

# Supporting Research Data Collection from YouTube with *TubeKit*

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

## Abstract

We present *TubeKit*, a query-based YouTube crawling toolkit. This software is a collection of tools that allows one to build one's own crawler that can crawl YouTube based on a set of seed queries and collect up to 17 different attributes. *TubeKit* assists in the phases of this process starting with database creation to finally giving access to the collected data with browsing and searching interfaces. We further demonstrate how we used this toolkit to collect elections related data from YouTube for nearly two years. Some analysis of the collected data relating to the elections is also given.

**Keywords:** YouTube crawling, video data collection, presidential elections 2008

## 1 Introduction

Ever since its inception in 2005, YouTube has emerged as a premium forum for hosting online videos. In this time, YouTube has become much more than posting, viewing, and sharing digital videos; it has become a platform where people express their opinions, participate in discussions, and voice their issues in many creative ways (Gomes, 2006).

While the YouTube platform caters to the video publishing and consuming needs of anyone, it has also become an essential tool for political parties and campaigns for getting their messages and propaganda out to its audience. The 2008 presidential election was unique in that it was the first election where a tool like YouTube was used very extensively, creatively, and methodically for the first time (Dalton, 2007; Jarvis, 2007; Seelye, 2007). Due to its large impact on political movements and public opinions, it became essential for anyone - political and social scientists,

archivists, curators, information scientists, journalists, and librarians - interested in studying the elections to monitor and analyze YouTube activities around the elections.

Our interest in such analysis was initially motivated from the preservation point of view. As a part of VidArch project,<sup>1</sup> funded by the Library of Congress, we wanted to collect and archive election-related videos from YouTube. Our interest was not only in harvesting the videos, but also collecting their attributes, such as title, tags, ratings, and comments, and do so over a period of time. During the spring of 2007, when we embarked upon this project, we did not find good tools to collect such data from YouTube. We, therefore, started building our own set of tools. The result was *TubeKit* - a toolkit that assisted us in creating customized crawlers that could harvest the videos and related attributes based on running a set of queries.

As we continued collecting this data from YouTube, we realized that the kind of rich information we were gathering could help us analyze the aspects of the data beyond those relating to preservation. This paper depicts our journey to creating the tools to harvest such data, the collection that we developed, the analysis that we performed, and the lessons that we learned in this process lasting nearly two years.

## 2 Development

As mentioned before, we were interested in not only collecting pages and videos from YouTube, but a set of specific attributes, such as title, description, tags, ratings, and comments. Using typical crawling tools such as 'wget'<sup>2</sup> or 'Heritrix'<sup>3</sup> on YouTube could extract links and other information. However, a major problem with such an approach was the constantly changing site and page structure of YouTube. Ever since Google acquired YouTube, we have seen many modifications in YouTube's interface. This makes extracting specific attributes hard. We were also not interested in broad crawling; rather, we wanted to crawl the data that related to the elections only. Due to these two major criteria, we decided to use a query-based focused crawling approach and use YouTube APIs as much as possible. Such focused crawlers are highly desired in narrow domains, vertical portals, and for mining the web-spaces for specific entities (Chakrabarti, Berg, & Dom, 1999).

---

<sup>1</sup><http://ils.unc.edu/vidarch/>

<sup>2</sup><http://www.gnu.org/software/wget/>

<sup>3</sup><http://crawler.archive.org/>

The design of our crawler is shown in Figure 1 and was first presented in (Shah & Marchionini, 2007). Following is a brief description of its workflow.

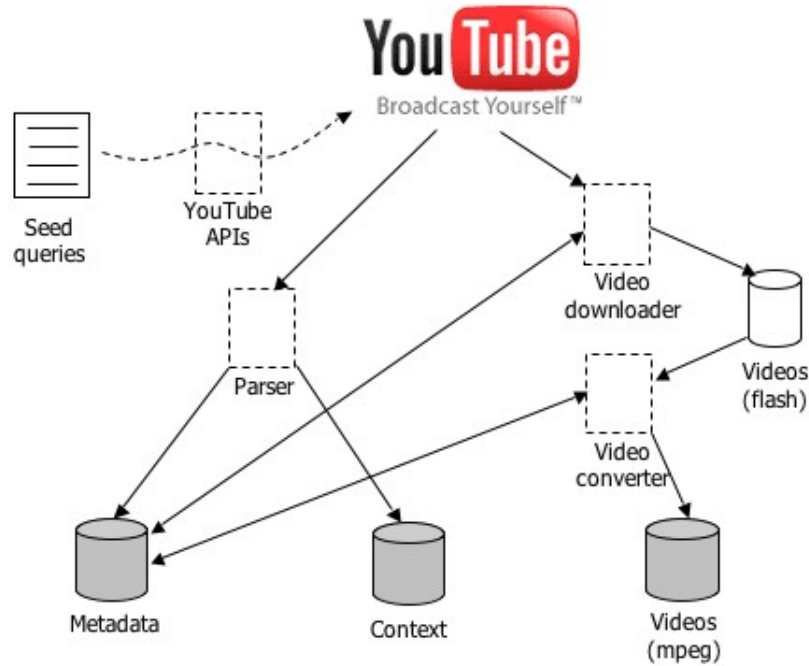


Figure 1: Our scheme for query-based YouTube crawling

1. The user provides a set of seed queries to monitor.
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.
4. The video downloader component checks the metadata table to see which videos have not been previously downloaded and collects those videos in flash format from YouTube.
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 contextual information about the video items for which the metadata is already collected. Each time

such social context is captured, a time-stamp is recorded. We define social context as the data contributed by the visitors to a 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 (discussed later). The context capturing component runs periodically and updates time-sensitive data such as new comments or video postings, thus capturing temporal context.<sup>4</sup>

Thus, there are four major processes of our focused crawler: (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 facilitate consistency and integrity of the whole system.

