The International Journal of Digital Curation
Issue 2, Volume 4 | 2009

ContextMiner: Supporting the Mining of Contextual Information for Ephemeral Digital Video Preservation

Chirag Shah

School of Information & Library Science (SILS)

University of North Carolina at Chapel Hill

Summary

Emerging information media present new challenges to the curators. While archiving objects, and building meaningful collection for long-term preservation and access, have been well-understood practice for centuaries, digital objects present new issues. In the previous article (Shah, 2009) I identified a number of these issues related to digital objects, specifically digital videos of an ephemeral nature. I argued that while preserving such objects, adding contextual information is essential. One of the interesting challenges is to identify what to collect and preserve as contextual information. For ephemeral digital videos, I proposed to harvest four kinds of relevances and five kinds of contexts. In order to implement this proposal, I presented ContextMiner, a framework and a system to support digital video curation. In this article, I will take a closer look at ContextMiner, analyzing it for its functionalities and usability. This is done by usability inspection and content analysis. For the former, we simulated two curatorial tasks, asked our users (curators) to use ContextMiner, and provide us feedback on its usability and functionalities. For the latter, we mined a collection prepared by ContextMiner for its potential usage in preservation. Finally, I have summarized the lessons learned from developing and using our system, providing implications for digital library curators interested in collecting and preserving digital objects of an ephemeral nature.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

The usage of digital videos is rapidly inclining. Due to the advent of the technology to produce, publish, and consume, such digital content is not only windespread, but also intertwined with many of the cultural and socio-political aspects of today. A curator interested in documenting and preserving a cultural or social phenomenon may want to include such video content, as well as the context surrounding it as well in the collection. For instance, one of our projects, funded by the Library of Congress, was to collect and preserve YouTube videos relating to the 2008 presidential elections in the United States. We used the ContextMiner system, described in a previous article (Shah, 2009), to acccomplish this task.

Since we began our work on the issue of mining contextual information for digital curation as a part of The VidArch Project,[1] we have made several efforts to evaluate our methods, interfaces, and collections. Here I will present some of our endeavors and findings in the direction of evaluating the usability of ContextMiner. In addition to the system itself, I will present various analyses of the data harvested using ContextMiner. This will allow us to comment more on the potential usage, lessons, and implications of ContextMiner framework for helping the curators in preserving digital objects of an ephemeral nature.

Usability Inspection of ContextMiner

To evaluate the usability of our curator's interface[2], we asked a few people to use the interface and gave them a couple of curatorial tasks. A total of 11 people participated in this usability inspection study. They were doctoral students of faculties in the School of Information & Library Science (SILS) at UNC Chapel Hill, each with at least some experience of preservation and curation. Users first viewed a video tutorial[3] explaining the functionality of the interface. They were then asked to perform the following two tasks using the curator's interface in ContextMiner:

1. Collect a small set of records for a topic related to the Hubble Space Telescope.
2. Collect a small set of records for a topic related to post-war German industrial reconstruction.

The users were allowed to work from anylocation and to their own convenience. Having completed the above two tasks, they filled in a feedback form. The first six questions on the form asked the subject to rate a variety of factors on a scale of 1 to 5. The summary for these questions is provided in Table 1. We can see that the interface did not receive a very high rating for ease of use, some of which feedback is revealed in the open-ended comments in the second part of the feedback form. The subjects seemed to be quite happy with the speed of searching and automatic extraction of metadata.

[1] VidArch at SILS http://www.ils.unc.edu/vidarch/

[2] See Figures 4-6 in my previous article (Shah, 2009).

[3] Thanks to Sarah Jordan for preparing this tutorial, which can be seen at http://idl.ils.unc.edu/~chirag/ContextMiner/CM_tutorial2.mov

#	Question	Avg. Rating
1	Ease of interface use (1 very hard, 5 very easy)	2.73
2	Clarity field terms (1 no clarity, 5 well clarified)	3.45
3	Speed of searches (1 very slow, 5 very fast)	4.45
4	Speed of automatic metadata extraction (1 very slow, 5 very fast)	4.54
5	Familiarity with the subject of topic-1 (space telescope) (1 no knowledge, 5 expert)	1.72
6	Familiarity with the subject of topic-2 (postwar German industrial reconstruction) (1 no knowledge, 5 expert)	2.00

Table 1. Summary of the results for questions 1 to 6 on the feedback form.

The remainder of the form gave participants an opportunity to express their opinions in free-form text. Question 7 asked the subjects to list the fields they felt were redundant. Three subjects reported they did not know what the 'Genre' field was because it extracted a single-digit number. This happened in the case of OpenVideo since internally OpenVideo represents the genre of a video as a number. This is a good example of the classical interoperation difficulties in mapping field names and codes across databases. In practice, curators will likely have to do some translation manually for data sources that are highly specialized. One user indicated that "Production date" and "When" fields, representing the time aspect, were not required. Question 8 asked the subjects to list any additional fields they felt should be included. One subject listed "Identifier" and "Subject" and wanted to know the format of the output of the record. Another subject suggested using more specific terms for the contextual fields, for example instead of "Who", using "Person" and/or "Group". One user suggested having a "Series" field to indicate if a video belonged in a series with other videos.

