REFERENCES

[1] FFmpeg. FFmpeg. http://ffmpeg.mplayerhq.hu/ [Accessed: April 27, 2007], 2007.

[2] Ricardo Garcia Gonzalez. youtube-dl: Download videos from YouTube.com. http://www.arrakis.es/ rggi3/youtube-dl/ [Accessed: April 27, 2007], 2006.

[3] Vicky Reich and David S. H. Rosenthal. LOCKSS: A Permanent Web Publishing and Access System. *DLib Magazine*, 7(6), June 2001.

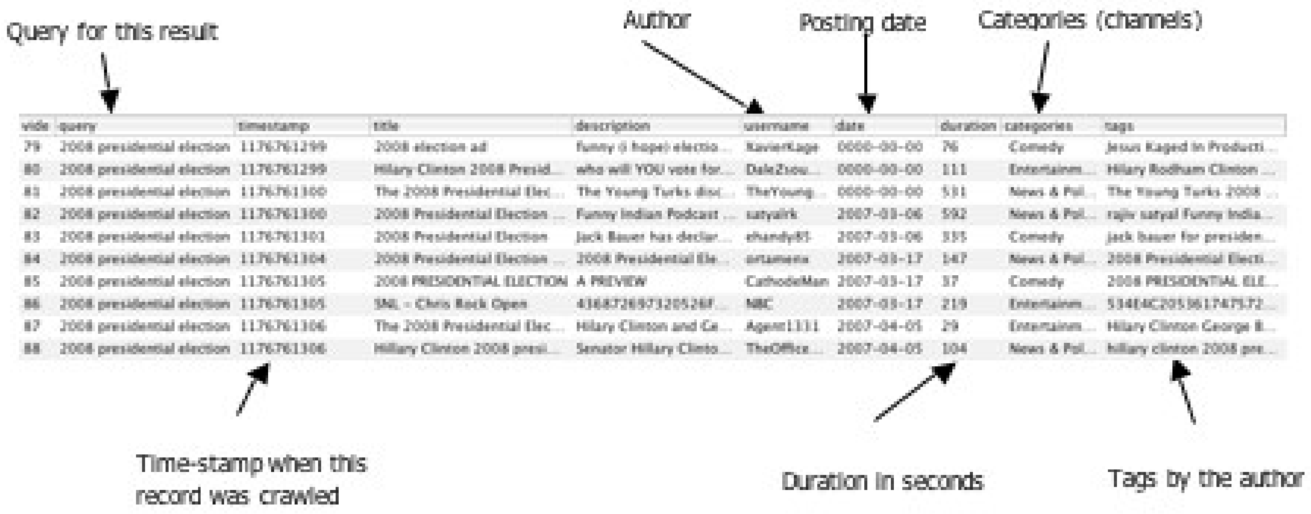
[4] Steve Rubel. Micro Persuasion: YouTube by the Numbers. http://www.micropersuasion.com/2006/08/youtube_by_the_.html [Accessed: March 1, 2007], August 2006.

[5] Helen R. Tibbo, Chirstopher A. Lee, Gary Marchionini, and Dawne Howard. VidArch: Preserving Meaning of Digital Video over Time through Creating and Capture of Contextual Documentation. In *Proceedings of Archiving*, pages 210–215, 2006.

[6] WikiPedia. United states presidential election, 2008 - wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/United_States_presidential_election,_2008#Candidates_and_potential_candidates [Accessed: April 27, 2007], 2007.

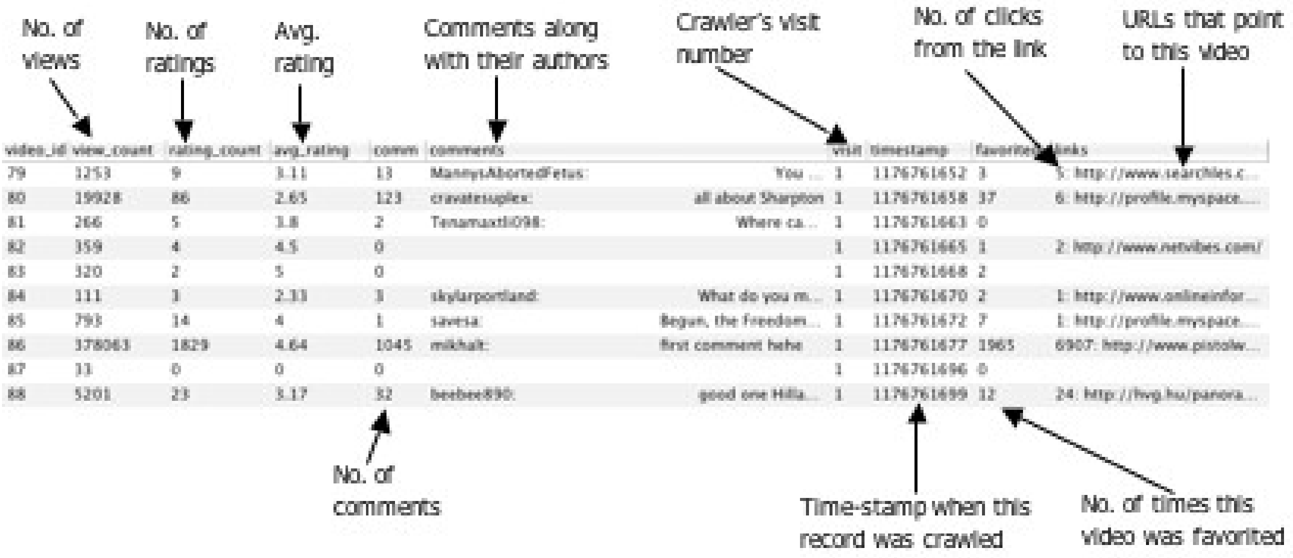
[7] YouTube. YouTube: Broadcast Yourself. http://www.youtube.com [Accessed: March 1, 2007], 2007.