As we finished building our YouTube crawler for the elections, we also had the need to create such focused YouTube crawlers for other topics. Instead of building these crawlers individually, we created *TubeKit* - a toolkit that can let anyone build a crawler based on the scheme given in Figure 1.<sup>5</sup> *TubeKit* is primarily built using PHP and MySQL. *TubeKit* has been tested on Linux and Mac and should work fine with any other UNIX-based system. Its web-based interface lets one configure and monitor a crawler with minimal efforts. *TubeKit* is available to the public for free at <http://www.tubekit.org> under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States License. The following section describes the usage of *TubeKit*.

### 3 Work Flow

In this section we will demonstrate how to create a query-based focused YouTube crawler using *TubeKit*. Following are the steps to build such a crawler.

1. *Provide basic information (project name, directory to store the crawler, etc. Figure 2).*

*TubeKit* uses MagpieRSS<sup>6</sup> for some of the parsing processes and youtube-dl<sup>7</sup> for downloading

---

<sup>4</sup>Now on we will refer to social or temporal context as simply contextual information.

<sup>5</sup>The original election crawler was created before *TubeKit*, but we introduced many enhancements to it after building *TubeKit*. Today, the crawlers created using *TubeKit* can expect similar interface and functionalities as shown here.

<sup>6</sup><http://magpierss.sourceforge.net/>

<sup>7</sup><http://www.arrakis.es/~rggi3/youtube-dl/>

flash videos. One needs to provide the locations of these freely available tools during the basic configuration.

Basic Configuration	
Name of the project	elections Prefix el
Directory to store the crawler	/home/projects/elections2008
Path to Magpie RSS parser	/home/tools/magpieRSS
Path to youtube_dl	/home/tools/youtube_dl (only required if you want to download flash videos)

Figure 2: *TubeKit*: setting basic preferences

2. Set up the database (Figure 3). *TubeKit* uses MySQL database for storing all the collected data. Given enough information, *TubeKit* can create a new database and required tables for the crawler being created.

Database Setup	
Host name	localhost
Name of the database	elections2008 (should not already exist)
User name	elections
Password	*****

Figure 3: *TubeKit*: giving the database details

3. Select different attributes to collect for a YouTube video (Figure 4). There are 17 such possible attributes that *TubeKit* can collect from YouTube for a given video. One can not only select which attributes to crawl, but also if it should be crawled only the first time, or every time the crawler process is run. This is useful while monitoring the videos over a period of time as several of its attributes, such as the user who posted the video, will not change over time, and there is no need to record it every time. *TubeKit* comes with a default setting that has all the attributes marked appropriately for the common use.
4. Set up various schedules for crawling (Figure 5). *TubeKit* provides full flexibility for scheduling various processes. One can specify how often or when exactly the four processes of the crawler should run. Once again, *TubeKit* comes with default values for these parameters that schedules to run the processes during the night.

**YouTube Setup**

Indicate how often each of these items should be crawled    Never    Once    Every time

	Never	Once	Every time
Title	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Description	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Username	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Time when video added	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Duration in seconds	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Category	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Keywords	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Video-page URL	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Thumbnail URL	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Number of views	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Number of ratings	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Average rating	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Number of comments	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Text comments	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Number of video responses	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Number of times favorited	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Rank in the rank-list for a query	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>

Figure 4: *TubeKit*: selecting the attributes to crawl (default setting shown)

**Crawling Setup**

Event	Month (1-12)	Day of month (1-31)	Day of week (0-6)	Hour (0-23)	Min (0-59)
Execute queries	<input type="text" value="*"/>	<input type="text" value="*"/>	<input type="text" value="*"/>	<input type="text" value="0"/>	<input type="text" value="0"/>
Crawl	<input type="text" value="*"/>	<input type="text" value="*"/>	<input type="text" value="*"/>	<input type="text" value="1"/>	<input type="text" value="0"/>
Download videos in Flash format	<input type="text" value="*"/>	<input type="text" value="*"/>	<input type="text" value="*"/>	<input type="text" value="2"/>	<input type="text" value="0"/>
Convert videos to MPEG	<input type="text" value="*"/>	<input type="text" value="*"/>	<input type="text" value="*"/>	<input type="text" value="4"/>	<input type="text" value="0"/>

Figure 5: *TubeKit*: scheduling various events (default setting shown)

**Add a seed query**

Queries being monitored:

1. elections 2008
2. decision 2008
3. Barack Obama
4. John McCain

Figure 6: Monitoring and adding queries to your customized crawler

5. *Access your crawler and enter seed queries (Figure 6)*. Finally, one needs to enter a set of queries that the crawler created with *TubeKit* can continue running as per the schedule chosen.