Question 9 asked the subjects what they liked the most about the ContextMiner interface. Two subjects responded that they liked the arrow buttons and how the fields were automatically populated with metadata, which of course is one of the primary functionalities this system offers. Three subjects listed they liked the speed and simplicity of the interface, as well as the ease of navigation. One subject noted that they could easily add and replace metadata. In general, our subjects liked the functionality of being able to search in different sources from the same page. Question 10 asked the subjects what they disliked most about the interface. A majority of the subjects were unhappy that the interface did not save search results and that there was no way to tell which objects had already been added. A couple of subjects disliked the fact that there was no way to view or edit the records they submitted. This feedback reinforces the need for the interface to be tightly coupled to the local content management scheme.

Overall the subjects felt the interface was easy to use and that some changes such as saving searches and having a means to track what had been added would make for a more effective interface. Furthermore, changing some of the manually added field names such as "Who" and "What" to nouns with more specific contextual categories such as "Person" or "Group" and "Event" and "Occurrence" would help ContextMiner be closer to ideals of the Encoded Archival Context initiative (Lee, 2007).

An interesting incident occurred during our study. YouTube suddenly changed its coding for formatting webpages. Since our search and extraction component for YouTube depends heavily on their website structure, it started having problems and we had to fix them immediately to carry on the study. This is a further example of the challenges digital library interoperation faces with respect to adapting to interfaces for digital libraries with evolving technologies, standards, and needs. Keeping this in mind, we do not claim that we have built a perfect interface for a digital curator. However, we have constructed a framework that is constantly evolving and which will continue to serve as an important platform to try and test a variety of ideas relating to mining contextual information for digital curation.

Content Analysis of the Harvested Data and Context

For more than a year we have been running our harvesting component of ContextMiner to automatically collect videos, metadata, and contextual information from YouTube on several topics. There were three major processes involved in this project as shown in Figure 1: (1) planning, (2) collection, and (3) analyses. The flows indicated in this figure also represent how each of the stages informs the other stages. In the previous article (Shah, 2009), I presented the details of the first two stages. Here I will talk about the third stage, which involves analyzing the collected data to inform the planning and the collection processes, as well as discovering some useful and interesting patterns. In addition to these analyses, I will also summarize a few things that we learned from our experience with the harvesting component of ContextMiner.

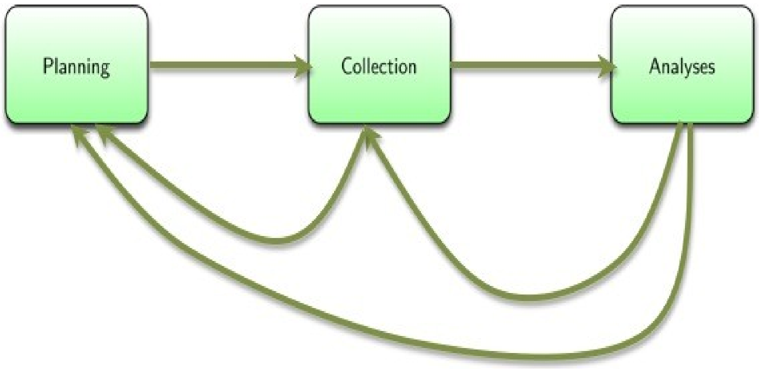


Figure 1. Three stages of our collection process.

	# Queries	# Total Videos
Election	57	20,637
Energy	48	8,737
Epidemics	5	682
Health	14	10,644
Natural Disasters	18	12,758
Truth Commissions	4	1,447

Table 2. Sizes of different collections (as of May 20, 2008).

In the past year, we performed about 300 crawls and collected thousands of videos with dozens of attributes in each crawl. The sizes of the collections generated by these crawlers are reported in Table 2.

The values of parameters such as views, comments, and ratings for the election collection, along with the size of the collection over the past 12 months is plotted in Figures 2-4.[4]

