Using Web Archives to Enrich the Live Web Experience Through Storytelling

Yasmin AlNoamany
Old Dominion University
Norfolk, VA, USA
yasmin@cs.odu.edu

ABSTRACT

The web has become an integral part of our lives, shaping how we get news, shop, and communicate. When critical events occur, social media and news websites cover the stories as they break and continually revise them as the story evolves. Unfortunately, much of the content around these stories is vulnerable and prone to loss. Thus, web archives have become a significant repository of our recent history and cultural heritage. Content from web archives can be used to fill in the gaps in the live web about the evolution of the story of an important event. Every story is made up of a sequence of events. In this research, events are exemplified through corresponding web pages from the live web and web archives, (semi-)automatically discovered, arranged in a narrative structure ordered by time, and replayed through an appropriate visualization interface.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: Web Archives—*Retrieval models*

General Terms

Archives, Logs, Methodology, Analysis, Human Factors

Keywords

Web Archiving, Information Visualization, Web Usage Mining, User Access Patterns, Content Mining, Storytelling, Information Retrieval

1. MOTIVATION

Do you recall the relevant dates, people, and web pages to replay the story of the Egyptian Revolution? We posit that there are three information needs for either learning about or telling a story: overview information, recency of information, and replaying the story. Overview is well served by Wikipedia (Figure 1(a)), and news web sites serve recency (Figure 1(b)), Google can help us to discover these, but replaying the story as it captured by news web sites has not yet been served.

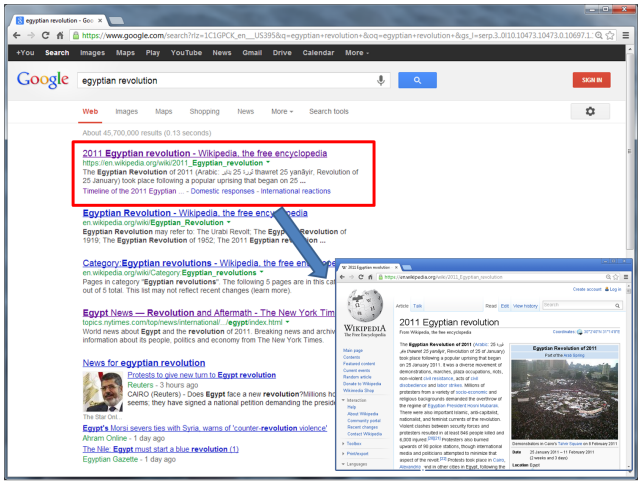
1.1 Problem Statement

The web has become an integral part of our lives, shaping how we get news, shop, and communicate. For many, it has also become the first resource when an important event occurs [26]. Most news sites cover important events and revise their reporting as the story evolves. Unfortunately, the nature of the web is ephemeral, and the expected lifetime of a web page is short [14, 21, 30]. This can cause access to information about an event to decay rapidly after a while and make it difficult to retrieve how the story of an important event evolved over time. The evolution of the story and the context in which it was reported are important for preserving our cultural heritage. Because of this, web archives have become a significant resource for preserving our recent history. Even though web archives can fulfill this important function, they may be under-utilized by human users [3], as the general public is either not aware of their existence or of how to access archived web pages. One way to enhance the use and utility of web archives is to use them to enrich the live web experience through the re-telling of important stories.

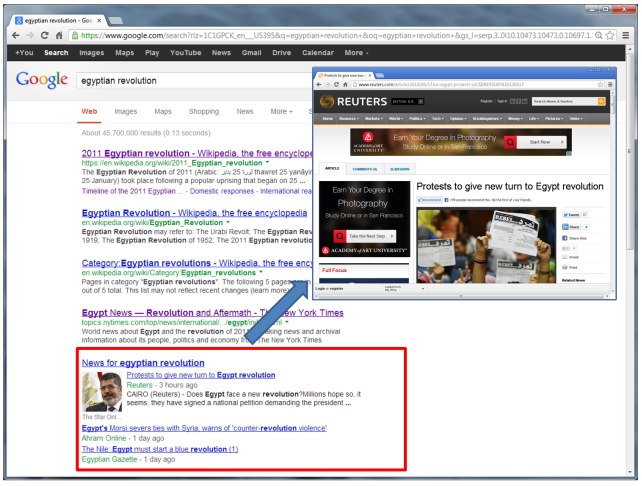
1.2 Current Capabilities

To demonstrate the limitations of the current capabilities in replaying a story, we will use the "Egyptian Revolution" case study with real news data. Searching Google is the first thing that the people do when they want to know about a story. However, searching Google with the "Egyptian Revolution" query term is not enough to replay the story as it happened. Figure 2(a)[1] shows the results of the first page of Google search. As expected, the Wikipedia article, which has a summary of the entire story, appears as the first result. Wikipedia articles give high level details, but how the story unfolded in the media and how it started is still not clear in the summary. Searching by date will not be easy after many years because people usually remember event names more than dates. Regarding our earlier question, few people remember the exact date of the Egyptian Revolution. Specifying the search in Google by date from Jan. 1, 2011 - Feb. 28, 2011 does not provide any clues for the beginning of the Egyptian Revolution (Figure 2(b))[2]. Searching on social media brings the latest results easily, but the past results are hard to retrieve. Figure 2(c)[3] shows an important page in the Egyptian Revolution story because this page was used to invite youth to the demonstration events in Egypt. Even

[1]Captured on June 17, 2013
[2]Captured on June 17, 2013
[3]Captured on June 14, 2013

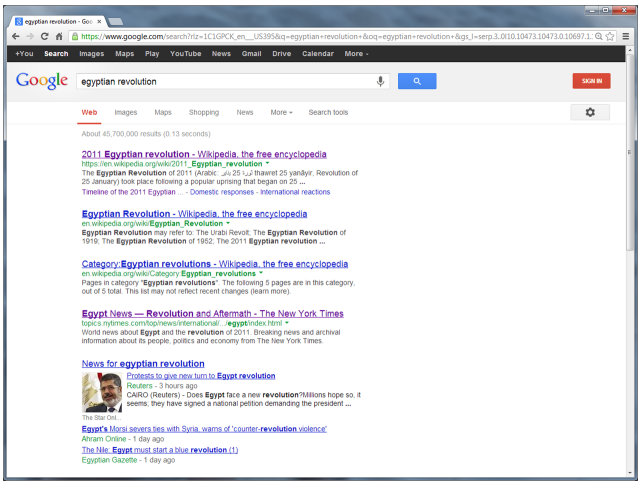


(a) Overview of Information

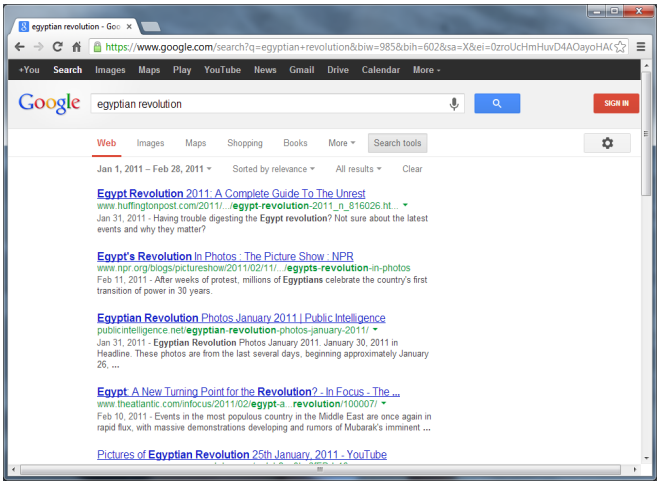


(b) Recency of information

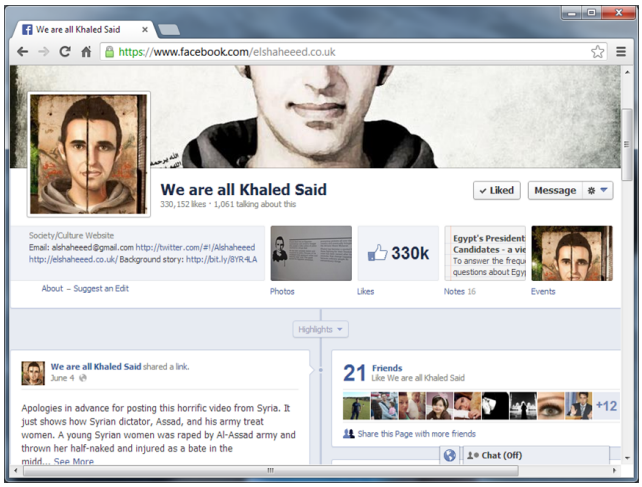
Figure 1: Google provides overview and recency of information (Captured on June 17, 2013).