Once the queries are entered, the crawler is ready and should start harvesting the videos along with a variety of attributes based on the configuration of the crawler and other settings. *TubeKit* also generates a front-end of the crawler that can be accessed using a browser. This interface allows one to monitor the collection being built by the active crawler. The following section describes this interface with a crawler that we created for collecting election-related videos from YouTube.

## 4 *TubeKit* and the 2008 Elections

This section presents the details of our election crawler built using *TubeKit*, along with some analysis of the collected data. As mentioned before, we were interested in documenting presidential elections of 2008 from the perspective of an archivist concerned with preserving online digital media. Given its popularity, usage, and market penetration, YouTube was our natural choice for this. In addition, most of the proposed or possible candidates have their own channels on YouTube. CNN had also paired up with YouTube for hosting candidates' debates and getting public responses to those videos (YouTube, 2008).

We built a crawler using *TubeKit* as described in the previous section and entered 56 queries. Of these queries, 6 were general queries such as 'election 2008', and the rest were the names of possible candidates at that time (March 2007) obtained from Wikipedia (Wikipedia, 2007). For each of these queries, we decided to collect the top 100 results from YouTube every day. This means every day our crawler would send 56 queries to YouTube, get the top 100 results, and store the results that we do not already have. Thus, we get only new videos every day. However, we do collect the contextual information for *all* the videos that we have every time we run our crawler. As noted before, such contextual information includes time-sensitive attributes such as number of views, comments, and ratings.

Figure 7 shows some of the queries being monitored by our crawler along with the number of YouTube videos it has collected for each of these queries.

Figure 8 displays a query-wise summary of additional attributes for the collected videos. In

#	Query	Setup	Total results so far
1	election 2008	Setup	1636
2	US election 2008	Setup	1044
3	United States election 2008	Setup	785
4	presidential election 2008	Setup	1018
5	campaign 2008	Setup	1101
6	decision 2008	Setup	722
7	Joe Biden	Setup	565
8	Hillary Rodham Clinton	Setup	544
9	Christopher Dodd	Setup	510
10	John Edwards	Setup	1246

Figure 7: Partial list of queries for the election crawler

this display, we can see that as of December 27, 2008, we had finished more than 500 crawls<sup>8</sup> and collected nearly 25000 unique videos. We can also see query-related statistics. For instance, as of that day, we had collected 534 videos related to Hillary Rodham Clinton, with average views of 27708, and average comments of 185 per video. The crawler updates these statistics after every crawl and prepares a front-end to present it.

Latest crawl on: 12/27/2008, Total crawls: 528, Total videos: 27097, Unique videos: 24347, Downloaded videos: 23338, Converted videos: 18838						
#	Query	Videos	Views	Ratings	Comments	Favorited
1	election 2008	1550	26405.1378	2.1612609948231	140.2874	64.1921
2	US election 2008	964	12381.0872	2.3948825543759	99.1309	45.6393
3	United States election 2008	708	17985.0991	2.3004185047969	73.3018	30.8370
4	presidential election 2008	942	6749.3778	2.1905333354738	41.2533	21.3333
5	campaign 2008	1093	8672.1379	2.1311877417838	33.0096	21.4502
6	decision 2008	705	7505.0127	2.4034394891399	36.2739	12.7580
7	Joe Biden	562	8589.4360	2.7083430165468	35.9448	17.8169
8	Hillary Rodham Clinton	534	27708.9363	1.8098726208043	185.9554	43.1338
9	Christopher Dodd	506	3933.4276	2.4079385993251	19.0263	3.6075
10	John Edwards	1237	10385.8673	2.8525585399332	59.0650	22.7615

Figure 8: Portion of collection summary (as of 12/27/2008). Columns with ‘Views’, ‘Ratings’, ‘Comments’, and ‘Favorited’ show the averages for 12/27/2008, and *not* for the entire collection up to that day.

An overview of our collection over 18 months is depicted in Figure 9. One might question - given that we are automatically collecting the videos based on running a set of queries, what is the guarantee that we are actually getting the videos on the given topic? There are several ways to validate the collected data. For instance, we can look at the genre of the videos and find out what

<sup>8</sup>We intended to run one crawl per day, but there were days we had to skip the crawling due to system maintenance.

the collection is mostly about. However, given that our collection is focused on election 2008, genre for the most videos is likely to be the same. What may be more interesting is looking at what these individual videos are about. One way of finding this *aboutness* is by analyzing the tags associated with these videos. Tags on YouTube are usually some keywords that are assigned by the author of the video while posting that video. For instance, for video titled ‘John Edwards Feeling Pretty’,<sup>9</sup> the tags are “John Edwards Hair Style”. This tells us that the video is about John Edwards and also has something to do with hair styling.<sup>10</sup>

