ContextMiner - Collect Different

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

ABSTRACT

We present *ContextMiner*, a web-based service for collecting contextual information for digital objects from a variety of sources. *ContextMiner* lets one run campaigns that can include a set of queries that *ContextMiner* can run on various sources, such as YouTube and blogs, and keep extracting and adding contextual information to the collected objects based on their usage. Such contextual information can help to make sense of digital objects and better preserve them.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*Collection*; H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*Web-based interaction*

General Terms

Design, Human Factors, Management

Keywords

Digital curation, Digital Preservation, Contextual information

1. INTRODUCTION

ContextMiner is a framework to collect, analyze, and present contextual information along with the data. It is based on an idea that while describing or archiving an object, contextual information helps to make sense of that object or to preserve it better (Tibbo, Lee, Marchionini, & Howard, 2006; Marchionini, Tibbo, Shah, & Lee, 2007). This idea has been realized as a web-based service, called *ContextMiner*,[1] that provides tools to collect data, metadata, and contextual information off the web by automated crawls (Figure 1).

ContextMiner helps one (1) run automated crawls on various sources on the web and collect data as well as contextual information, (2) analyze and add value to collected data and context, and (3) monitor digital objects of interest over a period of time.

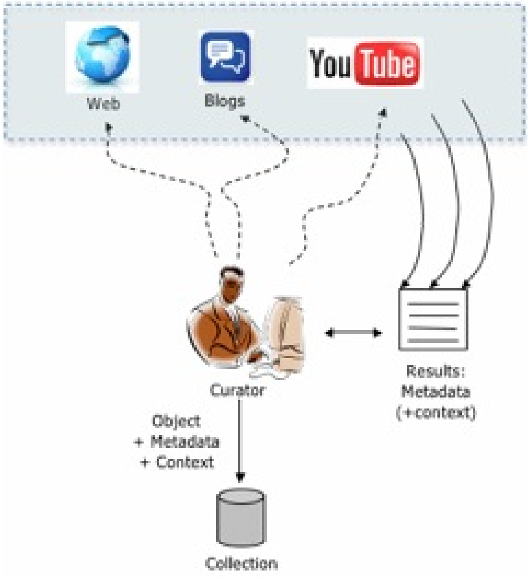


Figure 1: *ContextMiner* architecture

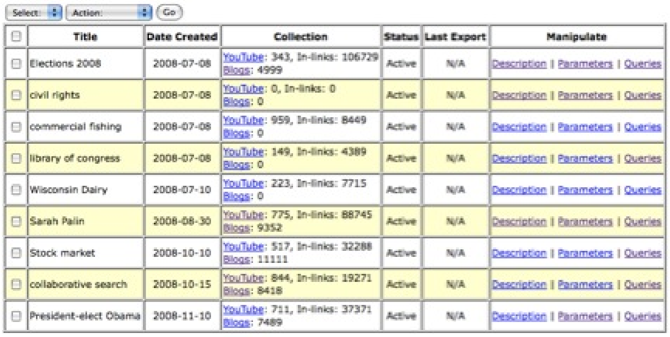
2. USING CONTEXTMINER

Once a user signs up for a free account, he/she can immediately start creating campaigns. A campaign in *ContextMiner* is a project that is based on running several automated processes and collecting data, metadata, and contextual information. Following is a typical flow of using *ContextMiner*:

1. Start a new campaign based on some story, concept, or an object.
2. Choose the sources (Web, Blogs, YouTube) that you want *ContextMiner* to do your searches and crawls on.
3. Once you provide all the required parameters, *ContextMiner* can immediately start running your campaign. You can access all your campaigns and collected data as well as contextual information through its website (Figure 2).

[1] http://www.contextminer.org/

4. You can manipulate individual items as well as related items that are collected by the above processes to add your interpretation and meaning to the campaign.