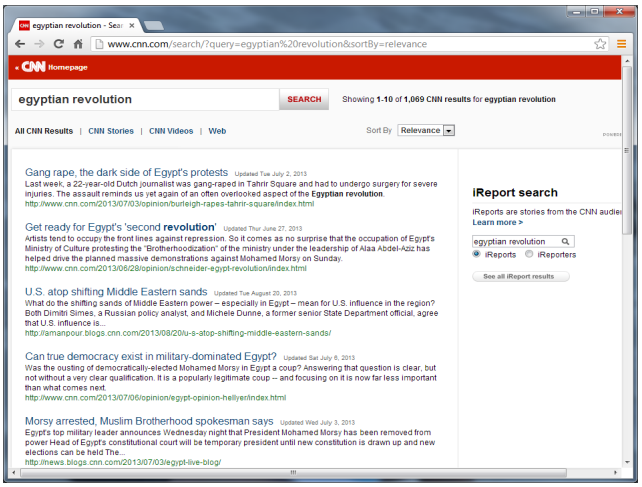
(a) Searching Google



(b) Searching Google between Jan. 1, 2011 - Feb. 28, 2011



(c) The Facebook page that started the events of the Egyptian Revolution



(d) Searching CNN

Figure 2: Searching for "egyptian revolution".

knowing about the page requires a background knowledge about the event; the current status of the page does not represent the story. Also news web sites are hard to search. As we see in Figure 2(d)[4], the recent pages about the story are at the top of the results and there are not many options to specify the search. Furthermore, most news sites retain only the summary of the story.

Storify[5] is a free social media service that allows people to manually create stories from Facebook, Twitter, and other web resources. Figure 3(a) contains a story related to the Egyptian Revolution on Storify. One of the problems that faces the reader of this story is that many links of this story are no longer available [38], for example the link in Figure 3(b). Thus, Storify is for bookmarking, not for preserving the links, and the user will not be able to determine the link's content, especially if the context of the text around the link does not contain enough information about the link.

A subscription service of the Internet Archive[6] (IA), called Archive-It[7] (Figure 4), allows institutions to build, manage, and preserve their own web collections. Half of the Archive-It collections are for government sites, and the rest are event-specific events (such as Pakistani flood, Boston Marathon Bombing, the Egyptian Revolution, etc.). The event-specific collections seed URIs are manually collected by people based on domain knowledge, which means there is no policy for automatically collecting the seed URIs. The other issue with Archive-It collection is the crawling time. In many cases, seed URIs are submitted and crawling begins well after the start of the event, so important information may not be captured. For example, Figure 4(b) shows the date of crawling of the first web site of the Egyptian Revolution collection (Figure 4(a)) on Archive-It, which is after the important events of the revolution had passed.

1.3 Egyptian Revolution's Hand-Crafted Story

We used IA's Wayback Machine [32] to manually retrieve copies of http://cnn.com with the datetime of the Egyptian Revolution. The awareness of the event datetime and the existence of web archive materials are two limitations that may prevent casual users from building these stories.

Figures 5 and 6 contain different snapshots of the timeline of "Egyptian Revolution" as it appeared on cnn.com. Note that all of these mementos (archived web pages) are for http://cnn.com except the mementos in Figures 5(a) and 6(f). These mementos are of specific articles. Presumably because IA did not adjust their crawling strategy in time, there are no archived copies for Feb. 11, 2011, which contains the important news that captured that Mubarak stepped down.

January 25, 2011 was the beginning of anti-government protests in Egypt (Figure 5(a)). Newspapers started full coverage of the protests with the increasing number of protesters because of violent clashes between security forces and protesters (Figures 5(b), 5(c), 5(d), 5(e)). Figure 5(c) shows the reaction of the people toward the curfew, and Figure 5(e) shows the beginning of military interaction. Figures 6(a) and 6(b) show other ways the government tried to stop the protests. Social media played a significant role in the revolution. The organization of the protests started on Facebook. The government shut down access to the Internet and suppressed the media to close the communications (Figure 6(a)). On the second day, Mubarak supporters attacked the protesters in Tahrir Square, riding on camels and horses (Figure 6(b)). On the Friday of departure, Mubarak stepped down after 30 years of rule (Figure 6(f)). As we can see through these figures, replaying the events can be more compelling than a summary.

1.4 Research Questions

This research aims to integrate the past with the present by automatically creating, identifying, and linking stories culled from the past web that are related to the content of a live web page or a specific event. Based on the terminology that was introduced earlier, we have the following resarch questions:

- How do we define the time frame of a story?
- How do we identify the individual events that make up a story?
- How do we identify, evaluate, and select candidate (archived) web pages to support the events?
- How do we visualize the result?

This raises some further questions:

- Can we leverage the content of social media services to discover stories?
- Can we extract stories based on user access patterns of the Wayback Machine?
- Can we associate the names that people give particular events with their datetimes in order to find them in web archives?

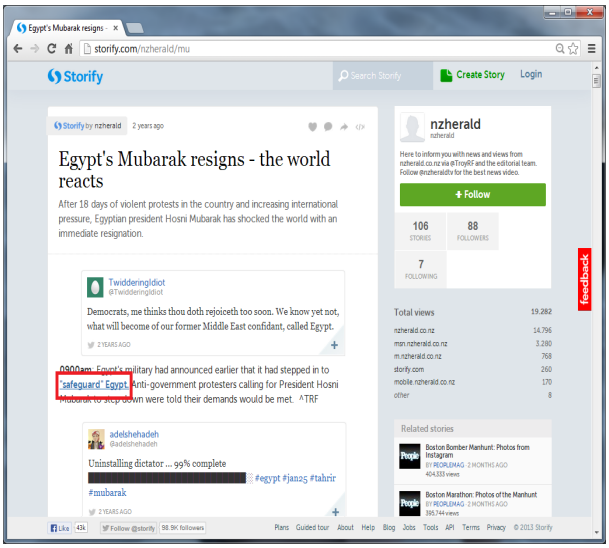
2. RELATED WORK

In recent years, there has been much interest in the use of web pages and social media to create stories. For example, Storify is a social network service launched in 2010 that allows users to create stories or timelines manually using tweets, Facebook artifacts, YouTube videos, and Instagram. The service depends on users' efforts in creating a timeline for the stories. It may demonstrate the widespread interest in an event, but the evolution of the story about the event as it appeared in the media is still not clear.

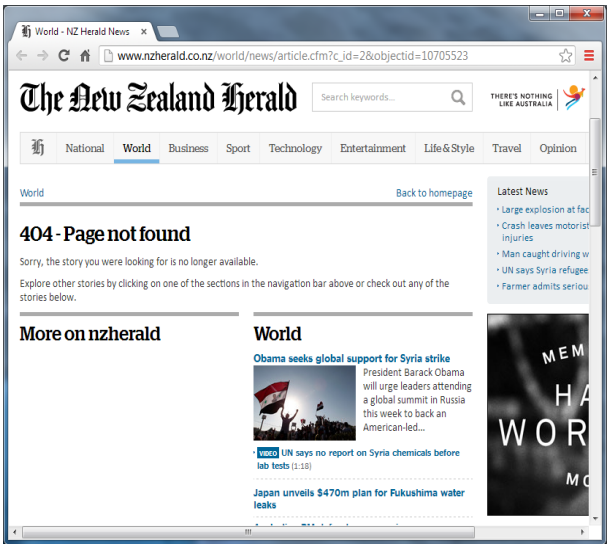
2.1 Time Series Visualizations

Many studies have been conducted recently in the visualization community for exploring and visualizing online stories. Most of these studies have been devoted to summarizing the text data and its evolution over time [11, 29, 22, 37, 23]. Dou et al. developed LeadLine, an interactive visual analytics system to automatically identify meaningful events in

[4]Captured on August 21, 2013
[5]http://storify.com
[6]http://archive.org
[7]http://www.archive-it.org