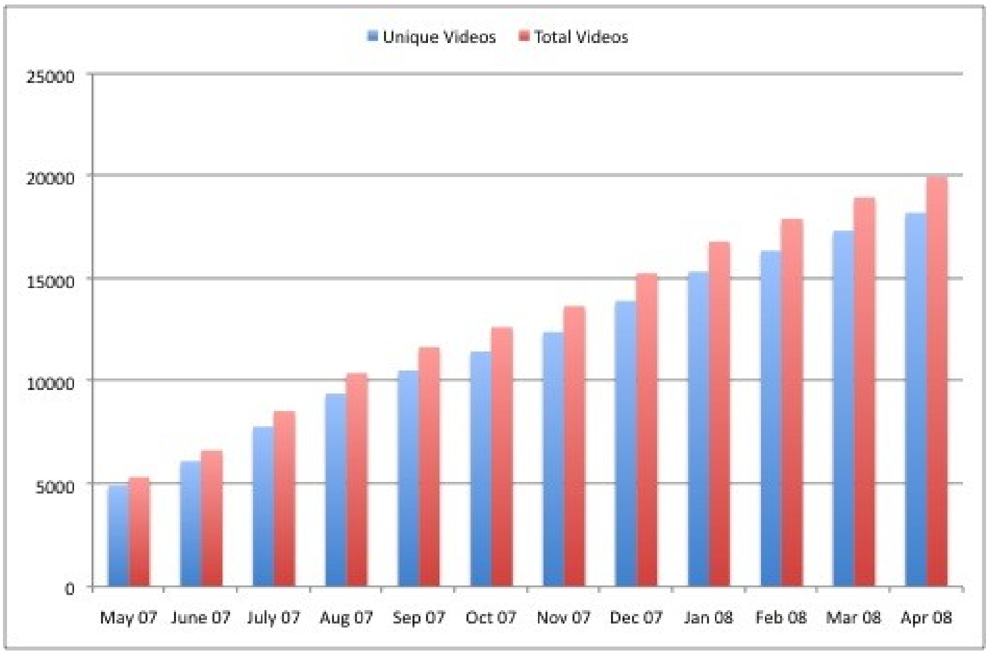


Figure 2. Total and unique number of videos over 12 months.

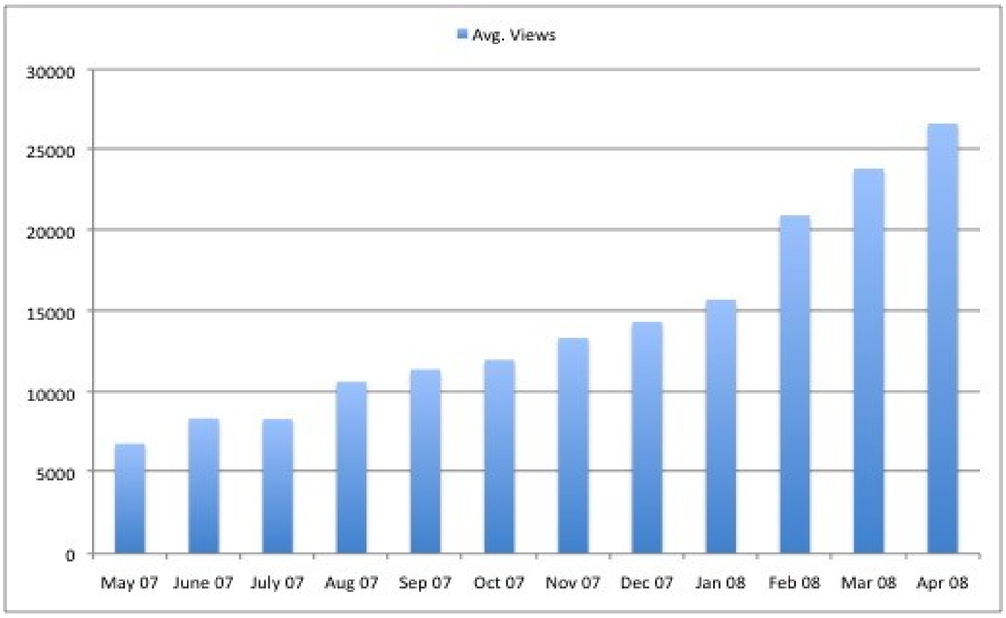


Figure 3. Average number of views for the collected videos over 12 months.

[4] Since some videos appear for more than one query, different statistics for total and unique videos are reported.

