Hunting for Hip, Hipsters, and Happenings on YouTube

Chirag Shah†
School of Information and Library Science, University of North Carolina at Chapel Hill, CB# 3360, 100 Manning Hall, Chapel Hill, NC 27599-3360. chirag@unc.edu

Gary Marchionini
School of Information and Library Science, University of North Carolina at Chapel Hill, CB# 3360, 100 Manning Hall, Chapel Hill, NC 27599-3360. march@ils.unc.edu

Abstract
The changing nature of information and evolving role of information sources have made it possible for almost anyone to be a consumer as well as a producer of information. Thus, many information services are focused on user participation and support different user roles. It has become essential for information scientists, social analysts, and digital library curators to recognize and study these social factors while analyzing the content of these information sources. In this paper we present these ideas in the light of our work with collecting and analyzing election videos from YouTube. Over the course of more than 8 months and 200 passes of data collection, we have gathered about 15000 videos along with nearly two dozen attributes for each video relating to US presidential elections 2008 from YouTube. Using this collection, we demonstrate how various social attributes such as tags, ratings, and comments can be used to detect significant trends, people, and events. This detection can help us gaining a better understanding of not only the content, but also the population that produces and consumes it.

Introduction
One of the goals for future libraries that the early thinkers in the field of information and library science such as Vannevar Bush and JCR Licklider suggested (Hauben, 2004) was developing a single human-machine-knowledge system that would make the body of knowledge more useful and accessible. While this vision that was conceptualized nearly 50 years back is still relevant, the landscape of information and library science has changed significantly. In the span of the last few decades we have seen a mass production of digital information, a worldwide networking of people, machines, and information, and constantly emerging new outlets of information and user participation.

As the type and sources of information are evolving rapidly, the role of an information scientist is likewise evolving. In recent years we have seen big burst of information on relatively new media services such as Flickr and YouTube and sources such as blogs and wikis. In this paper we argue that in order to realize the goal of making such a body of knowledge more useful, we need to build tools and techniques that are more in line with the nature of such information. In other words, simply representing the information in bits, indexing it with simple syntax, and retrieving it with unrealistic assumptions about human behavior, is not enough. We need to go beyond these traditional approaches to understand the nature of the information as well as the users producing it. This understanding is essential to an information scientist who is interested in developing or analyzing a collection of information. In order to bring these ideas to a

† Any inquiry regarding the work reported here should be directed to Chirag Shah (chirag@unc.edu)
realization, we present our work with developing a collection of digital videos relating to US presidential elections 2008, collected from YouTube. Using this collection and the corresponding process, we attempt to address the following research questions. Given a collection derived from a source where there is a high level of user participation,

1. How can we find what is popular?

2. How can we identify key people in that population?

3. How can we detect significant events?

4. How do we report the above findings to a digital library curator who is interested in creating and maintaining such a collection or in selecting portions of it as contextualizing information for their collections?

The rest of the paper is organized as the following. In the next section we present our methodology, describing how we collected a rich set of data from YouTube. Given this data, we then focus on understanding certain significant attributes of it. We achieve this by analyzing the data for significant trends, people, and events. This is demonstrated in Section 3, along with possible implications of such an analysis. Finally, some concluding remarks and pointers to future work are given in Section 4.

**Methodology**

We were interested in building a collection of videos related to the 2008 US presidential election. YouTube was a natural choice for us given the amount of information (Gomes, 2006), the level of user participation, and its coalitions with other channels such as CNN (YouTube, 2008). We decided to build a tool to harvest election related videos from YouTube and analyze them. Since the election is one of the most popular current topics on YouTube, the content as well as user participation keep changing constantly. Realizing this, we decided to collect this data frequently, if possible, everyday. The design of this tool, the process of data collection, and the description of the data are provided in this section.