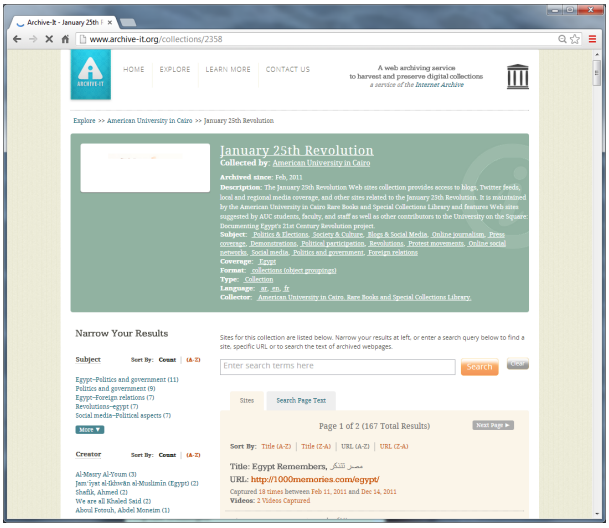


(a) Related story to the Egyptian Revolution

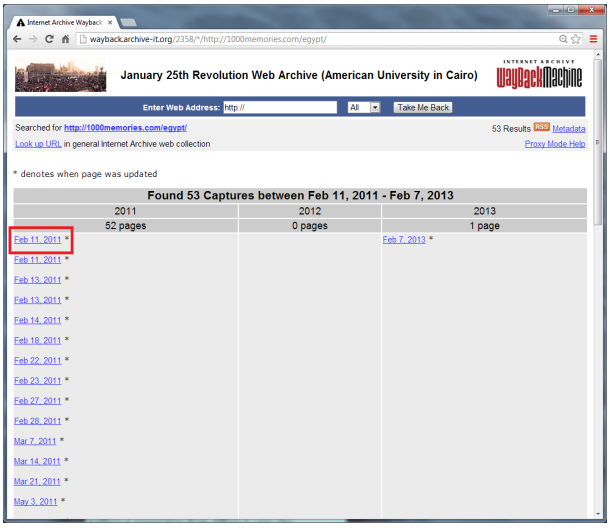


(b) The bookmarked link is broken

Figure 3: Storify is for bookmarking, not for preserving. When the annotated link (on the left) is requested, it results a 404 (on the right.)

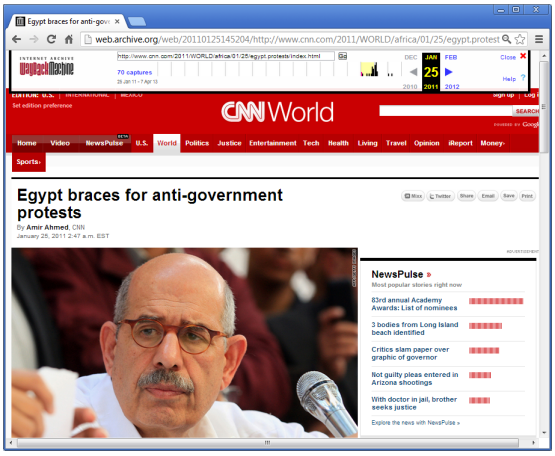


(a) Egyptian Revolution collection on Archive-It

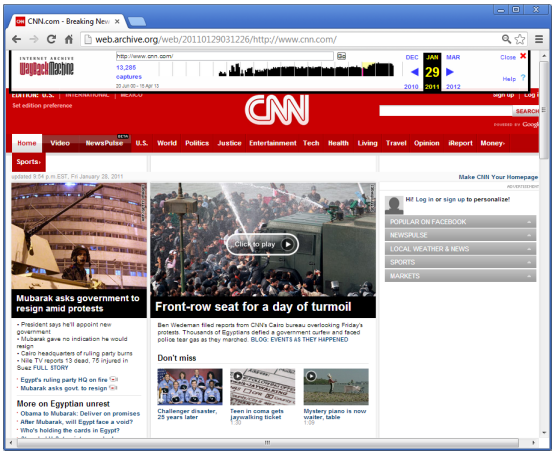


(b) Crawling date is Feb. 11, 2011

Figure 4: The coverage of Archive-It to the Egyptian Revolution. The figure on the right shows that the crawling date for one of the URIs in the collection of the Egyptian Revolution started Feb. 11, 2011, which was after Mubarak stepped down.



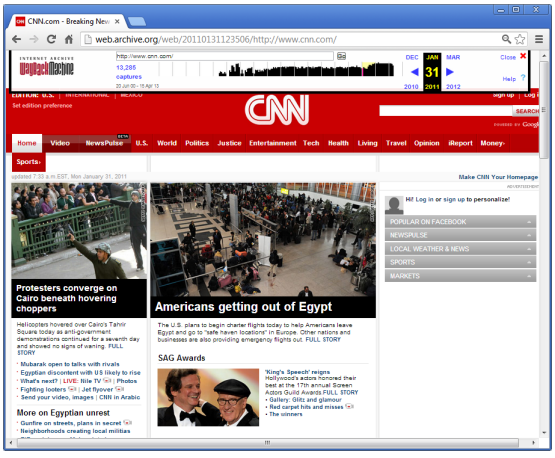
(a) 25 January 2011 2:47 a.m. EST



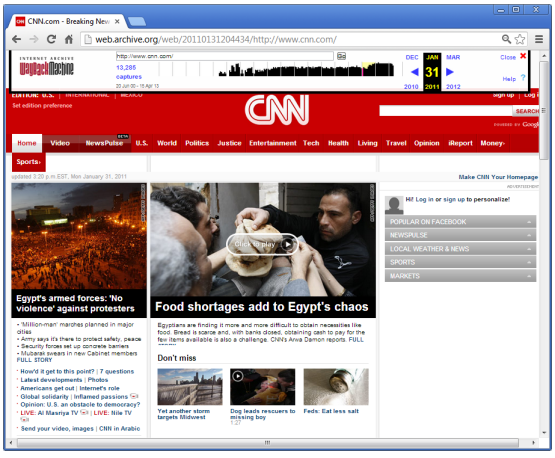
(b) 28 January 2011 9:54 p.m. EST

(c) 28 January 2011 12:49 p.m. EST

(d) 29 January 2011 7:12 a.m. EST

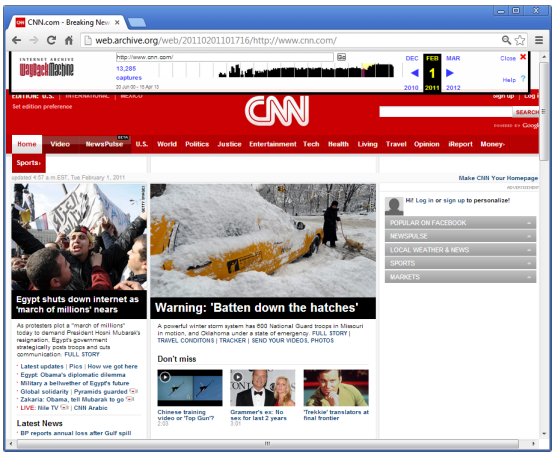


(e) 31 January 2011 7:33 a.m. EST

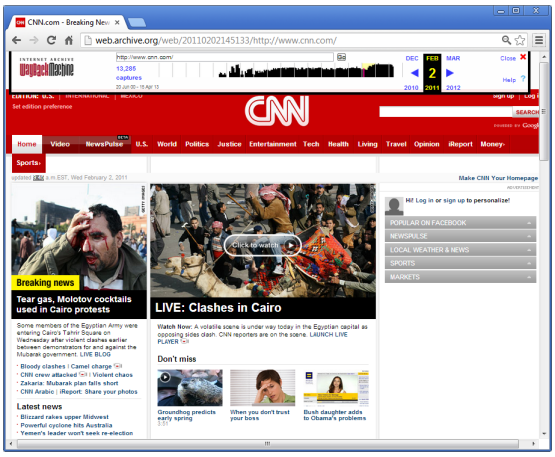


(f) 31 January 2011 3:20 p.m. EST

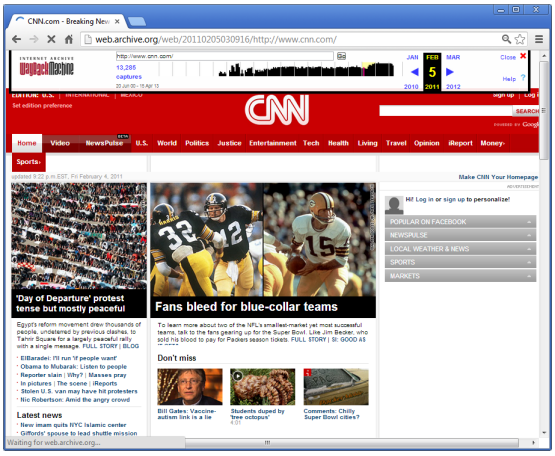
Figure 5: Coverage of 25 Jan. Egyptian Revolution on cnn.com.