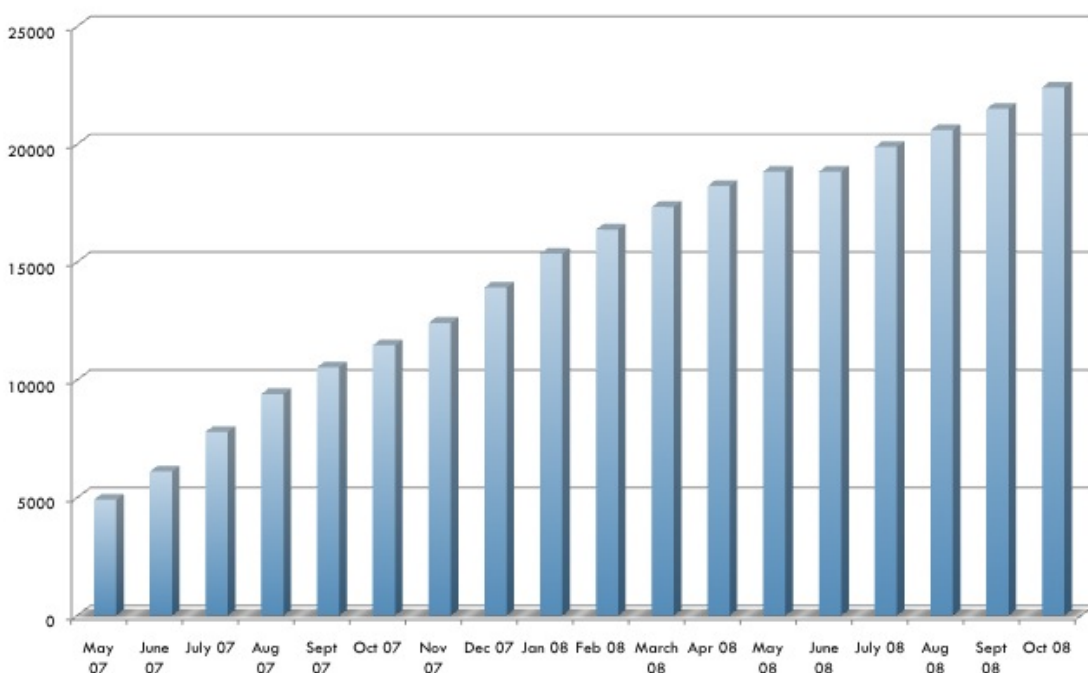


Figure 9: Overview of our election collection

A popular way of visualizing the tags is using a tag cloud (Halvey & Keane, 2007; Kaser & Lemire, 2007), which is extensively used in several of the Web 2.0 websites (Bielenberg & Zacher, 2006). We generate tag cloud after each crawl from all the unique videos collected so far. The size of a tag term on a tag cloud is proportionate to its frequency in the collection. A snapshot of our tag cloud on January 5, 2009 is given in Figure 10. In order to make it feasible for usable

<sup>9</sup><http://youtube.com/watch?v=2AE847UXu3Q>

<sup>10</sup>Note that at present, YouTube considers multi-term concepts as individual keywords; thus, a two term concept such as *hair style* is considered two separate terms by the retrieval system.

display, we ignored the tags that occurred fewer than 50 times in the collection. Thus, we can retain important tags such as “Edwards” and remove less significant tags such as “hair” for this collection.

Over time as new videos keep appearing in our collection, this tag cloud keeps changing and in a way, reflects what is gaining or losing popularity in terms of content production and posting. This not only helps us in visualizing the trends in our collection, but also provides a verification that indeed, the most of the videos in our collection are about the topics that we would expect.



Figure 10: Snapshot of the tag cloud on January 5, 2009

*TubeKit* also prepares a front-end for browsing or searching in the collected videos. A snapshot of this interface is shown in Figure 11. Some basic information such as title, description, and genre of each of the videos is displayed here. The last column in the given table has letters ‘M’ and ‘C’, which link to the metadata and contextual information respectively.

As noted before, we treat any static information about a video as metadata, and any dynamic or time-dependent information as the contextual information. Snapshots for such metadata and

Search

Total results found: **1776** with **barack obama** in titles      [First Page] [Prev] Showing page **1** of **89** pages [Next] [Last Page]

Query	Title	Category	Crawl date	Info
1	<a href="#">Barack Obama 2008 - Las Vegas Nevada - Feb. 18, 2007</a> Part of Barack Obama's speech at his first major campaign stop since he announced his candidacy for president. For more <a href="http://www.randinternational.net/obama">www.randinternational.net/obama</a> .	News & Politics	2007-05-11	<a href="#">M</a> <a href="#">C</a>
1	<a href="#">BARACK OBAMA @ DNC: Calls for End to War and for Hope</a> Likely presidential candidate Barack Obama spoke at the DNC's winter meeting on February 2, 2007.	News & Politics	2007-06-26	<a href="#">M</a> <a href="#">C</a>
1	<a href="#">Another message for Senator Barack Obama Bin Laden</a> This is my response to remarks by Barack Obama Bin Laden. WARNING -- this clip maybe harmful to political correctness and a little humor is required.	News & Politics	2007-07-01	<a href="#">M</a> <a href="#">C</a>
1	<a href="#">Barack Obama's Walk for Change: South Carolina</a> On June 9, hundreds of South Carolinians walked door to door in their neighborhoods to spread the word about Barack Obama. Here's a video of a canvass in North Charleston.	News & Politics	2007-07-01	<a href="#">M</a> <a href="#">C</a>
1	<a href="#">Barack Obama @ Jim Webb Rally in Richmond</a> *View Obstructed by Someone's Head and it's shakey because I had to hold the camera up in the air for 4 minutes and I am not that strong, but the words are what's best* Yeah man, I went to the Political Rally and it had basically every Democrat ever elected in VA and Barack Obama. It was a great rally. And I got within three feet of him and I LOVED IT!!!	News & Politics	2007-07-01	<a href="#">M</a> <a href="#">C</a>
1	<a href="#">Barack Obama Announcement For President</a> Barack Obama's announcement on forming a presidential exploratory committee. What it basically means is that he can raise funds for a possible presidential bid, pretty much a "test the waters" thing.	News & Politics	2007-07-01	<a href="#">M</a> <a href="#">C</a>
1	<a href="#">Ted Sorensen on Barack Obama</a> Legendary speechwriter and adviser to President John F. Kennedy on why he supports Barack for President.	News & Politics	2007-07-01	<a href="#">M</a> <a href="#">C</a>
1	<a href="#">Barack Obama Will Run for President</a> Watch Barack's statement on forming a Presidential Exploratory Committee	News & Politics	2007-07-01	<a href="#">M</a> <a href="#">C</a>