[8] YouTube. Youtube: Broadcast yourself. http://www.youtube.com/dev [Accessed: April 27, 2007], 2007.



vide	query	timestamp	title	description	username	date	duration	categories	tags
79	2008 presidential election	1176761299	2008 election ad	funny (i hope) electio...	XavierKage	0000-00-00	76	Comedy	Jesus Kaged In Producti...
80	2008 presidential election	1176761299	Hilary Clinton 2008 Presid...	who will YOU vote for...	DaleZsou...	0000-00-00	111	Entertainm...	Hilary Rodham Clinton ...
81	2008 presidential election	1176761300	The 2008 Presidential Elec...	The Young Turks disc...	TheYoung...	0000-00-00	531	News & Pol...	The Young Turks 2008 ...
82	2008 presidential election	1176761300	2008 Presidential Election ...	Funny Indian Podcast ...	satyalrk	2007-03-06	592	News & Pol...	rajiv satyal Funny India...
83	2008 presidential election	1176761301	2008 Presidential Election	Jack Bauer has declar...	ehandy85	2007-03-06	335	Comedy	jack bauer for presiden...
84	2008 presidential election	1176761304	2008 Presidential Election ...	2008 Presidential Ele...	ortamera	2007-03-17	147	News & Pol...	2008 Presidential Electi...
85	2008 presidential election	1176761305	2008 PRESIDENTIAL ELECTION	A PREVIEW	CathodeMan	2007-03-17	37	Comedy	2008 PRESIDENTIAL ELE...
86	2008 presidential election	1176761305	SNL - Chris Rock Open	4368726973205 26F...	NBC	2007-03-17	219	Entertainm...	534E4C205361747572...
87	2008 presidential election	1176761306	The 2008 Presidential Elec...	Hilary Clinton and Ge...	Agent1311	2007-04-05	29	Entertainm...	Hilary Clinton George B...
88	2008 presidential election	1176761306	Hillary Clinton 2008 presi...	Senator Hillary Clinto...	TheOffice...	2007-04-05	104	News & Pol...	hillary clinton 2008 pre...

Figure 3: Sample of collected metadata



No. of views
No. of ratings
Avg. rating
Comments along with their authors
Crawler's visit number
No. of clicks from the link
URLs that point to this video
No. of comments
Time-stamp when this record was crawled
No. of times this video was favorited

video_id	view_count	rating_count	avg_rating	comm	comments	visit	timestamp	favorites	links
79	1253	9	3.11	13	MannysAbortedFetus: You ...	1	1176761652	3	5: http://www.searchles.c...
80	19928	86	2.65	123	cravatesuplex: all about Sharpton	1	1176761658	37	6: http://profile.myspace....
81	266	5	3.8	2	Tenamaxtli098: Where ca...	1	1176761663	0	
82	359	4	4.5	0		1	1176761665	1	2: http://www.netvibes.com/
83	320	2	5	0		1	1176761668	2	
84	111	3	2.33	3	skylarportland: What do you m...	1	1176761670	2	1: http://www.onlineinfor...
85	793	14	4	1	savesa: Begun, the Freedom...	1	1176761672	7	1: http://profile.myspace....
86	378063	1829	4.64	1045	mikhalt: first comment hehe	1	1176761677	1965	6907: http://www.pistolw...
87	33	0	0	0		1	1176761696	0	
88	5201	23	3.17	32	beebee890: good one Hilla...	1	1176761699	12	24: http://hvg.hu/panora...

Figure 4: Sample of collected contextual information