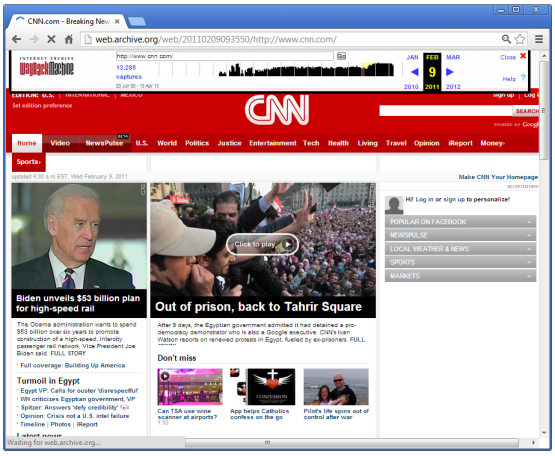
(a) 1 February 2011 4:57 a.m. EST



(b) 2 February 2011 9:45 a.m. EST



(c) 4 February 2011 9:22 p.m. EST



(d) 9 February 2011 4:30 a.m. EST

(e) 10 February 2011 7:23 p.m. EST

(f) 11 February 2011 2:14 p.m. EST

Figure 6: Coverage of 25 Jan. Egyptian Revolution on cnn.com (continued).

news and social media data and support exploration of the event [11]. LeadLine summarizes and visualizes events over time based on the 4Ws (who, what, when, where) of each event, then allows users to interactively explore these events.

Luo et al. proposed EventRiver, a visual analytics approach for event-based automated text analysis and visualization [29]. EventRiver allow users to browse, searche, track, associate and investigate the events. It presents events in a river-like metaphor in which the semantics and the temporal influences of the events are visually depicted in a temporal context to reveal the narrative arcs of the log-term stories in a display that looks like a river of events flowing over time.

CloudLines is a visual analytics technique to visualize context as a continuous flow [22]. CloudLines provides a compact visualization for time series event data with a lens-based interaction for direct access to overlapping events.

Rose et al. introduced a visual analytics system to identify and understand trends and changes from streaming information over time and for linking essential content from information streams over time [37].

Further research conducted on understanding the growth and the evolution of news corpora, through integrating topic evolution algorithms in interactive visualization methods for supporting effective analysis of the growing news corpora [23]. The system clusters documents in streaming news data based on similarity of textual content, then extracts the important topics and stacks the main keywords of the news in columns. The system uses node-link-based visualization and depicts the topical change over time.

2.2 Memento Framework

Memento [44] is an HTTP protocol extension that enables time travel on the web through interlinking current resources with their prior state. Memento introduces content negotiation in the datetime dimension using a special HTTP header, Accept-Datetime [43]. Memento is unique in that a "time traveling" browsing session begins with the current URI, for example http://cnn.com, instead of requiring the user to navigate to a specific archive. Although Memento is an effective tool for discovering these sites, many users may not know the times of the events, so they want to see the events as narrative-based more than time-based (the way Memento is currently constructed). Memento defines the following terms, which we will adopt in the rest of the paper:

- URI-R denotes the original resource. It is the resource as it used to appear on the live web; it may have 0 or more mementos (URI-Ms).
- URI-M is an archived snapshot, or memento, for the URI-R at a specific datetime, which is called Memento-Datetime. e.g., URI-M$_i$= URI-R@t_i.
- URI-T denotes a TimeMap, a resource that provides a list of mementos (URI-Ms) for a URI-R with their Memento-Datetimes, e.g., $URI{-}T(URI{-}R) = \{URI{-}M_1, URI-M_2, ..., URI-M_n\}$.

2.3 Determining Datetime of Web Pages

For the purpose of this research, we need to identify the datetime of the suggested web pages that will serve the context of the story. The related web pages should be ordered and presented to the user by datetime. There has been research in the area of automatically estimating the creation dates of content elements of pages [17, 18]. Most of these studies were browsing applications for Web archives. Estimating the date of a web page by looking at the pages that link to it has been done by Jatowt et al. [16] and Nunes et al. [33]. SalahEldeen et al. [38] also presented "carbon date", a simple web application that estimates the creation date of a URI by polling a number of sources of evidence and returning a machine-readable structure with their respective values. We will use these tools where possible.

2.4 Web Mining

Web mining is the application of data mining techniques to extract knowledge from Web data. Web mining is divided into three research areas according to the kinds of web data to be mined [12]:

- Web content mining, which is the process of extracting information from the content of web documents.
- Usage mining, also known as web-log mining, studies user access information from server log data in order to extract interesting usage patterns and to understand and better serve the needs of Web-based applications.
- Structure mining, which uses hyperlinks within web pages for discovering structure information from the Web.

The research on the area of web mining is massive and increasing [4, 25, 39, 5, 40, 31, 19]. The obtained results from web mining can be used in different applications, such as web traffic analysis, site modification, system improvement, personalization and business intelligence, and usage characterization.

Lui et al. [28] proposed an approach to automatically classify user navigation patterns and predict user future requests. The approach is based on the combined mining of Web server logs and the character N-grams for the representation of the content of the retrieved Web pages. In their study, they integrated an off-line mining system into an online Web recommendation system to observe and to calculate the degree of user satisfaction on the generated recommendations derived from the predicted requests by the system.

Adams et al. explored the usage patterns of scientific and historical data repositories [1]. However, their study focused on a variety of archive types (e.g., public vs. private, digital but non-web resources) and does not directly address the issues of archiving the web. The usage of web archives in general has not been widely studied. We will highlight the few studies that have been conducted on web archive usage.

The characterization of search behavior and the information needs of web archive users have been studied by Costa et al. [8, 7] based on quantitative analysis of the Portuguese

Web Archive (PWA) search logs. The authors introduced a comparison between search patterns of web archives and web search engines. Despite the different information needs for web archives and web search engine users, the search patterns for web archives had shown adoption of web search engine technologies. In their research, they also found that most web archive users conducted short sessions.

One of the challenges that faces web usage mining is detecting the robots who camouflage their identity and pretend to be humans. The robot detection problem has been examined in several studies [42, 9, 27, 13]. Doran et al. characterized robot detection techniques into four categories: syntactical log analysis, traffic pattern analysis, analytical learning techniques, and Turing test systems [10]. In our research, we used syntactical log analysis (simple processing by finding the self-identified robots) and traffic pattern analysis (specifying features for contrasting robots with humans).

3. PRELIMINARY WORK

In this section, we describe the work that has been done as steps toward achieving our goal. We considered several aspects in our analysis to serve user needs. We studied presenting the data of web archives efficiently to the users, understanding how users access web archives, why they come to web archives, what links to web archives, and why these things do so. In the next subsections, we examine each of these aspects and explain how they are shaping our understanding of the problem that we are studying.

3.1 Visualizing Web Archive Collections

Presenting web archive data to users is a significant issue to attract more users toward using web archives. In [34], we proposed multiple visualizations, namely image plot with histogram, wordle, bubble chart, and timeline for Archive-It collections. As mentioned earlier, Archive-It is a subscription service developed by the Internet Archive to allow institutions to harvest and preserve collections of digital content. The visualizations help to provide an overview of each collection and highlight the collection's underlying characteristics, allowing the user to progressively gain insight into the collection. For those collections that lack a curator-defined grouping, we also provided a heuristics based categorization to make the new visualizations more meaningful. We picked several Archive-It collections that differ widely from each other to test the proposed visualizations. The timeline visualization provides insight into how the collection developed over time. We plan to integrate the timeline visualization for displaying the created stories of our final framework.

3.2 User Access Patterns in Web Archives

User navigation patterns provide useful information on how users satisfy their needs. Understanding the current demand for access to web archives can provide insight into how to make the best use of limited archiving and access resources. We had multiple questions regarding user access in web archives, such as:

- How do users go through web archives?
- Do they go in deeply from URI-R_1 to URI-R_2?
- Do they browse broadly from URI-M_1 to URI-M_2 for the same URI-R?
- Do they use a combination between the previous two patterns?
- Do robot accesses are similar to human accesses?

We analyzed samples of the IA's Wayback Machine web server logs to get answers of these questions [3]. To extract user access patterns for web archives from the Wayback access logs, first we applied data preprocessing techniques (data cleaning, user identification, session identification) to determine the server sessions from the log file [6]. Then, we performed feature extraction, robot detection, and user access pattern detection.