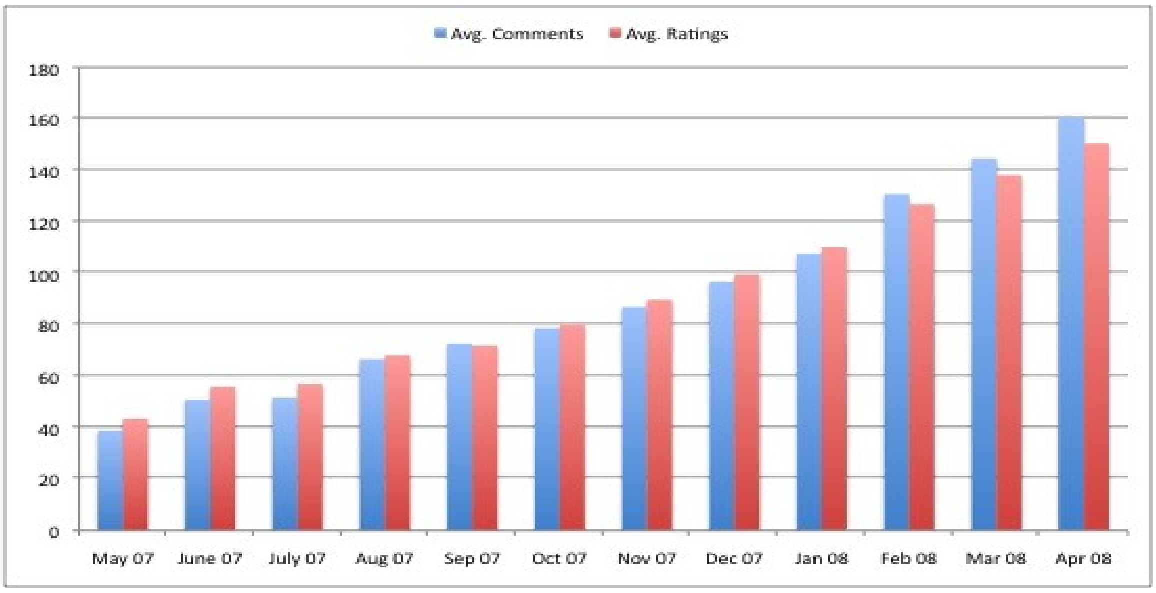
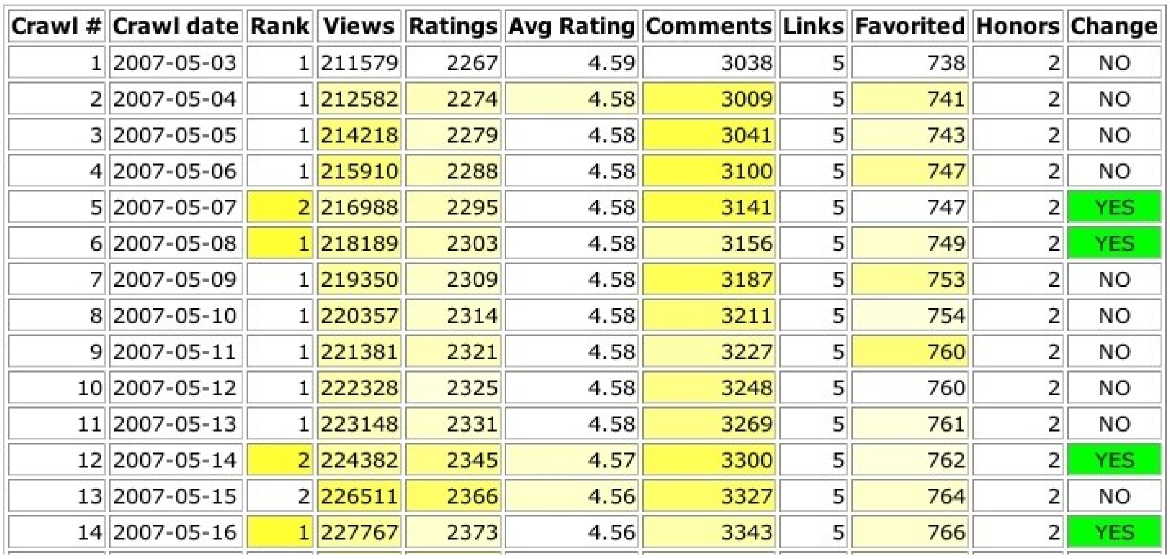


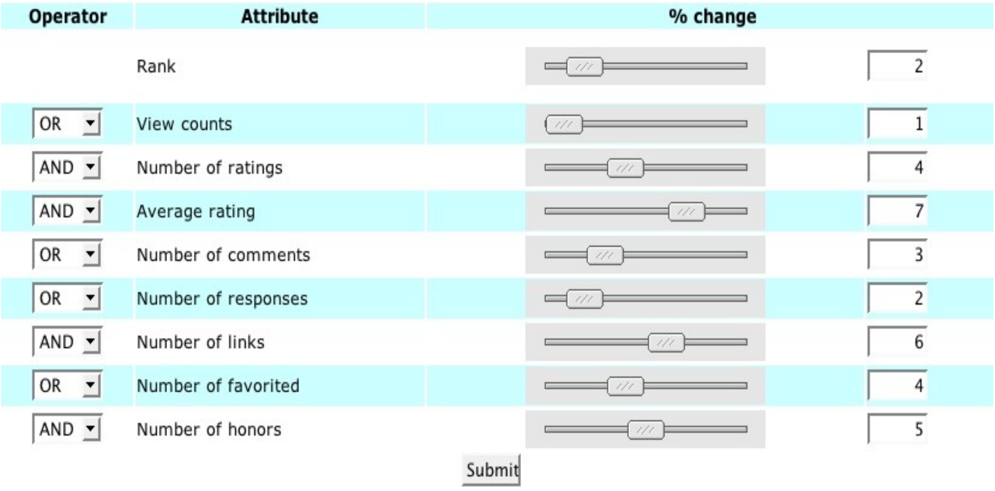
Figure 4. Average number of comments and ratings for the collected videos over 12 months.



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

Figure 5. ContextMiner reports the changes in various parameters between the consecutive crawls. The amount of change is represented by the intensity of the background yellow color. The right-most column details the change (YES or NO) based on the parameters set by the curator (see Figure 6).

Given what we have in our collection, there are several directions we could take in analyzing it. These analyses, among other things, also helped us revise our planning and collection processes.



Operator	Attribute	% change
	Rank	2
OR	View counts	1
AND	Number of ratings	4
AND	Average rating	7
OR	Number of comments	3
OR	Number of responses	2
AND	Number of links	6
OR	Number of favorited	4
AND	Number of honors	5

Figure 6. Setting monitoring preferences for a query.

