DEPARTMENT OF COMPUTER SCIENCE

MASTER'S PROJECT

Different Methods of Visualizing Thumbnails Of Archived Web Pages

Author:

Surbhi Shankar

Advisor:

Dr. Michele C. Weigle

April 24, 2017

Acknowledgement

I express my gratitude to my project advisor **Dr. Michele C. Weigle**, Associate Professor, Department of Computer Science at Old Dominion University, for being a remarkable source of inspiration to me and for guiding me throughout the project. Without her creative ideas, detailed guidelines and her depth of knowledge in the field of Visualization, this project would not be complete. She has been very understanding and supportive at every step and hence, I could keep track of my progress and finish the project as intended.

I take this opportunity to thank **Mr. Mathew Kelly**, PhD Candidate, Web Sciences and Digital Libraries lab, Computer Science Department, Old Dominion University, for spending time in helping me understand some of the critical parts of the project which supported my work.

Finally, my sincere thanks the **Department Of Computer Science** for providing all the required resources and infrastructure to complete this project successfully.

Contents

Table of Figures

1 Preface

The web pages on the internet are dynamic and the content on the web changes very frequently. These web pages are stored in the form of Internet archives, which are used for research purposes. The idea behind the visualizations in this project is to make it easier to access the web archived pages.

There may be so many archives among which only some are important and significantly used in research. The visualizations presented in this project represent the collections from New York Art Resources and Consortium (NYARC), and Columbia University Libraries (CUL). The NYARC and CUL have active web archive collections hosted with Internet Archive's Archive-It service.

The goals that this project aims to fulfill are as follows:

- Through visualizations, showing how a single web page has changed over time, without the user having to open every memento separately.
- Reducing the search time required to look for a memento, as images are displayed for quick access.
- Highlighting the temporal and dynamic nature of a website through these visualizations.

2 Motivation

This project is a part of a bigger idea of representing internet archived pages. The idea of visualization was proposed by Dr. Michele C. Wielgle. Opening each memento to observe the changes that the web page has had, is a trivial job. To avoid this, we can use many different innovative visualization methodologies. This project has three such major designs implemented.

Love for visualization started during course Information Visualization in Spring 2016 under Dr. Weigle. Visual representations are always powerful and more effective when compared to other methods of representing data. Users can understand the dynamic nature of the web and the differences between each memento just by looking at the images. This whole concept was very interesting for me because these visualizations are different from the usual ones which are usually used in every application. I was excited to see how data can be represented differently and what the final product would look like. The archived pages are easily accessible based on their appearance as snapshots of the mementos that are distinct are being used.

3 Sources of Data and Cleaning

Collecting data for this took a lot of time because NYARC and CUL have a vast collection of archived pages. Snapshots of every page has been stored and used for visualizations. All the mementos are chosen from collections of the New York Art Resources and Consortium (NYARC) and Columbia University Libraries (CUL). I selected some of the archived collections from NYARC and CUL which are distinct from each other. Links and images for the same are displayed below.

- https://wayback.archive-it.org/1068/*/http://www.cageprisoners.com/
- https://wayback.archive-it.org/1757/*/http://www.brooklynbridgepark.org/
- https://wayback.archive-it.org/1068/*/http://www.cofadeh.hn/
- https://wayback.archive-it.org/1068/*/http://gulflabor.org/
- https://wayback.archive-it.org/1757/*/http://www.atlanticyards.com/
- https://wayback.archive-it.org/4847/*/http:/peterblumgallery.com/
- http://wayback.archive-it.org/1068/*/http://www.amnesty.org/

The archived pages are hosted with the Internet Archive's Archive-It service. NYARC's collections are relatively smaller and more recent ones. On the contrary, CUL collection is vast and maintains a long-running human rights collection when compared to other collections.