Figure 11: Browsing/searching in the collection

contextual information for a video are shown in Figure 12 and 13. The metadata information is self-explanatory. Let us look at Figure 13 for the contextual information. Here, our election crawler has presented crawl-wise statistics of a variety of dynamic parameters such as number of views, ratings, and comments. In addition to this, it indicates the significance of changes for a given parameter between two crawls. This is done by using different shades of yellow for highlighting the values. The scale on the top of the table presents the relation between the highlighting color and the % change in the value of a given parameter from the previous crawl.

In addition to reporting the differences between two crawls, *TubeKit* also provides a way for the user to indicate what constitutes a significant change for him/her. Figure 14 shows the interface for setting such preferences. As we can see, the user can combine different dynamic parameters using AND or OR operators and set their individual values that help decide if the present crawl is reporting a significant change from the previous crawl or not. Once such parameters are set for a query<sup>11</sup>, the crawler provides a binary decision regarding whether a given crawl is significantly

<sup>11</sup>See Figure 7 where the queries are displayed. With each query, there is a link to 'Setup', which brings up the interface shown in Figure 14.

**Query:** Barack Obama  
**Title:** Barack Obama: My Plans for 2008  
**Description:** Visit BarackObama.com for more information.  
**Username:** BarackObamadotcom  
**Posted date:** 2007-04-02  
**Duration:** 187 seconds  
**Categories:** News & Politics  
**Tags:** barack obama politics  
**Recording location:** lincoln park, west 11th and kenilworth, in tremont, cleveland, ohio,  
**Collection date:** 2007-05-03



**Surrogate:**

Figure 12: Metadata for the video ‘Barack Obama: My Plans for 2008’

Color coding for % changes

<	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	5.0	>
---	------	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	---

Crawl #	Crawl date	Rank	Views	Ratings	Avg Rating	Comments	Links	Favorited	Honors	Change
1	2007-05-03	1	211579	2267	4.59	3038	5	738	2	NO
2	2007-05-04	1	212582	2274	4.58	3009	5	741	2	NO
3	2007-05-05	1	214218	2279	4.58	3041	5	743	2	NO
4	2007-05-06	1	215910	2288	4.58	3100	5	747	2	NO
5	2007-05-07	2	216988	2295	4.58	3141	5	747	2	YES
6	2007-05-08	1	218189	2303	4.58	3156	5	749	2	YES
7	2007-05-09	1	219350	2309	4.58	3187	5	753	2	NO
8	2007-05-10	1	220357	2314	4.58	3211	5	754	2	NO
9	2007-05-11	1	221381	2321	4.58	3227	5	760	2	NO
10	2007-05-12	1	222328	2325	4.58	3248	5	760	2	NO
11	2007-05-13	1	223148	2331	4.58	3269	5	761	2	NO
12	2007-05-14	2	224382	2345	4.57	3300	5	762	2	YES
13	2007-05-15	2	226511	2366	4.56	3327	5	764	2	NO
14	2007-05-16	1	227767	2373	4.56	3343	5	766	2	YES
15	2007-05-17	1	228835	2384	4.56	3348	5	767	2	NO
16	2007-05-18	1	229611	2394	4.56	3348	5	768	2	NO

Figure 13: Crawl-wise contextual information for the video ‘Barack Obama: My Plans for 2008’

different from the previous crawl or not. This is indicated in the last column of crawl listing (Figure 13).<sup>12</sup>

Operator	Attribute	% change	
	Rank		<input type="text" value="2"/>
<input type="button" value="OR"/>	View counts		<input type="text" value="1"/>
<input type="button" value="AND"/>	Number of ratings		<input type="text" value="4"/>
<input type="button" value="AND"/>	Average rating		<input type="text" value="7"/>
<input type="button" value="OR"/>	Number of comments		<input type="text" value="3"/>
<input type="button" value="OR"/>	Number of responses		<input type="text" value="2"/>
<input type="button" value="AND"/>	Number of links		<input type="text" value="6"/>
<input type="button" value="OR"/>	Number of favorited		<input type="text" value="4"/>
<input type="button" value="AND"/>	Number of honors		<input type="text" value="5"/>

Figure 14: Setting monitoring options for a query

We found such functionalities provided by *TubeKit* extremely useful in our analysis. For instance, we found that on August 26, 2007, there was a significant change reported on the crawl for the video ‘Barack Obama: My Plans for 2008’. On that Sunday, Barack Obama visited New Orleans and gave a speech presenting a plan aimed at hastening the rebuilding of New Orleans and restructuring how the federal government would respond to future catastrophes in America. He also took a walking tour of a city neighborhood. This event created many discussions in the news media (Zeleny, 2007) as well as the blogosphere (The Richmond Democrat, 2007). Reflecting this significant Obama event his flag-video on YouTube reflected much more than usual participation. Such incident also indicates high correlation between real-life events and participation around the related YouTube videos. The generality of such correlation, however, needs to be investigated further.