**Query-Based Harvester for YouTube**

Realizing the importance of people’s thinking and expressions on a politically and socially significant topic such as the presidential elections in the US, we decided to harvest YouTube (YouTube, 2007a) videos on this topic. This process that started in May 2007, and had been executed almost everyday since then, is described below and depicted in Figure 1.
1. Send queries to YouTube. We selected the names of possible candidates (Wikipedia, 2007) as well as 6 general, election-related phrases as our queries. This gave us total 57 queries.

2. Using YouTube APIs (YouTube, 2007b) and other scrapping process, download and parse result pages. We decided to collect the top 100 results for each query. For a given query, we do not collect any duplicate videos.

3. From the parsed data, certain information that are static in nature and provided by the author of the video, such as title, author, and tags, are stored as the metadata. This information is collected only once for each video page. There are other attributes, contributed by the visitors, such as comments, ratings, and number of views, which keep changing. We collect this information every time we run our harvester and store as contextual information.

4. Download the actual video in Flash format. Later we can convert this to MPEG format, which is more common and supported.

More information of this work can be found in (Shah & Marchionini, 2007), (Marchionini, Tibbo, Shah, & Lee, 2007) and the project website available through
www.ils.unc.edu/vidarch. A summary of the collected data can be accessed through ContextMiner website.¹ This data has also been made available for access via web-based APIs (Shah, 2007) under a Creative Commons License.

**Data Collection**

For over more than 8 months we have automatically crawled YouTube for election videos more than 200 times,² collecting about 15000 unique videos along with their associated metadata and contextual information. Each of these videos contains about two-dozen attributes such as title, description, views, and comments, and since we have been collecting these attributes almost everyday, we have gathered data that is very rich in nature and can be used for various kinds of interesting analysis. For instance, at the time of writing this paper (early January 2008), our harvester was collecting more than a million comments everyday for these 15000 videos collectively. Analyzing these many comments can give us a good understanding about several of the ideas and opinions that people have regarding candidates and issues for election 2008.

**Data Analysis**

In the work described here, our goal is to capture the popularity of the content, evaluate user participation, and understand how such participation reflects real-life events. These three issues are addressed in the following subsections along with their possible implications.

**Representing the Hip - What is Popular**

There are several ways to evaluating the popularity of the content using our data. For instance, we can look at the genre of the videos and find out what 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”.³ 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.⁴

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 10, 2008 is given in Figure 2. In order to make it feasible for usable display, we ignored the tags that occurred less than 50 times in

---


² We tried to run the crawler (context collector component) everyday, but there were some days we could not due to maintenance and debugging.

³ [http://youtube.com/watch?v=2AE847UXu3Q](http://youtube.com/watch?v=2AE847UXu3Q)

⁴ 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.
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 verification that indeed the most of the videos in our collection are about the topics that we would expect.

Finding the Hipsters - People Who Make a Big Difference

As one of the early projects, Bureau of Applied Social Research studied 1940 presidential election and found that ideas often flow from radio and print to opinion leaders and from these to the less active sections of the population (Lazarsfeld, Berelson, & Gaudet, 1944). In other words, those opinion leaders played a vital role in transforming information and helping people join the conversation.

Figure 2: Snapshot of the tag cloud on January 10, 2008
Figure 3: Plotting connectors: x-axis represents number of videos on which a user posted at least one comment, and y-axis shows number of users.

'Connectors': a few people who posted at least one comment on many videos related to presidential election 2008.
Figure 4: Plotting mavens: x-axis represents number of comments on a single video and y-axis shows number of users. Note: the y-axis in this graph is on the log scale.
There is a reason why attractive models, sports-stars, and film celebrities are used for promoting a product. Some people have a large influence on the society than the rest for various reasons. In general, this group of people possess specific “powers” based on their positions. In social sciences this phenomenon is studied under the light of social capital of key people (Burt, 1999). For our work, we found Gladwell’s book The Tipping Point (Gladwell, 2002) very relevant, where he identifies three categories of significant people: connectors, mavens, and salesmen. These three kinds of highly influential people are described below.

- **Connectors**: They know a large number of people and possess special gifts for bringing the world together. Connectors are defined by having many acquaintances, a sign of social power.