The URIs collected from these Archived collections and snapshots are stored. The path for the images and also the URIs are stored in a JSON file. JSON file is created using a python script. Input for the python code will be a text file containing all the URIs listed. A snapshot of a sample JSON file for the same is given below.

```
[
  - {
        website: "http://www.amnesty.org/ ",
        photo: "20110902210647.png",
        content: "http://wayback.archive-it.org/1068/20110902210647/http://www.amnesty.org/ ",
        start: "02-09-2011",
        time: "21:06:47",
        id: "1"
    },
  - {
        website: "http://www.amnesty.org/ ",
        photo: "20120302210436.png",
        content: "http://wayback.archive-it.org/1068/20120302210436/http://www.amnesty.org/ ",
        start: "02-03-2012",
        time: "21:04:36",
        id: "2"
    },
  - {
        website: "http://www.amnesty.org/ ",
        photo: "20130307001447.png",
        content: "http://wayback.archive-it.org/1068/20130307001447/http://www.amnesty.org/ ",
        start: "07-03-2013",
        time: "00:14:47",
        id: "3"
    },
  - {
        website: "http://www.amnesty.org/ ",
        photo: "20141001174452.png",
        content: "http://wayback.archive-it.org/1068/20141001174452/http://www.amnesty.org/ ",
        start: "01-10-2014",
        time: "17:44:52",
        id: "4"
    },
  - {
        website: "https://www.amnesty.org/en/",
        photo: "20151001222753.png",
        content: "http://wayback.archive-it.org/1068/20151001222753/https://www.amnesty.org/en/",
        start: "01-10-2015",
        time: "22:27:53",
        id: "5"
    }
]
```

Figure 1: JSON file format for Image Slider and Image Grid views

The images below are the samples of snapshots of the archived pages which are used for visualizations.

Figure 2: Sample Thumbnails used for visualizations.

The JSON file that is used for Timeline Visualization is slightly different due to the way the data is used in JavaScript files. The structure of the file is shown below.

```
[
  - {
        timestamp: 1298321593,
        event_series: "Mementos",
        event_html: "<img src='http://www.cs.odu.edu/~sshankar/MS_Project/timeline/photos/BrooklynBridgePark/20110221155313.PNG' height='300px' width='300px' />",
        event_date: "Feb. 21, 2011",
        event_display_date: "",
        event_description: "",
        event_link: "https://wayback.archive-it.org/1757/20110221155313/http://www.brooklynbridgepark.org/"
    },
  - {
        timestamp: 1298322085,
        event_series: "Thumbnails",
        event_html: "<img src='http://www.cs.odu.edu/~sshankar/MS_Project/timeline/photos/BrooklynBridgePark/20110221160125.PNG' height='300px' width='300px' />",
        event_date: "Feb. 21, 2011",
        event_display_date: "",
        event_description: "",
        event_link: "https://wayback.archive-it.org/1757/20110221160125/https://www.brooklynbridgepark.org/"
    },
```

Figure 3: JSON format for Timeline Visualization.

4 Introduction

Visual aspects are important for scholars and researchers of art history. As they are extending research in the field of born-digital domain, visual results for their searches are more effective and navigable than any current methods of presentations. The designs used in this project are more user-friendly and helps to access the web archives easily and is faster when compared to other methods.