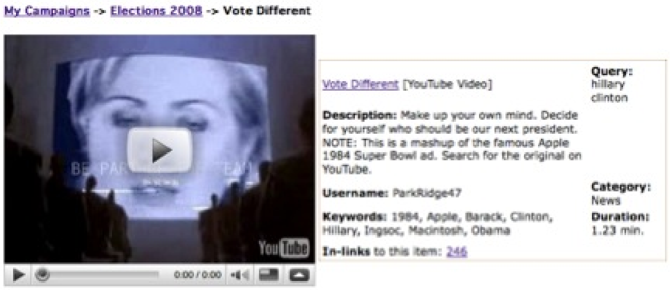


□	Title	Date Created	Collection	Status	Last Export	Manipulate
□	Elections 2008	2008-07-08	YouTube: 343, In-links: 106729 Blogs: 4999	Active	N/A	Description | Parameters | Queries
□	civil rights	2008-07-08	YouTube: 0, In-links: 0 Blogs: 0	Active	N/A	Description | Parameters | Queries
□	commercial fishing	2008-07-08	YouTube: 959, In-links: 8449 Blogs: 0	Active	N/A	Description | Parameters | Queries
□	library of congress	2008-07-08	YouTube: 149, In-links: 4389 Blogs: 0	Active	N/A	Description | Parameters | Queries
□	Wisconsin Dairy	2008-07-10	YouTube: 223, In-links: 7715 Blogs: 0	Active	N/A	Description | Parameters | Queries
□	Sarah Palin	2008-08-30	YouTube: 775, In-links: 88745 Blogs: 9352	Active	N/A	Description | Parameters | Queries
□	Stock market	2008-10-10	YouTube: 517, In-links: 32288 Blogs: 11111	Active	N/A	Description | Parameters | Queries
□	collaborative search	2008-10-15	YouTube: 844, In-links: 19271 Blogs: 8418	Active	N/A	Description | Parameters | Queries
□	President-elect Obama	2008-11-10	YouTube: 711, In-links: 37371 Blogs: 7489	Active	N/A	Description | Parameters | Queries

Figure 2: Running campaigns with *ContextMiner*

3. USAGE EXAMPLE

Let us now look at an example of capturing contextual information with *ContextMiner*. One of the sources *ContextMiner* works with is YouTube. 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'.[2] 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 3. Some of the basic information about this video, including description, author name, and keywords, can also be seen.


My Campaigns -> Elections 2008 -> Vote Different

Vote Different [YouTube Video]

Description: Make up your own mind. Decide for yourself who should be our next president. NOTE: This is a mashup of the famous Apple 1984 Super Bowl ad. Search for the original on YouTube.

Username: ParkRidge47

Keywords: 1984, Apple, Barack, Clinton, Hillary, Ingsoc, Macintosh, Obama

In-links to this item: 246

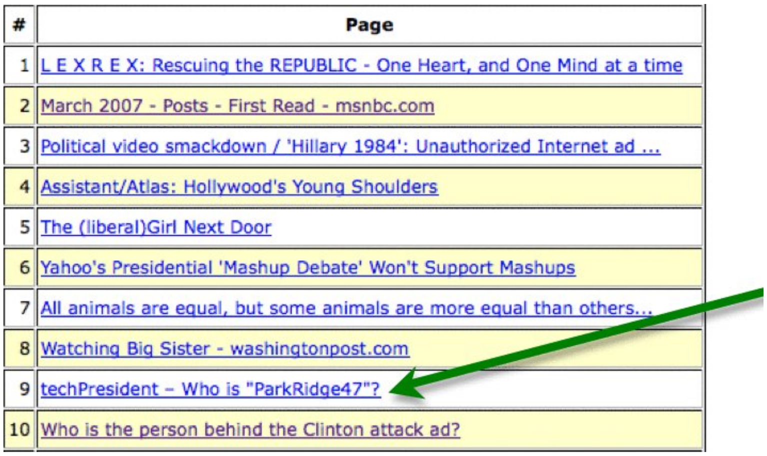
Query: hillary clinton

Category: News

Duration: 1.23 min.

Figure 3: Captured video from YouTube along with some contextual information

One of the kinds of contextual information that *ContextMiner* captures is the links from other webpages on the web (in-links) to a given digital object. Now if we look at the in-links collected to this YouTube video (Figure 4), 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 webpage, we can see that it talks about who created this video, why, and what is the background for the video. We can also find 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.



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

Figure 4: In-links to the video 'Vote Different'

4. CONCLUSION

We have been using the *ContextMiner* framework and services for harvesting videos and contextual information relating to the presidential elections 2008 (Shah & Marchionini, 2007). In addition to this, we have also been running crawls for collecting data and contextual information on a variety of topics, such as energy, epidemics, health, natural disasters, and truth commissions.

At the time of writing this, there are more than 200 users who have been using *ContextMiner* for several months, and have collected millions of objects (YouTube videos, blogs) and related contextual information. *ContextMiner* is also in use by several members of the National Digital Information Infrastructure Preservation Program (NDIIPP)[3] and can be used by teachers or others who wish to harvest content on specific topics. Further development providing access to more sources, and tools for information exploration is underway. *ContextMiner* is available as open source code or a web-based service from http://www.contextminer.org.

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).*

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.*

Shah, C., & Marchionini, G. (2007, June 29). Preserving 2008 US Presidential Election Videos. In *International Web Archiving Workshop (IWAW).* Vancouver, BC, Canada.

Tibbo, H. R., Lee, C. A., Marchionini, G., & Howard, D. (2006). VidArch: Preserving Meaning of Digital Video over Time through Creating and Capture of Contextual Documentation. In *Proceedings of archiving* (p. 210-215).

[2] http://www.youtube.com/watch?v=6h3G-lMZxjo

[3] http://www.digitalpreservation.gov/