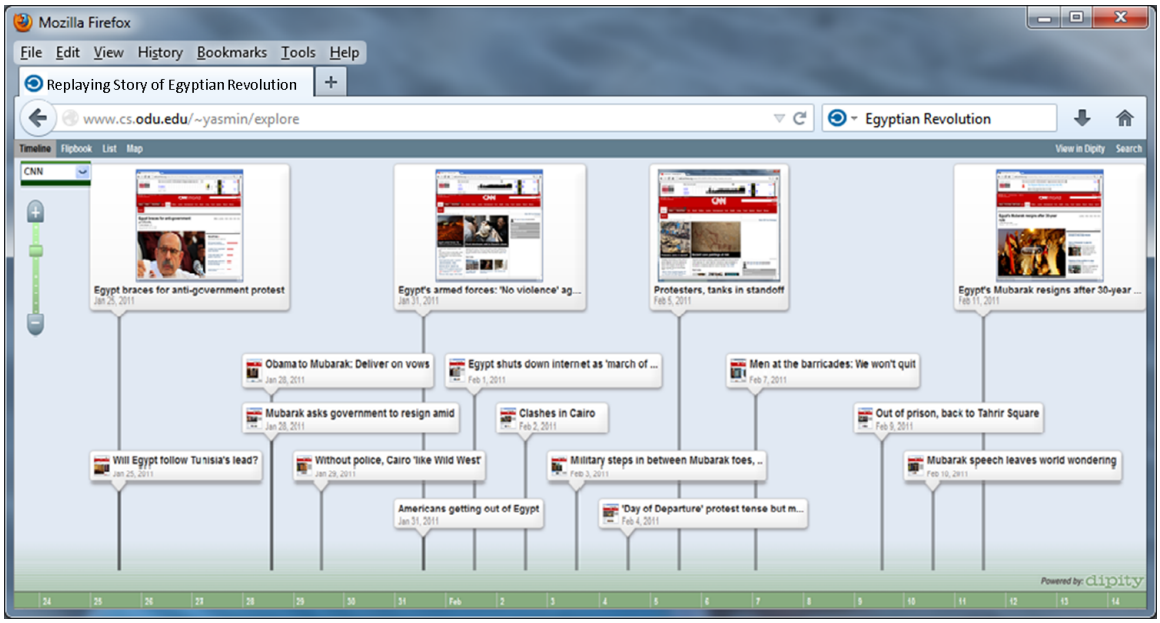
We discovered four basic usage patterns (Figure 7): Dip (a single access), Slide (the same page at different archive times), Dive (different pages at approximately the same archive time), and Skim (accessing only lists of what pages are archived). Robots are limited almost exclusively to Dips and Skims, but human accesses are more varied between all four types. We plan to extend our study on a large data set to detect stories that humans might create from their access patterns. We expect that Slides and Dives that users create on web archives may create stories around a particular event. For example, when I was creating the example in section 1, I created 2 dives (for the articles) and 1 slide (for cnn.com) in the same session. We will analyze these two patterns and examine the content of the web pages of each pattern to find stories.

Another finding was that robots outnumber humans 10:1 in terms of sessions, 5:4 in terms of raw HTTP accesses, and 4:1 in terms of megabytes transferred. Robots almost always access TimeMaps (95% of accesses), but humans predominately access the archived web pages themselves (82% of accesses). We also wanted to know users' time preferences toward the archived web pages. In terms of unique archived web pages, there is no overall preference for time, but the recent past (within the last year) shows significant repeat accesses.

3.3 Linking to Web Archives

After discovering the user access patterns in web archives for robots and humans, we wanted to study new research questions related to linking to web archives: What content languages are web archive users looking for? Why do users come to web archives? Where do web archive users come from? Who links to web archives? How do sites link to web archives? Do sites link deeply to specific archived pages or link to the repository? Why do sites link to the past? We found that users request English pages the most, followed by the European languages. Based on our analysis [2], most human users come to web archives because they do not find the requested pages on the live web. About 65% of the requested archived pages no longer exist on the live web. From analyzing the referrers of humans, we find that more than 82% of human sessions connect to the Wayback Machine from web sites, while only 15% of robots have referrers. Most of the links (86%) from websites are to individual archived pages at specific points in time, and of those, 83%

Figure 7: User access patterns in web archives (Dip, Dive, Slide, and Skim).



(a) Mockup for timeline interface

(b) Mockup for slideshow interface

Figure 8: Mockup of possible user interfaces.

no longer exist on the live web. The missing pages on the live web are big motivation for developing our framework. The need to leverage web archives into the live web becomes necessary everyday as web pages disappear.

4. RESEARCH PLAN

My plan of work is motivated by our notion of a likely usage scenario in trying to discover and browse the mementos such as those in Figure 5 and 6. After the scenario is described, the methodology to achieve this solution will be explained.

4.1 Usage Scenario

The research output could be summarized with the following scenario. Lori wants to show her friend the Egyptian Revolution story, as narrated by Figures 5 and 6 ordered by dates. She has an add-on/web application installed to create automatic stories for topics of interest or related to the content of currently browsed page. The add-on/application provides a textbox for the query, in which Lori types "Egyptian Revolution". The first step will be calculating the datetime of the story dynamically. The framework gets the URI seeds for the story and determines the datetimes of the web pages. Suppose the framework finds 10,000 related pages for each event of the story. The next step is to choose the best candidates for each event of the story and then visualize these candidates. The framework can provide many visualizations for the story. The timeline UI mockups in Figure 8(a) provides a conceptual model for the output of the framework, and Figure 8(b) shows another possible visualization, a slideshow interface. Lori has the ability to choose among different visualizations, and she is also able to specify the boundary times of the story. After creating the story, Lori has the facilities to save and share the story with her friends.

4.2 Research Methodology

In this section, we describe the general methodology for addressing the research questions and constructing a list of archived pages and web pages (Figure 9) that represent the story, arranging them in a narrative structure ordered by time, then visualizing them. The entry point for the automatic creation of the story is to define the story name, for example the "Egyptian Revolution". This step needs user interaction through entering the story name using the web browser extension (e.g., Firefox add-on or search textbox).

The next step is to specify the story beginning and ending dates. The challenge of this step is that some stories have no specific time because history is not static. For example, we thought that the Egyptian Revolution ended with Mubarak's stepping down, but Egyptians are seeing that their revolution has not yet finished. We plan to investigate different methods for computing the time of the event dynamically. This step may need user interaction for suggesting start and end dates for the story, because in many stories people's perspectives of the story are different. If we look at our Egyptian Revolution example, some people in Egypt saw what happened on June 30, 2013 as an extension for the revolution and others saw it as a coup. In this example, the beginning and ending dates of the story of the Revolution will be different from one person to another, according to their political opinions. For finding the time range of the story, we will investigate different methods for detecting the date of events, such as using the time in Wikipedia infoboxes by Hoffart et al. [15], extracting temporal expressions from unstructured text using time and event recognition algorithms [41, 20], or looking up in news sites (e.g., wikinews). We plan to evaluate this step on a gold standard dataset of stories that have previously specified datetimes.

The next problem is to specify the top K related pages automatically by obtaining the URI seeds for the story. We plan to use the list of references on Wikipedia pages and check older versions of the page for these references' URIs. We also plan to investigate using Google's API to search for the term of the story and also query the archives full text search (such as Archive-It and UK Web Archive). Because different sources give us different URI seeds for the story with different perspectives, we plan to use social media web sites (such as Storify, Twitter, and Topsy) as well. Figure 11 shows a story related to the Egyptian Revolution on Storify which contains the three main reasons (the hope for change, police brutality, and Tunisian revolution) behind the Egyptian Revolution. Such hand-crafted stories on Storify have important related URIs to the story evolution that may not be indexed by Google search or may be indexed but with low rank (Figure 2(a) and Figure 11). If we leverage the data (Figure 11) from Storify and the data from the web archives (Figures 5 and 6), we will have insight about the whole story as it happened and how it was documented by users, which may lead to significant details that could not be seen at the time of the event.

After getting the URI seeds, we will determine information retrieval techniques for finding relevant links from the seeds. We plan to evaluate this step by testing the aboutness of the web pages through checking the anchor text and its relation to the content. Furthermore, we will examine the existence of the URI in the live web and its aboutness by looking to the page content in the web archives. Another part of the evaluation will be checking the coverage of Google search against the coverage of the links from social media and from the web archives, and contrast the cost against the quality (for example, Google may have bias, but it may be faster).

There are many notions of times for each web page (e.g., creation date, modification date, archiving date, etc.). After specifying the top related pages to the story, the web pages' datetimes will be determined. Figure 10 shows the "We are all Khalid Saeed" page[8], the first page that called for the protests of Egyptian Revolution on Facebook. This page is very important for the story, but the archiving date is March 14, 2011 (i.e., two months after the start of the Revolution) after Mubarak's stepping down, while the creation date of this web page is before the revolution. Estimating the creation date of a resource is challenging problem that has been investigated in many studies before [38, 18], and we will use these results where appropriate. For evaluation, we plan to use a gold standard dataset of articles that have clear timestamps to be extracted.