Detecting such events can help us in spawning off other processes. For instance, one can think of having an automated system that can go out and explore various information outlets such as the New York Times and CNN.com when a change of certain magnitude for a query (candidate) or a video occurs.

Let us now look at the videos specifically contributed by Barack Obama and John McCain campaigns. To identify these videos, we looked at the videos posted by ‘BarackObamadotcom’ and

<sup>12</sup>Note that the display in Figure 13 was generated with different values than what is shown in Figure 14.

‘JohnMcCaindotcom’ users respectively.<sup>13</sup> Using this approach, we found that as of October 20, 2008, Barack Obama’s campaign had posted 577 videos, which was the highest number of videos posted (2.6% of 22,104) by any individual or organization in our collection. On the other hand, John McCain’s campaign had posted only 94 videos (0.4%), ranking 21 among the authors in our collection. It is not surprising that Obama’s videos had a view count of more than 34 million, whereas McCain’s videos had a view count of less than 2 million. This gives Obama’s videos nearly 18 times more views than that of McCain’s. Other statistics about the YouTube videos of these two candidates can be seen in the table below.

Table 1: Obama and McCain on YouTube (based on our collection as of 10/20/2008)

<b>Parameter</b>	<b>Obama</b>	<b>McCain</b>
Videos posted	577	94
Number of views	34,387,028	1,919,855
Number of comments	69,188	23,711
Number of ratings	219,876	15,622
Number of times favorited	13,517	3,791

Since Obama had significantly more videos posted than McCain, it may be unfair to compare their views etc. directly. We, therefore, present the averages for both the candidates in the Table 2. As we seen in that table, on average, an Obama video was viewed nearly three times as much as McCain’s. Sure, McCain’s videos have more comments per video than Obama’s, but without analyzing them, it is hard to say anything about the opinionated nature of those comments.

Table 2: Averages for Obama and McCain on YouTube (based on our collection as of 10/20/2008)

<b>Parameter</b>	<b>Obama</b>	<b>McCain</b>
Avg. views per video	59,596	20,424
Avg. comments per video	120	252
Avg. ratings per video	381	166
Avg. number of times a video favorited	23	40

---

<sup>13</sup>Note that we do not claim that we have *all* the videos related to these two candidates.

## 5 Additional Tools and Analysis

In addition to the primary component of *TubeKit*, which incorporates a suit of tools to perform query-based YouTube crawling, we have developed a few small tools<sup>14</sup> that lets one collect various forms of information off YouTube without running queries. These tools are listed below.

- *Extract YouTube video URLs*

A script that takes a set of YouTube URLs (or URLs to almost any webpage), and extracts the embedded URLs that point to YouTube videos. One can use this generated list to harvest various attributes about those videos using the ‘Harvest videos’ or ‘Download YouTube videos’ tools described below.

- *Download YouTube videos*

This script lets one download the YouTube videos using ‘youtube\_dl’ tool. One can write these URLs manually, or use the output of the ‘Extract YouTube videos URLs’ tool.

- *Harvest videos*

This script lets one collect a number of attributes of a YouTube video. All one need to do is put the URLs of those videos in a text file and pass the name of that file as an argument on the command line. One can write these URLs manually, or use the output of the ‘Extract YouTube videos URLs’ tool.

- *Harvest profiles*

This script reads username handles from a table in which the data is collected by ‘Harvest videos’ tool, and collects a set of attributes from that user’s profile.

- *Crawl inlinks*

This script takes a URL of a YouTube video (or any URL) and finds the webpages that link to it or embed it. This tool uses the web crawl by Yahoo! to find such inlinks.

‘Harvest videos’, ‘Harvest profiles’, and ‘Crawl inlinks’ tools store the harvested data in a MySQL database, which can then be easily viewed or extracted.

---

<sup>14</sup>Available from <http://www.tubekit.org/tools.php>

In addition to the election crawler, we have created additional crawlers using *TubeKit* for collecting data from YouTube for our research on various topics. These crawlers are on topics such as energy, epidemics, health, natural disasters, and truth commissions. These crawlers have also been running for almost as long as our election crawler and have harvested about 30000 videos over a period of nearly two years.

The framework of *TubeKit* has been used to create *ContextMiner*, which allows one to run queries on different sources, including YouTube, without any installation on their side and with even less effort. The description of *ContextMiner* is beyond the scope of this article, but the reader is referred to <http://www.contextminer.org> for further details and exploration.