Some of the analyses that we performed are summarized below.

1. We were interested in looking at "significant" changes between two crawls for a video. This was accomplished by providing a mechanism in our interface to report the intensity of a change (Figure 5), as well as letting a curator decide what constitutes a "significant" change (Figure 6).
2. Not everyone who participates on YouTube plays an equally vital role. We were interested in identifying those key people who do play such a vital role who might affect a video (and through that, the population) significantly. We adopted an idea of identifying collectors, mavens, and salesmen in a population from (Gladwell, 2002) and applied it to our collection (Shah & Marchionini, 2008). We showed that a small number of people had a huge influence on production, consumption, and participation relating to the videos in our collection.
3. The majority of the videos in our collection follow a "uniform" pattern, but there are a few videos that exhibit a different behavior. This includes an unusually high number of comments (popular on YouTube), or in-links (popular on the Web). We found these videos doing linear regression among a set of parameters. Identifying such outliers can help a curator in explaining and documenting various trends and anomalies in a collection.
4. We did some analysis on a collection of blog posts as well as our own collection of YouTube videos to understand how we could help digital library curators with decisions on what to collect and what are the trade-offs with different processes. This is reported in Capra et al. (2008) as well as Clemens, Capra, Lee, & Sheble (2008).
5. We are now looking at the relevance of our collected videos on a given topic. We know intuitively that the most of the videos we collected with our election crawler are about the election, but if we zoom in further, do we also find that the query Barack Obama brought us videos that were about Barack Obama? How can curators quickly make such judgments to inform the processes they have employed? We are working on this kind of analysis at present.

Crawling YouTube for videos and contextual information is a part of a big picture of capturing contextual information for digital curation (Shah & Marchionini, 2007). A core idea of this research is to identify what we need to tell the whole story about a digital object (Marchionini, Tibbo, Shah, & Lee, 2007). This involves, among other things, selection of the objects, capture of the context, and preservation. We have learned quite a bit from our experiences in the past few months through this harvesting/crawling processes, some of which is enumerated here:

1. *Backup*. There were times during our collection process when we were afraid of losing our valuable data. Had that happened, all the hard work would have gone to waste.
2. *Get it before it is gone*. Many videos disappear from YouTube due to various reasons. Whenever we did not download the video or extract the contextual information immediately after encountering it, we ran a risk of losing it forever.
3. *Complete automation is an illusion*. The crawlers that we developed were designed to do everything automatically, including producing the analysis reports. However, the very nature of automation also made it difficult to debug when things went wrong. Our crawlers were running around the clock, automatically, and it was often hard to work around their schedules. As we mentioned earlier in the article, our goal was to provide enough information, tools and support to the curator for various curatorial tasks, but not to replace the curator altogether.
4. *There is nothing like too much data*. When we began crawling operations, we thought that obtaining the top 100 videos for a query would be too much. We expected we would reduce it to some reasonable number after a few weeks of crawling, but we continued the same process thereafter. In fact, later we found ourselves debating whether we should obtain even more.[5]
5. *Everything is contextual and anything could be context*. We have been collecting nearly two dozen attributes for each video including all the text comments during every crawl. This seems sufficient to provide the context for a video. But then we found blog posts about these videos, and other sorts of related in-links. Of course, at some point someone (the curator) has to make a decision where to stop, but potentially almost anything could serve as context.

The lessons learned from the YouTube harvesting experiments also helped us build a general purpose tool called TubeKit, designed to create YouTube crawlers. It allows one to build one's own crawler that can crawl YouTube based on a set of seed queries and collect up to 24 different attributes. TubeKit assists in all the phases of this process, starting with database creation and ultimately giving access to the collected data via browsing and searching interfaces (Shah, 2008).

Conclusions

Preserving ephemeral digital videos is not only a task that is intellectually engaging, but also a process that is both culturally and socially important and valuable. These videos which acquire importance, popularity, or attention for a certain short period of time can tell us a lot about the cultural and social values, opinions, sentiments, and community dynamics of that time. However, simply storing the videos is not sufficient. If we want to ensure that future generations can access and make sense of these videos, we need to add value to them. This goes beyond storing some metadata. In this article I argued that, in order to make sense of the videos being preserved, we need to store additional contextual information. This problem of capturing context brings up a host of research questions including defining context, capturing and validating it, and then presenting it. The following subsection

[5] And we indeed did during some extra crawls in January 2008.

summarizes the efforts reported in this article to address some of these questions.

Summary

A high-level view of the ContextMiner model is depicted in Figure 7. From this figure, various components of ContextMiner that are described in the present article can be divided into two major parts: curator/system side, and user side.