The next step is choosing the best candidates for each event

[8]https://www.facebook.com/elshaheeed.co.uk

```
var mementos = [{ ...
‘‘Title’’: ‘‘Egypt braces for anti-government protests’’,
‘‘Image’’: ‘‘im001.png’’,
‘‘Snippet’’: ‘‘Egypt’s authoritarian government is bracing itself ...’’,
‘‘Memento-Datetime’’: ‘‘25 January 2011 2:47 a.m.EST’’
‘‘URI’’: ‘‘ http://web.archive.org/web/20110125024787/http://www.cnn.com/ ... ’’
}, {
‘‘Title’’: ‘‘Will Egypt follow Tunisia’s lead?’’,
‘‘Image’’: ‘‘im002.png’’,
‘‘Snippet’’: ‘‘ Demonstrators protest in central Cairo, Egypt, on Tuesday ...’’,
‘‘Memento-Datetime’’: ‘‘25 January 2011 3:00 p.m.EST’’
‘‘URI’’: ‘‘ http://web.archive.org/web/20110125030042/http://www.cnn.com/ ... ’’
}, {
‘‘Title’’: ‘‘Obama to Mubarak: Deliver on vows’’,
‘‘Image’’: ‘‘im002.png’’,
‘‘Snippet’’: ‘‘President Obama spoke with Egypt’s president moments after ...’’,
‘‘Memento-Datetime’’: ‘‘28 January 2011 8:56 p.m.EST’’
‘‘URI’’: ‘‘http://web.archive.org/web/20112801085658/http://www.cnn.com/...’’
}, {
‘‘Title’’: ‘‘Without police, Cairo ’like Wild West’’,
‘‘Image’’: ‘‘im002.png’’,
‘‘Snippet’’: ‘‘ With local police effectively no longer on the ground in ...’’,
‘‘Memento-Datetime’’: ‘‘29 January 2011 3:00 p.m.EST’’
‘‘URI’’: ‘‘http://web.archive.org/web/20110129030000/http://www.cnn.com/ ... ’’}
... ];
```

Figure 9: Sample of candidates for the Egyptian Revolution story.

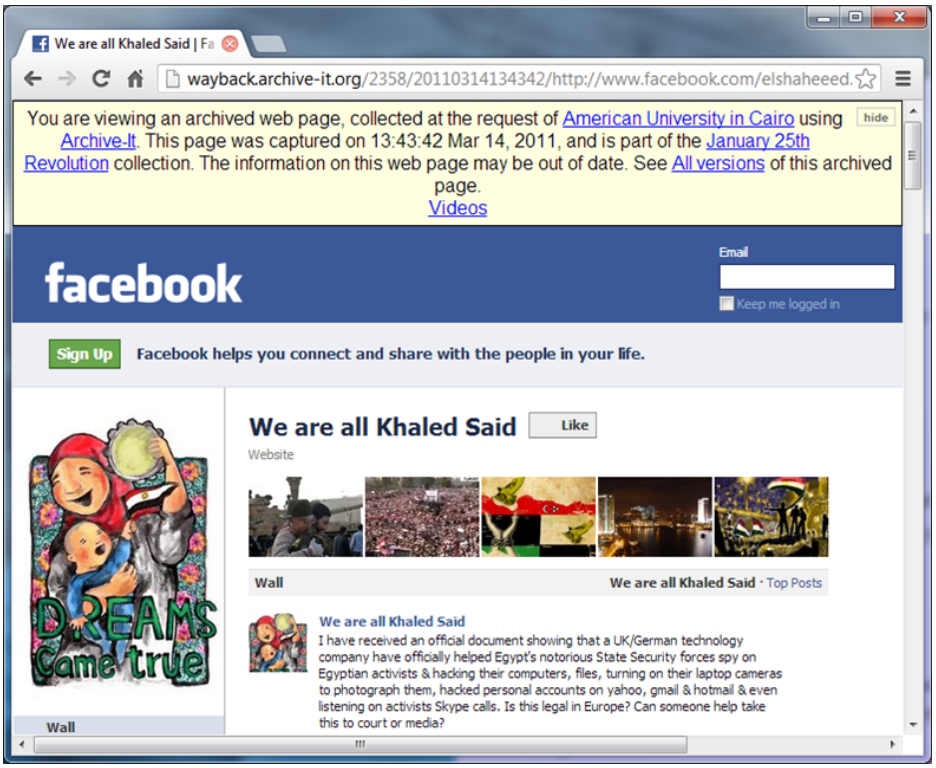


Figure 10: An example of web page datetimes problems.

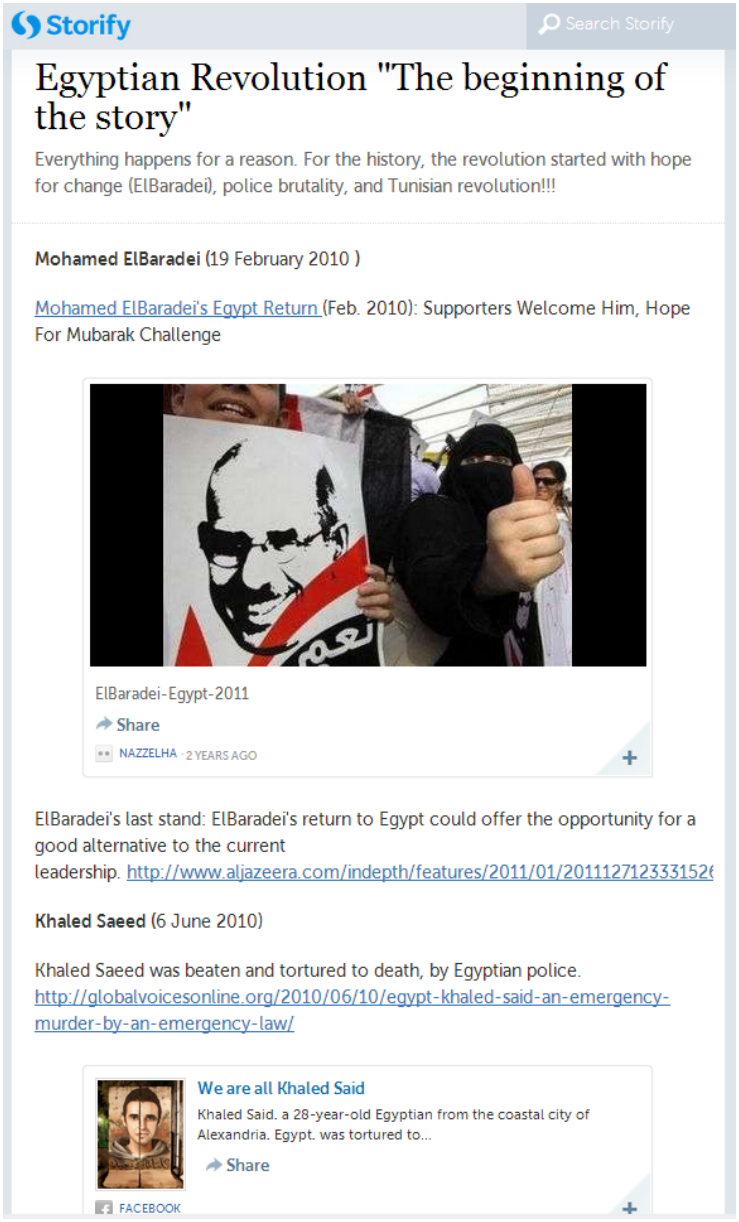


Figure 11: The Egyptian Revolution on Storify.
Source: `http://storify.com/yasmina_anwar/egyptian-revolution-the-beginning-of-the-story`

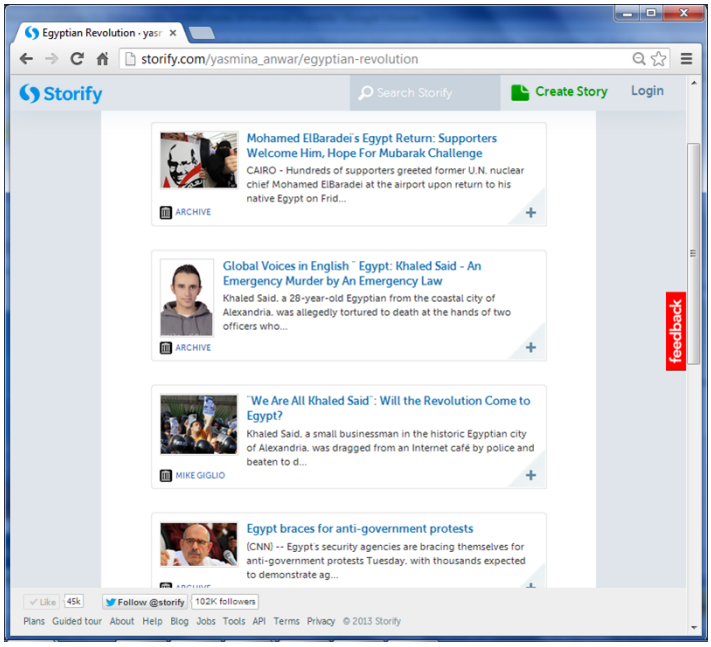


Figure 12: Mockup for the framework visualization using Storify interface.

of the story. The challenge of this step is to specify the best high quality pages that express the story. There are multiple dimensions of quality for choosing the source of the event when we have multiple candidates for the same event: web-based structural quality, such as page rank, and quality of replaying the archived page (Figure 13). We also will eliminate duplicates [35, 24, 36], such as the example in Figure 14. For evaluation, we plan to choose a small number of candidates for each event, then show all candidates and ask Mechanical Turkers if the candidates are good or not. After having the number of the best candidates for each event of the story, such as the sample in Figure 9, the next step will be creating a human-readable story and visualizing it. We will provide different interactive visualizations that enable exploring the story easily (such as the timeline visualization in Figure 8(a) and the slideshow visualization in Figure 8(b)) based on the previous work of visualizing Archive-It collections. We will also use any existing tools for visualizing the results, such as Storify in Figure 12. The user will have the ability to modify the story and specify the start and end dates. We plan to evaluate the interface by soliciting feedback from humanities researchers.