Currently, the web archives are presented in ways which may be difficult to search for a particular page that is required. It is difficult to search through every archived page to find what is needed. This process is more time consuming and also the interface may not be user-friendly. The image below shows how web archives are made accessible for research purposes, which may be inefficient sometimes.
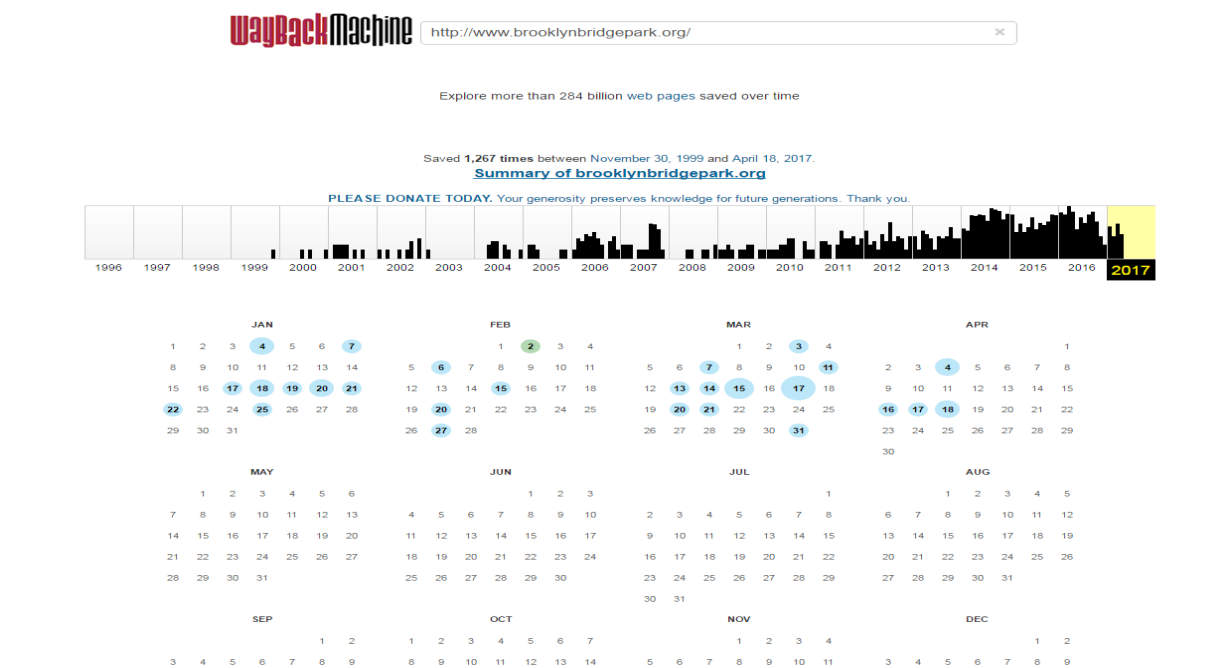

Figure 4: Visualizations on Wayback machine website using calendar view and bar graph.

Figure 5: One of CUL's Archived webpage collections, which needs a user to open every link to see how the website evolved.

This project is based on the archived pages from Archive-It collections which belongs to NYARC and CUL. This project can be used with an implementation of Ahmed AlSum's 2014 ECIR paper titled "Thumbnail Summarization Techniques for Web Archives"[1], which Mr. Mat Kelly and team worked on. This is an implementation for Web Archiving Incentive Program for Columbia University Libraries' grant, "Visualizing Digital Collections of Web Archives". Mat Kelly helped me understand the parts of the implementation which is useful for my project. Also, explained the algorithm and the way the output is designed to work.

Creating search result which has everything on a single screen is very helpful in terms of specific kinds of research and also as a tool for presenting the deployment of web archives in lectures and publications. This helps increase the use of web archives by the scholars, researchers and art historians, who have begun using the web archives extensively to aid their work. NYARC supports and sees the potential advantage of having a visual and user-friendly interface of comprehensive

---

[1] http://thumbnails.cs.odu.edu:15421/

snapshots of an artist's preferred work highlights over time. This is a better way to study and research about artistry and history which is digitally captured through some of the websites.

CUL supports research on building web archives in thematic areas corresponding to existing collection strengths. The Human Rights web archives is a huge collection that belongs to CUL. The evolution of public areas and environment in New York City are documented by the Avery Library Historic Preservation and Urban Planning web archive. They publish all the essential and important information captured by non-governmental organizations, advocacy campaigns, parks conservancies, neighborhood associations and other such organizations and projects that is usually found online only.

5   Overview

This is a system which contains three different parts in it – three different kinds of TimeMaps[2] representing the archived web pages. Having a visual representation or thumbnails[3] of each of the mementos[4] is very useful and effective in quickly determining the evolution of the webpage over a certain period and also helps a user to focus on the webpage. I used JSON file for the data and used technologies like HTML, JavaScript, JQuery and Ajax to create all the visualizations. All the visualizations are based on images of the web archived pages of websites from the archived collections of NYARC and CUL as discussed earlier. There are no graphs or trends used in this system and the visualizations show the best possible ways to be able to access the web archives.

The idea behind "Image Slider Visualization" is to imitate the photo roller property used in a photo album in iPhoto. Simply by moving the cursor over the images, which are the snapshots of the archived pages of a particular collection, we can see the images sliding and click on the page that we are interested in exploring more. There are additional properties added to this visualization, Play/Pause buttons, for example.