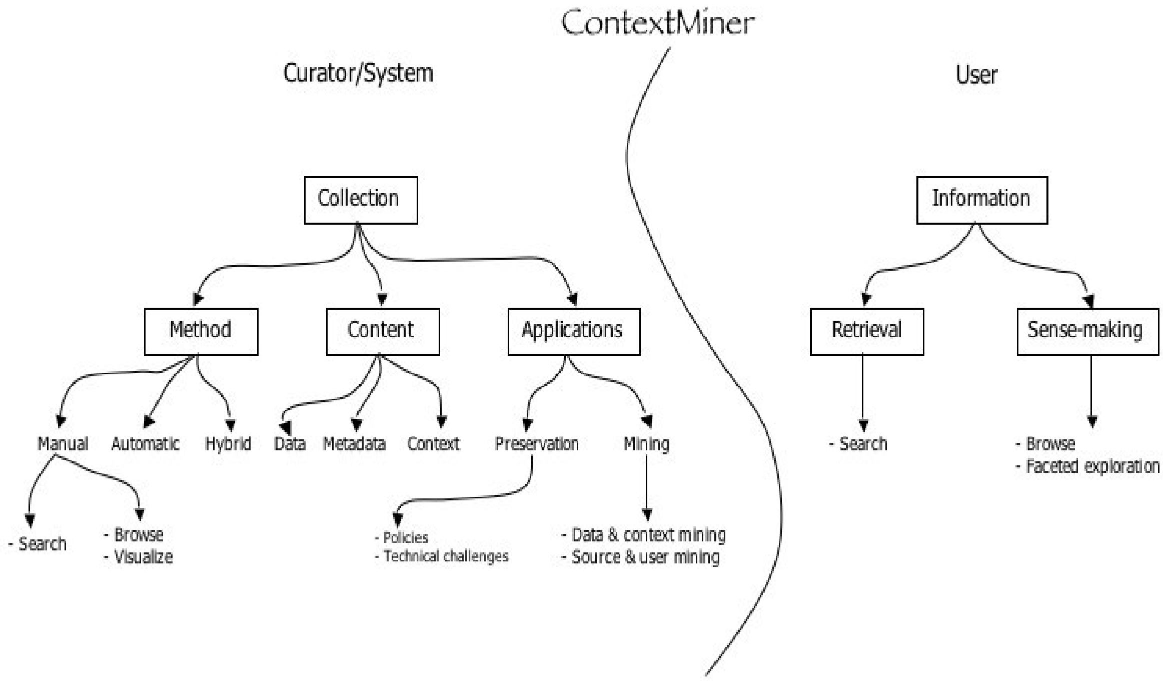


Figure 7. A higher-level view of the ContextMiner system.

The user side of ContextMiner aims to address some of the issues of providing end-user access to the collected and/or preserved information. There are some obvious questions about user interfaces here, and less obvious questions about the usability of preserved information and selecting contextual information for sense-making.

The majority of the work so far, however, has been on the curator/system side of ContextMiner. It revolves around a curator building a collection for a digital library or for preservation. In the present article I described several methods and interfaces to aid the curator in collecting information from different sources and build collections. There are three main aspects to our efforts: what to collect (content), how to collect (method), and what to do with it (applications). This collection building can be done by manually dealing with each record, doing it completely automated, or employing a hybrid approach.

So far our work has focused on using this content to understand various issues related to preservation, though it can be used for various other applications too, such as data mining. There are several interesting outcomes that can emerge through mining the crawled data in ContextMiner. For instance, Figures 8, 9 and 10 present normalized view counts, comments count, and ratings count for this collection. As we can see from these figures (Figure 8), there was an apparent change in activities relating to the videos in our collection between August 22 and 26, 2007. A curator monitoring such a collection can find it interesting to investigate this incident further.[6] It is likely that this sudden change in activity relating to the videos in the collection was triggered by a significant event. An important contribution to the work reported here is the proposal and implementation of the methods that can capture such ephemeral events for preservation.