- **Mavens**: They are those who accumulate knowledge and who have information on a lot of different products, prices, or places.

- **Salesmen**: These people have the skills to persuade others when they are unconvinced of what they are hearing. They may not have as much knowledge as a maven, but they have the skills to convince somebody.

As we can see, all of these three kinds of people have specific “powers” due to their position and background in the social network that they belong to. It may also be possible for a single person to be in more than one of the above-mentioned roles. Applying the ideas discussed in the previous section, we define our connectors, mavens, and salesmen in the
following manner. With these definitions, we also plot the corresponding data. For the results reported here, the data was extracted on July 2, 2007.

- **Connectors**: These people know a lot of people and have a wide range of knowledge about different topics. In our case, these may be the people who visited a large number of videos related to presidential election 2008 and left at least one comment there (Figure 3). These people make many postings (comments) on many videos.

- **Mavens**: These people specialize in a given (sub) topic and have a deeper knowledge or opinions about it. In our case, these may be the people who posted a significantly large number of comments on specific videos related to the presidential election 2008 (Figure 4). These people make many postings on a small number of videos.

- **Salesmen**: They reach out to as many places as possible and try to sell their ideas to a wide range of population. Mapping this definition to our case, a person who has posted many videos on the topic can be considered as a connector (Figure 5). These people post many videos to the site.

Identifying these key people can play an important role in understanding the nature of user participation and evaluating the content. Two such scenarios are presented below.

- It is very likely that several videos about a candidate are posted by their office. These authors will be identified as salesmen. This identification can be useful to recognize the “official” nature of some content and to classify the content in other ways.

- A person who has many things to say or has strong opinions about a candidate is likely to comment extensively on that candidate’s video(s). That person will be recognized as a maven. While analyzing the comments to understand people’s opinions, we can choose to filter out or normalize the opinions of a maven because those opinions, no matter how strong in nature, represent a single person and may bias our understanding of the information.

These three types of user behavior are automatically detectable and provide one set of data for information scientists, curators, and others to investigate more closely. Interpretation of the actual videos, comments, and other behaviors such as views will always require human processing, however, identifying the hot spots of behavior in huge volumes of evidence is a problem well suited to automation.

**Detecting the Happenings - Significant Events**
One interesting area for exploration in the YouTube crawls is detecting events. Since most of the videos on YouTube are open for public participation, we can hypothesize that this participation will be affected by real world events. In other words, we should be able to identify a significant event happening in the real world based on the users participation in online forums such as YouTube. These changes can be monitored for a video, a query, a topic (e.g., a specific candidate in the case here), or the entire collection. In order to detect such changes for a video, we propose to capture the videos crawl-progress by the following model.

\[ M = (\Theta, \Phi) \]  

(1)

Here, \( M \) is the model for a video, \( \Theta \) represents local changes for different parameters, and \( \Phi \) represents changes in participation. \( \Theta \) is defined below.

\[ \Theta = (\theta_1, \theta_2, \theta_3) \]  

(2)

\[ \theta_i = \tan^{-1} \left( \begin{pmatrix} y_2^i & y_1^i \\ x_2^i & x_1^i \end{pmatrix} \right) \]  

(3)
have seen some abrupt changes. This becomes apparent as we evaluate the parameters for this video in the days between July 31, 2007 and August 18, 2007 are depicted in Figures 6 and 7. As we can see from these figures, the number of views for this video has been steadily increasing, and so is the number of ratings. However, the numbers of comments, and θ₃ represents the change in number of ratings. Note that other parameters are possible as well (e.g., blog postings on the video; referrals in other media, etc.). For each of these, y₂ is the value of that parameter on crawl x₂ and y₁ is the value of that parameter on crawl x₁. The interval of interest can of course be small (two concurrent crawls) or large. Finding tan⁻¹ gives us an angle between Φ(θ₂, θ₃) that will indicate the amount of change for a given parameter between two crawls. Φ is defined below.

\[
Φ = \tan^{-1} \left( \frac{θ₂}{θ₃} \right)
\]  