The "Image Grid View" is a much simpler representation that shows a responsive grid view to show all the snapshots of archived pages at one glance. This helps in viewing and comparing all the snapshots and choosing the ones that are subject of interest. The "Timeline View" is a very effective visualization of web archives as it represents the snapshots of the archived web pages on a time line based on the date that it was archived. Users can easily search and select the page based on year that the page was archived and also can see how distinct the pages are.

---

[2] *TimeMap* is a list of all the mementos for a single webpage.
[3] *Thumbnail* is a small image representation of archived pages taken as a snapshot.
[4] *Mementos* are a selected set of archived versions of a webpage which can be explored individually.
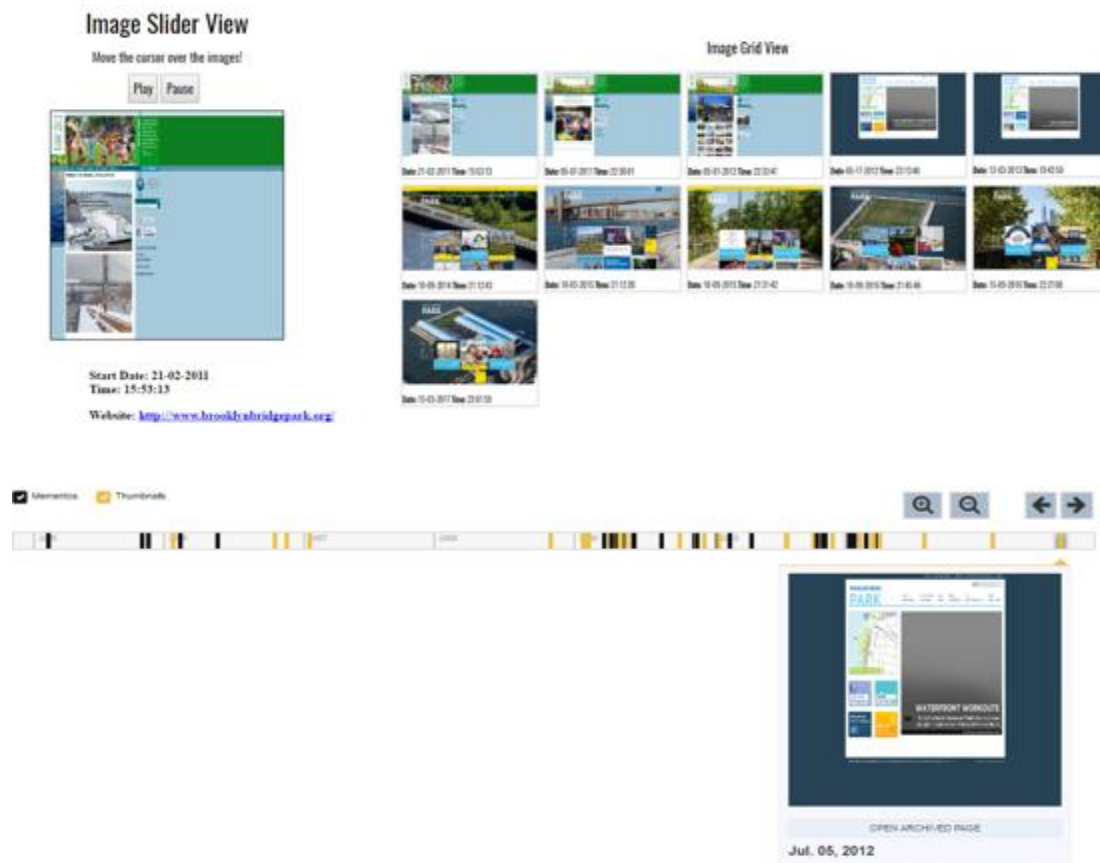
Figure 6: Overview of the system

To create timeline visualizations, I used the libraries that are used in Timeline Setter website - https://www.propublica.org/special/tbi-psycho-platoon-timeline. There is an attribute in the JSON file named as "Timestamp", using which the lines representing the archived time of the website have been marked on the timeline. Timeline attribute is nothing but the dates that are converted to milliseconds format. I used this website to convert Human readable date to timestamp - http://timestampconvert.com/

All the three visualization systems were put together using "iframe" tag in order to be able to compare the visual effectiveness of each of the representations. Insights for all of the visualizations are given in the section. Every archived collection is separately placed, the links for which are listed below.