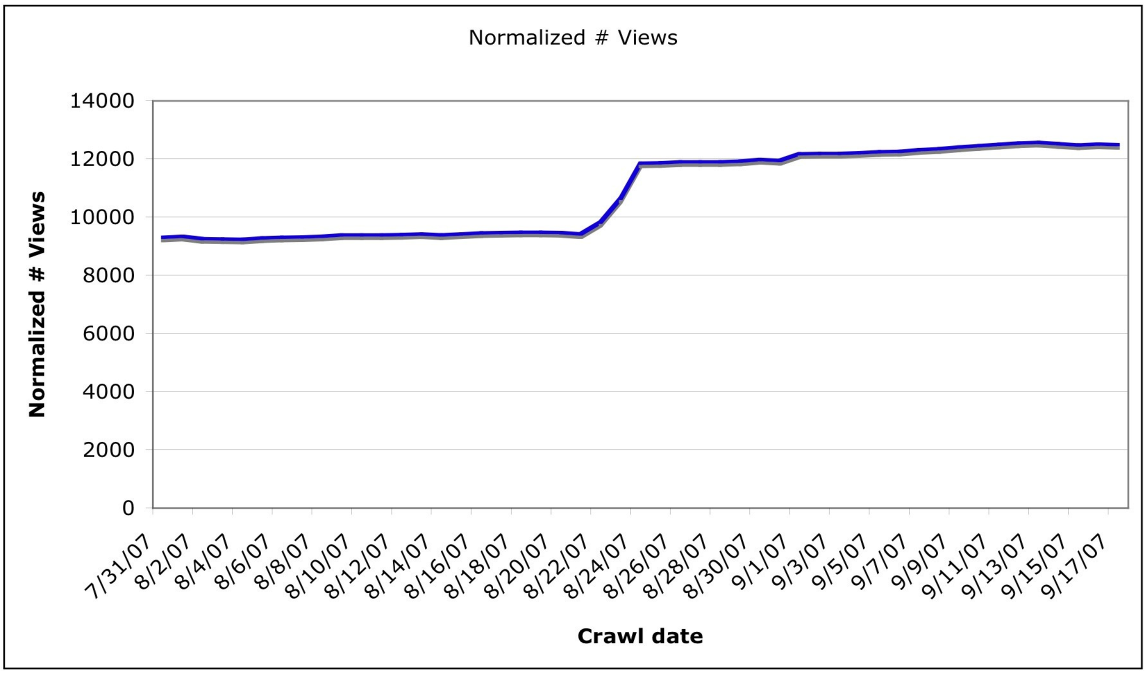


Figure 8. A snapshot of number of views (normalized) for the collection.

Ongoing and Future Work

The ContextMiner Project started with a vision of creating tools for a curator to capture contextual information relating to digital videos being preserved. In the process we not only achieved a sound understanding of defining and capturing context, but also developed a set of policies, interfaces, and data that have become extremely important resources in studying issues on digital video preservation. There follows some of the ongoing and planned activities using these resources.

[6] Possible actions may involve looking closely at some of the videos, or exploring external sources such as news sites and blogs.

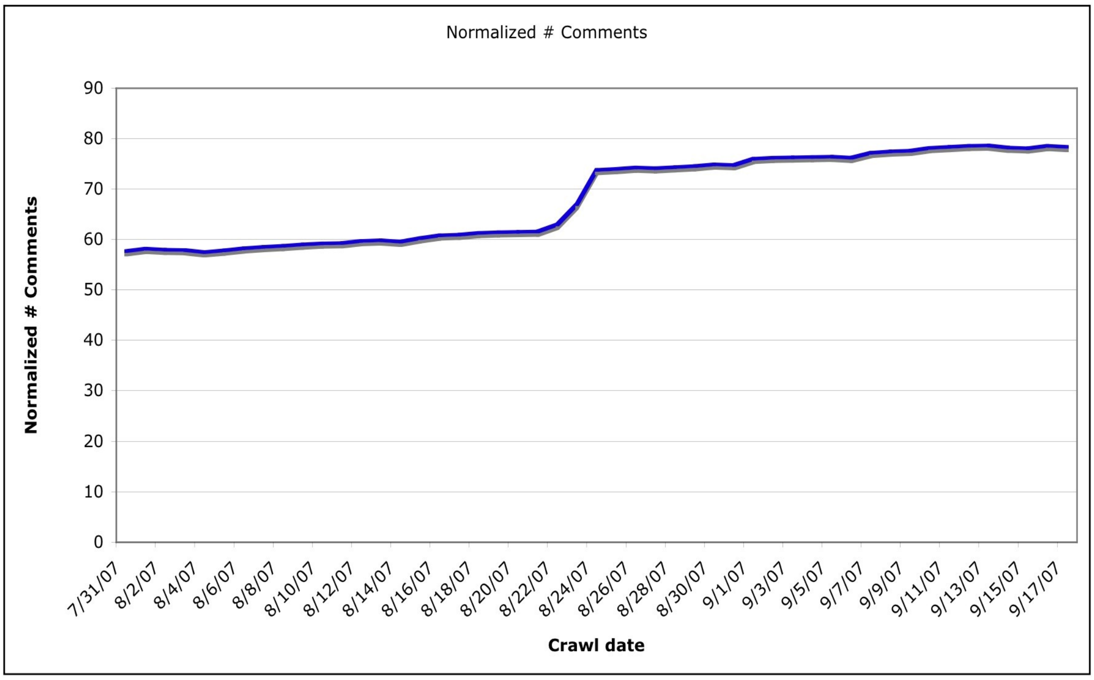


Figure 9. A snapshot of number of comments (normalized) for the collection.