The last step in this framework is making the created story accessible for others through allowing sharing of the story with others and allowing having feedback and updating of the story by others.

The estimated timeline of the research tasks is shown in Figure 15. Naturally, this plan is expected to evolve and be modified over time.

5. CONCLUSIONS

We propose to develop a new framework for serving user needs by enriching the live web experience by automatically creating, identifying, and linking topical stories culled from the past web. Creating automatic stories related to a query from the user or web page's content is the output of this dissertation. The story will be presented to the users in which the articles will be ordered by datetime. We will integrate the data from web archives and from social media services such as Storify to create a complete story for the user around a particular event.

At the completion of this work, the user will be provided with a useful framework to help in creating stories and save huge amount of time searching for the related articles to the story and at the same time, gaining an insight about the story evolution over time. Furthermore, leveraging web archives to compensate the missing web pages is a way to be able to retain the evolution of stories over time.

6. ACKNOWLEDGMENTS

I would like to thank my advisors, Dr. Michael L. Nelson, and Dr. Michele C. Weigle for their inspiring guidance in this research and for reviewing this abstract. This work was supported in part by the NSF (IIS 1009392) and the Library of Congress. We thank Kris Carpenter Negulescu (Internet Archive) for access to the anonymized Wayback Machine logs.

7. REFERENCES

[1] I. Adams, E. L. Miller, and M. W. Storer. Analysis of Workload Behavior in Scientific and Historical Long-Term Data Repositories. Technical Report UCSC-SSRC-11-01, University of California, Santa Cruz, Mar. 2011.

[2] Y. AlNoamany, A. AlSum, M. C. Weigle, and M. L. Nelson. Who and What Links to the Internet Archive. In *Proceedings of the 17th International Conference on Theory and Practice of Digital Libraries*, TPDL '13, Sept. 2013.

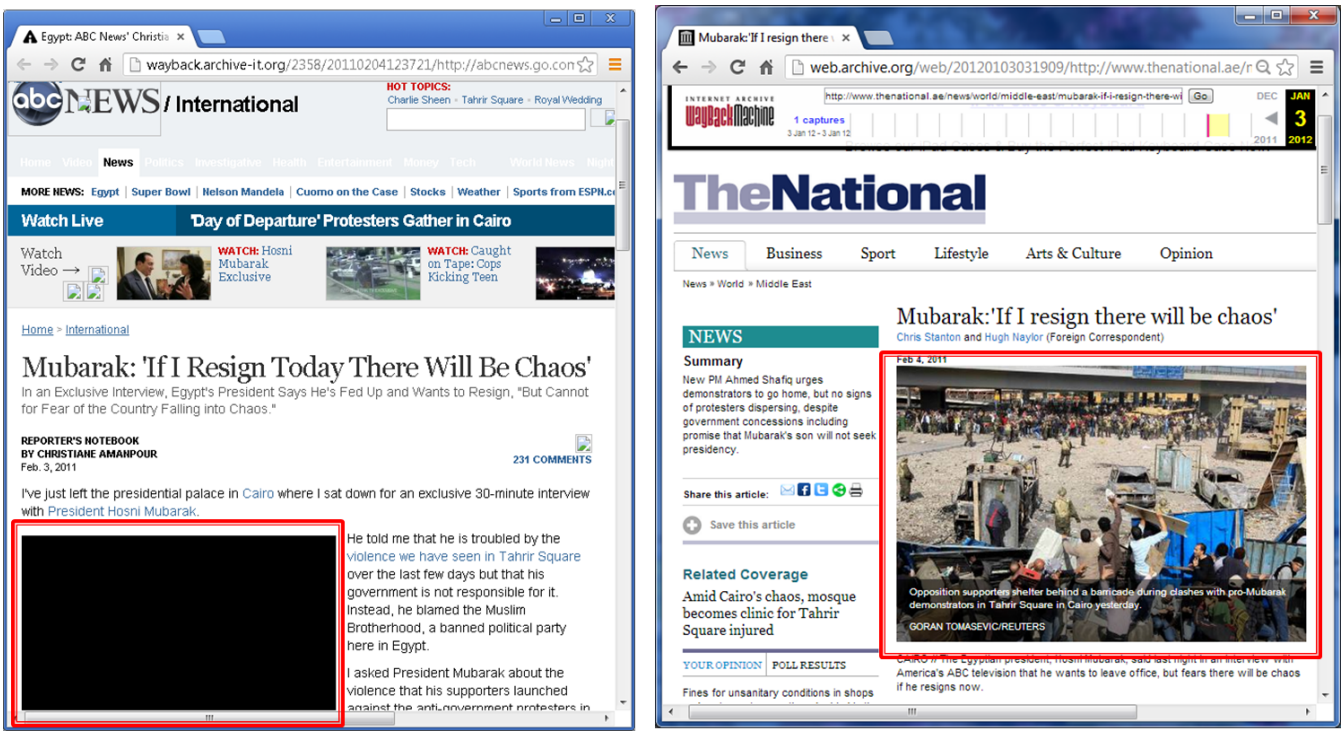


Figure 13: An example of the archive quality. The two archived pages have the same news, but the archived page on the left has a missing image, while the archived page on the right has an image.

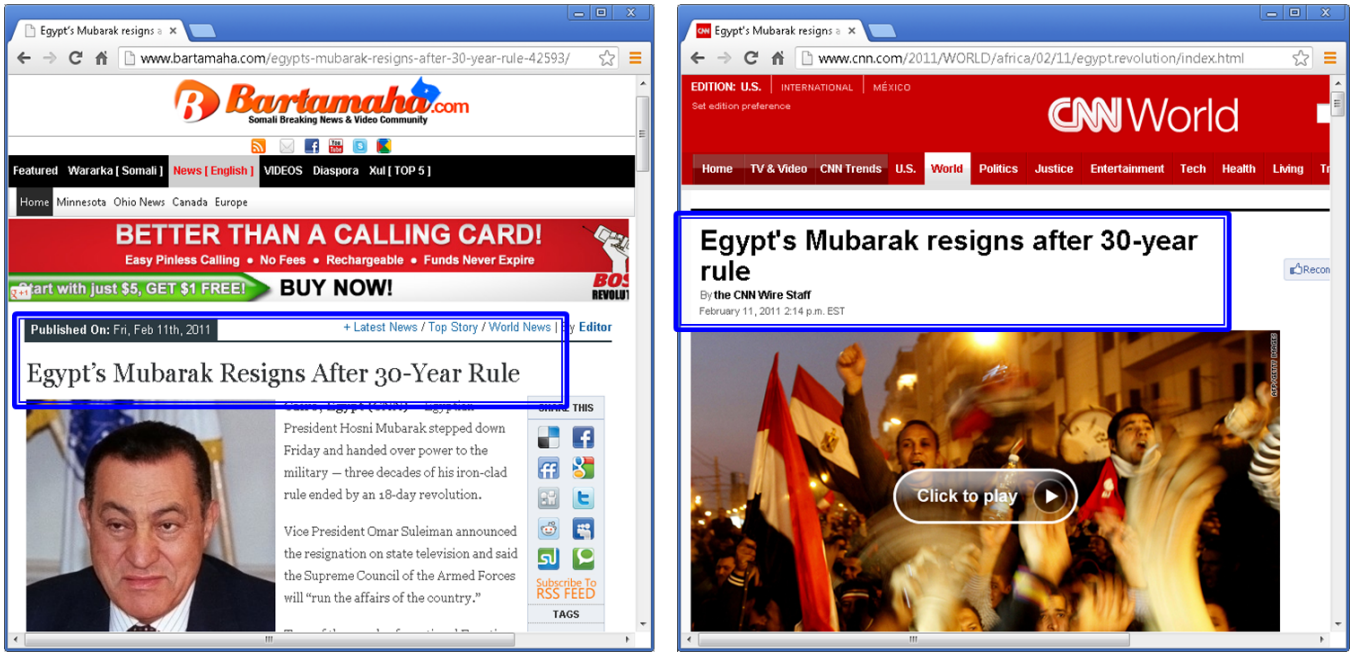


Figure 14: An example of the duplication of the same news.