While YouTube provides many valuable attributes relating to a video, we may need to explore other sources such as blogs to complete the picture (Capra et al., 2008). For instance, look at one of the most popular (viral) videos on YouTube: ‘Vote Different’.<sup>15</sup> To many people it is not clear where it came from - what the story is behind, who created it, and why. A screenshot of this item collected from YouTube by our system is shown in Figure 15. Some of the basic information about this video, including description, author name, and keywords, can be seen.

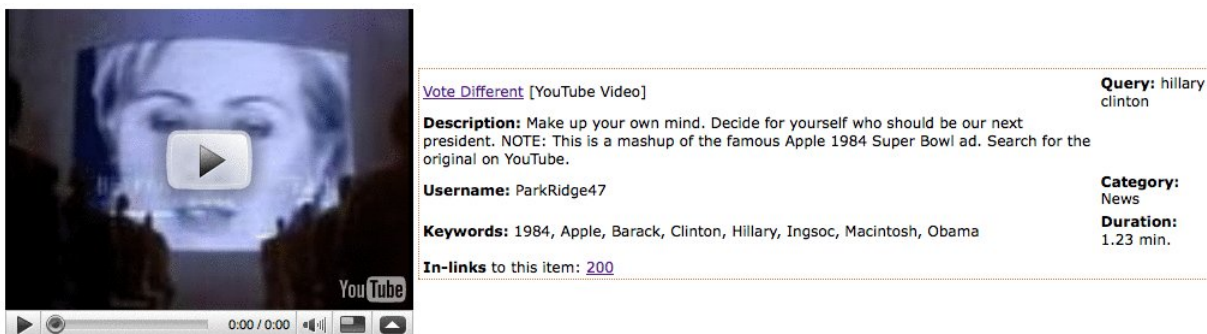


Figure 15: Metadata for the video ‘Vote Different’

Now if we look at the in-links collected to this YouTube video (Figure 16), we see that one of the articles linking to the above video talks about the author of this video. As we look at this article (Figure 17), we can see that it talks about who created this video, why, and what is the background for the video. We can also see the original ‘Think Different’ video embedded in the article. Together, these objects provide us good enough contextual information to document the given digital object in a more meaningful way.

<sup>15</sup><http://www.youtube.com/watch?v=6h3G-lMZxjo>

#	Page
1	<a href="#">L E X R E X: Rescuing the REPUBLIC - One Heart, and One Mind at a time</a>
2	<a href="#">March 2007 - Posts - First Read - msnbc.com</a>
3	<a href="#">Political video smackdown / 'Hillary 1984': Unauthorized Internet ad ...</a>
4	<a href="#">Assistant/Atlas: Hollywood's Young Shoulders</a>
5	<a href="#">The (liberal)Girl Next Door</a>
6	<a href="#">Yahoo's Presidential 'Mashup Debate' Won't Support Mashups</a>
7	<a href="#">All animals are equal, but some animals are more equal than others...</a>
8	<a href="#">Watching Big Sister - washingtonpost.com</a>
9	<a href="#">techPresident - Who is "ParkRidge47"?</a>
10	<a href="#">Who is the person behind the Clinton attack ad?</a>

Figure 16: Inlinks to the video 'Vote Different'

Article about the author of "Vote Different"

**Who is "ParkRidge47"?**

By Micah L. Sifry, 03/07/2007 - 3:06pm

The first piece of voter-generated video to make a splash in Campaign 2008 has hit, and with it comes a mystery: Is "Vote Different" really the work of an amateur, a civilian if you will? Or is it a shrewd move by someone who wants to stir up trouble between the Hillary Clinton and Barack Obama campaigns? After all, comparing Hillary to Big Brother, droning on about her "conversation" with America and portraying her supporters as silent automatons is hardly what Obama supporters want to say about the former First Lady. Or is it?

Yesterday, I was talking with several of TechPresident's contributing bloggers, and we all agreed that "Vote Different" was a terrific piece of viral video. Indeed, the minute-long mash-up of Apple's famous "1984" ad and some Hillary speeches has been viewed more than 100,000 times since being posted two days ago by someone using the handle "ParkRidge47."

One tip-off that this is not the work of a rank amateur is how well the video integrates not just Hillary's speeches into the rectangular TV boxes shown in the spot, but the placement of Obama's circular campaign logo on the woman runner's shirt (see below). Of course, it's totally possible that someone unconnected with any political operation could have these skills.

print

email

delicious

digg

technorati


Related links:

On "Hillary 1984": Phil de Vellis Speaks to YouTube and PoliticsTV

Breaking News: The Barocket Takes Off on YouTube

Mashup Madness, and the Rise of Videracy

**Vote Different**



Original "Think Different" video

Figure 17: Article about the author 'ParkRidge47'

## 6 Conclusion

In this paper we presented *TubeKit*, a toolkit that helps one create a customized focused crawler for YouTube. We demonstrated how we created a crawler for harvesting not only the videos relating to the election, but also several attributes over a period of many months. In this process that has lasted for nearly two years for us, we have learned several lessons, some of which are listed here.