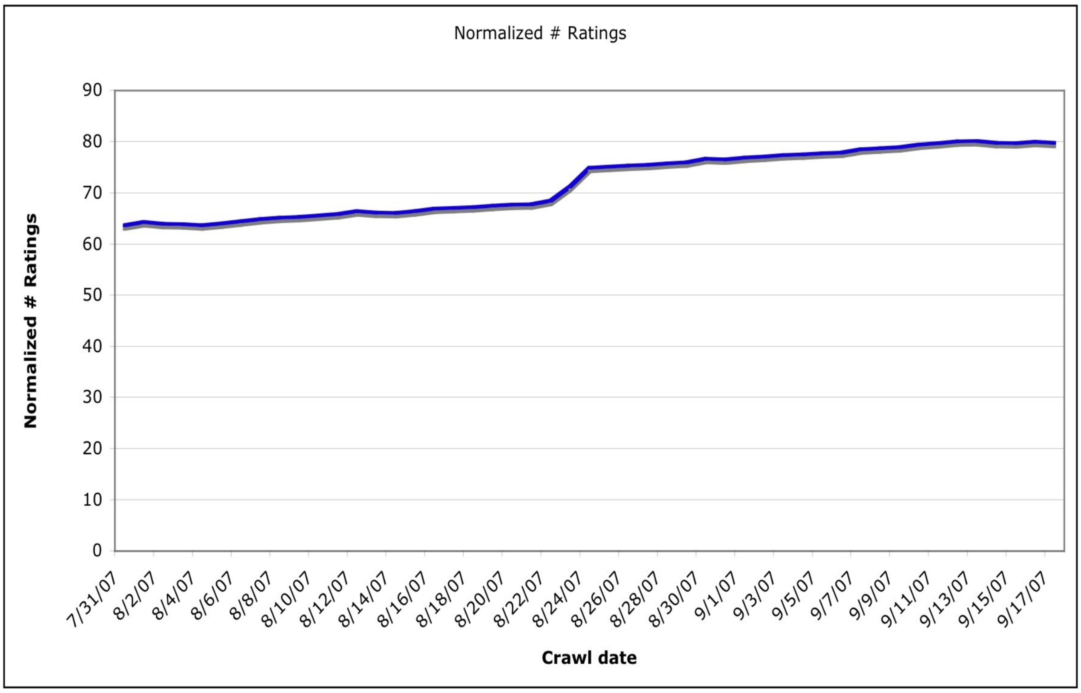


Figure 10. A snapshot of number of ratings (normalized) for the collection.

1. We are working on finding and explaining various scenarios that use ContextMiner to build a preservable collection of digital videos.
2. We are also testing the interface for a curator for its functionality, comprehensibility, and usability. We plan to accomplish this first by doing a cognitive walkthrough with the users who are adequately knowledgeable about the curatorial process. Given the interfaces, data, and results as reported earlier in this section, we want to investigate how a curator can make informed

decisions for building a collection. This will involve revising policy decisions (e.g., selecting sources, queries, formats), addressing some technical challenges (e.g., storage), and understanding the applications (e.g., event monitoring, preservation). We hope to exploit the experience to build a completely revised interface for the curator.

3. We have released a public beta of ContextMiner system,[7] which allows the curators to harvest data and contextual information from various sources, such as YouTube, blogs, Twitter, and Flickr. ContextMiner is also used by several members of the National Digital Information Infrastructure Preservation Program (NDIIPP),[8] and can be used by teachers or other related parties who wish to harvest content on specific topics.
4. We have made our data, collected using ContextMiner, available for research use under a Creative Commons License.[9] We have also developed TubeKit, a toolkit to develop harvesters for YouTube, and made it available under a Creative Commons License.[10]

Acknowledgments

The author is indebted to his advisor Gary Marchionini for his constant encouragement and guidance; and Barbara Wildemuth, as well as other members of the VidArch team, Rob Capra, Paul Jones, Sarah Jordan, Cal Lee, Terrell Russell, Laura Sheble, Yaxiao Song, and Helen Tibbo -for their feedback on and criticisms of this work. The work reported here is supported by NSF grant # IIS 0455970.

References

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)*.

Clemens, R., Capra, R., Lee, C., & Sheble, L. (2008). Contextual information from blogs in video digital curation. *Proceedings of Society of American Archivists 2008 Research Forum*.

Gladwell, M. (2002). *The tipping point: How little things can make a big difference*. Back Bay Books.

Lee, C. A. (2007). *From simply finding to making sense of digital objects: Toward an information model for contextual information*. Technical report, SILS, UNC Chapel Hill.

Marchionini, G., Tibbo, H. R., Shah, C., & Lee, C. A. (2007). Telling the whole story: Selecting and collecting web-based videos for archival collections. In *Proceedings of International Digital Curation Conference*, Washington D.C., December 2007.

[7] ContextMiner http://www.contextminer.org/

[8] Digital Preservation (Library of Congress) http://www.digitalpreservation.gov/

[9] Details can be found at http://idl63.ils.unc.edu/chirag/ContextMiner/

[10] TubeKit - A YouTube Crawling Toolkit http://www.tubekit.org/

Shah, C. (2008). TubeKit -A query-based YouTube crawling toolkit. In *IEEE ACM Joint Conference on Digital Libraries (JCDL)*.

Shah, C. (2009). Mining contextual information for ephemeral digital video preservation. *International Journal of Digital Curation*, *4(1)*, pp. 175-192.

Shah, C., & Marchionini, G. (2007a. Capturing relevant information for digital curation. In *IEEE ACM Joint Conference on Digital Libraries (JCDL)*, p. 496.

Shah, C., & Marchionini, G. (2008). Hunting for hip, hipsters, and happenings on YouTube. *ASIS&T 2008 Annual Meeting (AM08 2008)*.