6 Insights on the System

I examined each of the visualization while before and after developing them to see if the design is effective and if the design is achieving the basic goal of this system. The common goal of all these visualization is to provide a user-friendly interface to researchers and scholars who use web archives for their research. I have a brief description and observations of each of the Visualizations listed down here.

6.1   Image Slider Visualization

a.   Snapshots of archived pages which are distinct is displayed on a single thumbnail view, along with the information related to the page below the image itself.

b. The thumbnail images change when the cursor is moved over the grid which makes it easier for a user to pick their choice of archived page.

c. Play/Pause button is included to make the image slider work like a video. Also, when clicked on the image, it navigates to the corresponding archived page.

d. Image slider helps go through all the thumbnails of various archived pages and compare between the snapshots how distinct they are from each other.

6.2   Image Grid View

a. Snapshots are placed as a 5 X 5 grid with images placed next to each other giving room for comparison between the thumbnails showing the snapshots of the archived pages.

b. The interface is responsive so that more snapshots can be added at any point of time to the JSON file from where the data is pulled out from.

c. Under each image, date and time of archival is displayed. On clicking on any of the image, it navigates to the archived page itself.

    d. This visualization helps users to see all the thumbnails together and easily view the distinction between them.

6.3   Timeline Visualization

    a. The interactive timeline view includes all the thumbnails on a timeline which indicates the time when the page was archived.

    b. The lines on the timeline represent the thumbnails and mementos and on clicking on the line we can see a twitter card style thumbnail displayed below the timeline.

    c. This difference between this and other visualizations is that depending on the size of the TimeMap, the timeline includes mementos that are not selected as a part of the summary as well.

7 Challenges

This project was a great learning experience for me. I had an opportunity to work different kinds of visualization techniques and also learnt about web archiving concepts. However, the project was not very easy for me and there were a few challenges that I faced in the path. Some of the challenges I faced while I working on the project are mentioned below.

i.   Collecting data from each of the archived collections was a time consuming task. I had to open each page and determine if they are distinct and then select the mementos to be represented in the TimeMap. Initially, it was easy because I had to do a mock up for just one of the collections, but later this task got tougher because there are six different collections that I have represented in the TimeMap. Moreover, this data grows whenever it is archived and hence the data file has to be updated each time.

ii.   The python script that I wrote to get the data in a JSON file was working well with one of the collections as all the URIs were of same length. If the number of characters varies, then the script has to be changed a bit with respect to positions to get the JSON file correctly. This was also a difficult job considering the number of JSON files that had to be created.

iii. Algorithm was difficult to understand in the beginning. There were some of the Node.js and Phantom.js libraries and packages which were not compatible in MAC OS and also few packages which were outdated hence had to be replaced. This was also a difficult task as it was the first time for me to run a Node.js code.

iv.   Timeline visualization has a JSON file with a slightly different format. I used a separate source for date to timestamp conversion. Also, I had to create separate JSON files for each of the collection again as the structure is different.

8 Future Enhancements

Here are some of improvements and further implementation ideas, although the project is substantial.

- This project can be coupled with the Thumbnail Summarization implementation by just obtaining the dump of all the distinct URIs in a text file. This can be put into a JSON file in the format that is required and represented in the TimeMaps implemented in this project.
- For Image Grid view, we can also include a two dimensional responsive interface and have resizable thumbnails. This way, the thumbnails arrange itself depending on the number of records in the data file.
- Timeline implementation can be improvised by making it more interactive. Including filters to select years or months can be one such example. Also, we can include months and dates as we zoom into the timeline. We can also include zoom-in and zoom-out property with scrolling of the mouse.
- The data files have to be updated every time a new archival is added to the collection. Instead a script can be written to update the data file periodically such that the data is just appended to the existing data files.