- Looking at a YouTube video at a given time may not tell us the whole story behind it. It is important to observe it over a period of time to learn about its usage and impact.
- User participation is one of the defining factors of platforms such as YouTube. While studying the original information or objects (in this case, digital videos), we have to look at the user participation and the community built around it.
- Not all users are equal in their contribution on online mass media platforms such as YouTube. Shah & Marchionini (2008) showed how a small number of participants can make a huge impact on the overall information landscape due to their unique roles.
- There seem to be a high correlation between online participation on YouTube and real-life events.<sup>16</sup> For instance, we found that within a few days of the announcement of Sarah Palin as the republican party candidate for the vice-president, the number of YouTube videos relating to her went up by a significant number. We also saw a spike in the participation around those videos as measured by the views and comments.
- The core part of *TubeKit* is based on collecting data from YouTube by running queries. Unfortunately, we do not fully understand how YouTube’s relevance algorithm works, and thus, we are not sure what factors are considered while executing a query on YouTube. The functionality of *TubeKit* is also limited by the API and other support provided by YouTube’s website.

We are committed to continually develop and support open-source *TubeKit* for the research community, distributed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0

---

<sup>16</sup>This may not be true for other topics, and the truthfulness of this statement for general cases needs to be invested further.

United States License. Now in its third major release of public beta, *TubeKit* has helped many research groups and organizations in not only collecting valuable data from YouTube, but also in making sense of it. In about a year's time, *TubeKit* has been downloaded for nearly 300 times and used by researchers all around the world - from Library of Congress to University of Paris - for a variety of projects. We believe this toolkit and the associated tools will keep helping us accelerate our research related to YouTube.

## 7 Acknowledgment

This work was not possible without constant guidance and support of the other members of The VidArch Project team - Gary Marchionini, Rob Capra, Paul Jones, Sarah Jordan, Cal Lee, Terrell Russell, Laura Sheble, Yaxiao Song, and Helen Tibbo. The work reported here is supported by NSF grant # IIS 0455970.

## References

- Bielenberg, K., & Zacher, M. (2006). *Groups in social software: Utilizing tagging to integrate individual contexts for social navigation*. Unpublished master's thesis, Unisersitat Bremen.
- Capra, R., Lee, C. A., Marchionini, G., Russell, T., Shah, C., & Stutzman, F. (2008). Selection and Context Scoping for Digital Video Collections: An Investigation of YouTube and Blogs. In *IEEE ACM Joint Conference on Digital Libraries (JCDL)*.
- Chakrabarti, S., Berg, M. van den, & Dom, B. (1999). Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks*, 31(11-16), 1623-1640.
- Dalton, A. (2007). *The Digital Road to the White House*. Available from <http://www.pcmag.com/article2/0,1895,2109480,00.asp> [Accessed: 08/05/2007]
- Gomes, L. (2006, August 30). *Will All of Us Get Our 15 Minutes On a YouTube Video?* In *Wall Street Journal*. August 30, 2006. Available from [http://online.wsj.com/public/article/SB115689298168048904-5wWyrSwyn6RfVfz9NwLk774VUWc\\_20070829.html?mod=rss\\_free](http://online.wsj.com/public/article/SB115689298168048904-5wWyrSwyn6RfVfz9NwLk774VUWc_20070829.html?mod=rss_free) [Accessed: 08/30/2006]
- Halvey, M., & Keane, M. T. (2007, May 8-12). An assessment of tag presentation techniques. In *Proceedings of WWW Conference*. Banff, Alberta, Canada.

- Jarvis, J. (2007, February 5). *Why YouTube gets my vote for political punditry*. Available from <http://www.guardian.co.uk/media/2007/feb/05/mondaymediasection.politicsandthemedi> [Accessed 08/05/2007]
- Kaser, O., & Lemire, D. (2007, May 8-12). Tag-cloud drawing: Algorithms for cloud visualization. In *Proceedings of Tagging and Metadata for Social Information Organization (WWW 2007)*. Banff, Alberta, Canada.
- Seelye, K. Q. (2007, July 23). *Debates to connect candidates and voters online*. Available from [http://www.nytimes.com/2007/07/23/us/politics/23youtube.html?\\_r=2&ex=1186977600&en=d005e5d14080121d&ei=5070](http://www.nytimes.com/2007/07/23/us/politics/23youtube.html?_r=2&ex=1186977600&en=d005e5d14080121d&ei=5070) [Accessed 08/05/2007]
- Shah, C., & Marchionini, G. (2007, June 29). Preserving 2008 US Presidential Election Videos. In *International Web Archiving Workshop (IWAWS)*. Vancouver, BC, Canada.
- The Richmond Democrat. (2007). *Obama's trip to New Orleans*. Available from <http://richmonddemocrat.blogspot.com/2007/09/obamas-trip-to-new-orleans.html> [Accessed: 01/16/2008]
- Wikipedia. (2007). *United States presidential election, 2008 - Wikipedia, the free encyclopedia*. Available from [http://en.wikipedia.org/wiki/United\\_States\\_presidential\\_election,\\_2008#Candidates\\_and\\_potential\\_candidates](http://en.wikipedia.org/wiki/United_States_presidential_election,_2008#Candidates_and_potential_candidates) [Accessed: 03/15/2007]
- YouTube. (2008). *The CNN-YouTube Debates*. Available from <http://youtube.com/debates> [Accessed: 01/16/2008]
- Zeleny, J. (2007). *Obama's Plan to Restore New Orleans*. Available from <http://www.nytimes.com/2007/08/26/us/politics/26obama.html> [Accessed: 01/16/2008]