(4)

\[
Φ₁ = \frac{\text{Number of comments}}{\text{Number of views}}
\]  

(5)

\[
Φ₂ = \frac{\text{Number of ratings}}{\text{Number of views}}
\]  

(6)

Let us apply this model to a specific video. In our collection, video #4612 is titled “Barack Obama: My Plans for 2008”. We have been crawling this video since May 3, 2007. Different parameters for this video in the days between July 31, 2007 and August 18, 2007 are depicted in Figures 6 and 7. As we can see from these figures, the number of views for this video has been steadily increasing, and so is the number of ratings. However, the numbers of comments have seen some abrupt changes. This becomes apparent as we evaluate Φ (Figure 8) that shows that the variability of views is flat (constantly increasing is flat when change is the unit
being plotted), that comments are varying wildly across different days, and that ratings tend to be somewhat stable (generally regular increases). This still does not give us complete picture of participation about this specific video. But when we compute $\Phi$ values that normalize comments and ratings by view activity (viewing is a necessary but not sufficient condition for rating and commenting) (Figure 9), we can clearly see that while participation by rating shows no interesting changes in the given period of time, participation by comments shows spikes. These are the points that may indicate some kind of seminal event relating to this video. For instance, we can see that around crawl number 30, there was a jump in participation by comments. The corresponding date for this crawl was August 26, 2007. 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 responds 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.

Figure 8: Theta values for video titled “Barack Obama: My Plans for 2008”
Figure 9: Phi values for video titled “Barack Obama: My Plans for 2008”

We can apply the same model presented above for a query or the whole collection. While all the results for a query or the collection may have a lot of diversity in terms of content and quality, we have observed (Shah, 2008) that they still represent most of the concepts the way we would expect. A curator who is observing such a graph may find it interesting to investigate other sources around this time to see what event(s) influenced such a sharp change. One can also 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 occurs.

While capturing significant events in such an automatic way could be indicative, what constitutes as a real significant change can vary from situation to situation and requires human interpretation. Therefore, we also designed a way for an analyst or curator to specify parameters to investigate. A screenshot of this interface is given in Figure 10. Using this interface, a curator can indicate the amount of change for a given parameter between two crawls that can be considered as a significant change. The interface allows one to combine a number of parameters in different ways to represent more complex definitions of an event.
Conclusion

In this paper we argued that in order to get a better understanding of an information source, we need to consider factors that are behind and around that information - the context. We presented the details of our project of data collection from YouTube relating to election 2008. Our goal in this project is twofold: (1) build collection development models, tools, and services, and (2) understand these new dynamic media and user participation with them. Both of these goals are interconnected. This was realized by evaluating our collection for trends, key contributors, and seminal events. In turn, this realization helped us refining some of our policies for collection development. For instance, as we realized that videos relating to certain candidates had a higher degree of popularity and user participation, we executed an additional pass on data collection for them, collecting the top 1000 videos for them, instead of 100. We believe similar models can be adapted to other collection development scenarios where the collection and its surrounding context is dynamic in nature.

We also demonstrated how we can evaluate various social attributes associated with the objects in the collection. This can serve as an important tool for a social scientist to understand a population and a curator to define collection policies. One of the limitations of the work reported here is that each of the analysis is done for a certain interval in time. We are working on extending this to include a temporal dimension. We would then be able to analyze the evolution of the factors reported here over the time. Finally, as we approach the elections and as this event actually takes place, it will be very interesting to analyze how closely various parameters in our data set reflect real trends and events that take place.

Acknowledgment

The authors wish to thank other members of VidArch team - Rob Capra, Paul Jones, Sarah Jordan, Cal Lee, Terrell Russell, Laura Sheble, Yaxiao Song, and Helen Tibbo - for their feedback on this work. The work reported here is supported by NSF grant # IIS 0455970.

References

Burt, R. S. (1999). The social capital of opinion leaders. The ANNALS of the American Academy of Political and Social Science, 566(37), 36-54.


[Accessed: March 1, 2007]

[Accessed: April 27, 2007]