Figure 15: Timeline of the dissertation work.

[3] Y. AlNoamany, M. C. Weigle, and M. L. Nelson. Access Patterns for Robots and Humans in Web Archives. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, July 2013.

[4] T. T. Aye. Web Log Cleaning for Mining of Web Usage Patterns. In *3rd International Conference on Computer Research and Development*, ICCRD, pages 490–494. IEEE, Mar. 2011.

[5] L. D. Catledge and J. E. Pitkow. Characterizing Browsing Strategies in the World-Wide Web. *Computer Networks and ISDN Systems*, 27(6):1065–1073, Apr. 1995.

[6] R. Cooley and B. Mobasher. Data Preparation for Mining World Wide Web Browsing Patterns. *Knowledge and Information Systems*, 1:5–32, 1999.

[7] M. Costa and M. J. Silva. Characterizing Search Behavior in Web Archives. In *Proceedings of Temporal Web Analytics Workshop*, TWAW, 2011.

[8] M. Costa and M. J. Silva. Understanding the Information Needs of Web Archive Users. In *Proc. of the 10th International Web Archiving Workshop*, pages 9–16, Sept 2010.

[9] M. D. Dikaiakos, A. Stassopoulou, and L. Papageorgiou. An Investigation of web crawler behavior: characterization and metrics. *Computer Communications*, 28(8):880–897, May 2005.

[10] D. Doran and S. S. Gokhale. Web Robot Detection Techniques: Overview and Limitations. *Data Mining and Knowledge Discovery*, 22(1-2):183–210, June 2010.

[11] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. Zhou. LeadLine: Interactive Visual Analysis of Text Data through Event Identifcation and Exploration. In *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, VAST '12, pages 93–102. IEEE Computer Society, 2012.

[12] M. H. Dunham. *Data Mining: Introductory and Advanced Topics*. Prentice Hall PTR, 2002.

[13] W. Guo, Y. Zhong, and J. Xie. A Web Crawler Detection Algorithm Based on Web Page Member List. In *Proceedings of 4th International Conference on Intelligent Human-Machine Systems and Cybernetics*, pages 189–192. IEEE, Aug. 2012.

[14] T. L. Harrison and M. L. Nelson. Just-In-Time Recovery of Missing Web Pages. In *Proceedings of the 17th Conference on Hypertext and Hypermedia*, HYPERTEXT '06, pages 145–156. ACM, 2006.

[15] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artificial Intelligence*, 194(0):28 – 61, 2013.

[16] A. Jatowt, K. Kanazawa, S. Oyama, and K. Tanaka. Supporting analysis of future-related information in news archives and the web. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '09, pages 115–124. ACM, 2009.

[17] A. Jatowt, Y. Kawai, S. Nakamura, Y. Kidawara, and K. Tanaka. Journey to the past: proposal of a framework for past web browser. In *Proceedings of the 17th Conference on Hypertext and Hypermedia*, HYPERTEXT '06, pages 135–144. ACM, 2006.

[18] A. Jatowt, Y. Kawai, and K. Tanaka. Detecting Age of Page Content. In *Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management*, WIDM '07, pages 137–144. ACM, 2007.

[19] R. Jones and K. L. Klinkner. Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs. In *Proceeding of the 17th ACM Conference on Information and Knowledge Mining*, CIKM '08, page 699. ACM Press, Oct. 2008.

[20] N. Kanhabua, S. Romano, and A. Stewart. Identifying Relevant Temporal Expressions for Real-World Events. In *Processidng of SIGIR 2012 Workshop on Time-aware Information Access*, TAIA '12. Microsoft Research, 2012.

[21] W. Koehler. Web page change and persistence-a four-year longitudinal study. *J. Am. Soc. Inf. Sci. Technol.*, 53(2):162–171, Jan. 2002.

[22] M. Krstajic, E. Bertini, and D. Keim. CloudLines: Compact Display of Event Episodes in Multiple Time-Series. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2432–2439, 2011.

[23] M. Krstajic, M. Najm-Araghi, F. Mansmann, and D. A. Keim. Incremental Visual Text Analytics of News Story Development. volume 8294, pages 829407–829407–12, 2012.

[24] J. P. Kumar and P. Govindarajulu. Near-Duplicate Web Page Detection: An Efficient Approach Using Clustering, Sentence Feature and Fingerprinting. *International Journal of Computational Intelligence Systems*, 6(1):1–13, 2013.

[25] R. Kumar and A. Tomkins. A Characterization of Online Browsing Behavior. In *Proceedings of the 19th International World Wide Web Conference*, WWW '10, pages 561–570. ACM, 2010.

[26] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600. ACM, 2010.

[27] S. Kwon, M. Oh, D. Kim, J. Lee, Y.-G. Kim, and S. Cha. Web Robot Detection based on Monotonous Behavior. In *Proceedings of the Information Science and Industrial Applications*, volume 4, 2012.

[28] H. Liu and V. Kešelj. Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests. *Data Knowledge Engineer*, 61(2):304–330, May 2007.

[29] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. Keim. EventRiver: Visually Exploring Text Collections with Temporal References. *IEEE Transactions on Visualization and Computer Graphics*, 18(1):93–105, Jan. 2012.

[30] B. D. A. Michael L. Nelson. Object Persistence and Availability in Digital Libraries. *D-Lib Magazine*, 2002.

[31] B. Mobasher, H. Dai, T. Luo, Y. Sun, and J. Zhu. Integrating Web Usage and Content Mining for More Effective Personalization. In *Proceedings of the First International Conference on Electronic Commerce and Web Technologies*, EC-WEB '00, pages 165–176. Springer-Verlag, 2000.

[32] K. C. Negulescu. Web Archiving @ the Internet Archive. Presentation at the 2010 Digital Preservation Partners Meeting, `http://1.usa.gov/XSjDG8`, 2010.

[33] S. Nunes, C. Ribeiro, and G. David. Using Neighbors to Date Web Documents. In *Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management*, WIDM '07, pages 129–136. ACM, 2007.

[34] K. Padia, Y. AlNoamany, and M. C. Weigle. Visualizing Digital Collections at Archive-It. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '12, pages 15–18. ACM, 2012.

[35] Y. Plegas and S. Stamou. Reducing information redundancy in search results. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 886–893. ACM, 2013.

[36] F. Radlinski, P. N. Bennett, and E. Yilmaz. Detecting duplicate web documents using clickthrough data. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 147–156. ACM, 2011.

[37] S. Rose, S. Butner, W. Cowley, M. Gregory, and J. Walker. Describing Story Evolution from Dynamic Information Streams. In *In VAST 2009: IEEE Symposium On Visual Analytics Science And Technology*, pages 99–106, 2009.

[38] H. M. SalahEldeen and M. L. Nelson. Carbon Dating The Web: Estimating the Age of Web Resources. In *Proceedings of 3rd Temporal Web Analytics Workshop*, TempWeb '13, pages 1075–1082, 2013.

[39] D. S. Sisodia and S. Verma. Web usage pattern analysis through web logs: A review. In *Proceedings of 9th International Conference on Computer Science and Software Engineering*, JCSSE, pages 49–53, May 2012.

[40] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: discovery and applications of usage patterns from Web data. *ACM SIGKDD Explorations Newsletter*, 1(2):12, Jan. 2000.

[41] J. Strötgen, O. Alonso, and M. Gertz. IdentiïňĄcation of Top Relevant Temporal Expressions in Documents. In *Proceedings of the 2nd Temporal Web Analytics Workshop*, TempWeb '12, pages 33–40. ACM, 2012.

[42] P.-N. Tan and V. Kumar. Discovery of Web Robot Sessions Based on their Navigational Patterns. *Data Mining and Knowledge Discovery*, 6(1):9–35, Jan. 2002.

[43] H. Van de Sompel, M. L. Nelson, and R. Sanderson. HTTP framework for time-based access to resource states – Memento. `https://datatracker.ietf.org/doc/draft-vandesompel-memento/`, 2012.

[44] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. L. Balakireva, S. Ainsworth, and H. Shankar. Memento: Time Travel for the Web. Technical Report arXiv:0911.1112, 2009.