USING WEB ARCHIVES TO ENRICH THE LIVE WEB EXPERIENCE THROUGH STORYTELLING

by

Yasmin AlNoamany
B.S. May 2006, Mansoura University, Egypt
M.S. August 2009, Mansoura University, Egypt

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

COMPUTER SCIENCE

OLD DOMINION UNIVERSITY
August 2016

Approved by:
Dr. Michael L. Nelson (Director)
Dr. Michele C. Weigle (Member)
Dr. Hussein Abdel-Wahab (Member)
Dr. M'Hammed Abdous (Member)

ABSTRACT

USING WEB ARCHIVES TO ENRICH THE LIVE WEB EXPERIENCE THROUGH STORYTELLING

Yasmin AlNoamany
Old Dominion University, 2016
Director: Dr. Michael L. Nelson

Much of our cultural discourse occurs primarily on the Web. Thus, Web preservation is a fundamental precondition for multiple disciplines. Archiving Web pages into themed collections is a method for ensuring these resources are available for posterity. Services such as Archive-It exists to allow institutions to develop, curate, and preserve collections of Web resources. Understanding the contents and boundaries of these archived collections is a challenge for most people, resulting in the paradox of the larger the collection, the harder it is to understand. Meanwhile, as the sheer volume of data grows on the Web, "storytelling" is becoming a popular technique in social media for selecting Web resources to support a particular narrative or "story".

In this dissertation, we address the problem of understanding the archived collections through proposing the Dark and Stormy Archive (DSA) framework, in which we integrate "storytelling" social media and Web archives. In the DSA framework, we identify, evaluate, and select candidate Web pages from archived collections that summarize the holdings of these collections, arrange them in chronological order, and then visualize these pages using tools that users already are familiar with, such as Storify.

To inform our work of generating stories from archived collections, we start by building a baseline for the structural characteristics of popular (i.e., receiving the most views) human-generated stories through investigating stories from Storify. Furthermore, we checked the entire population of Archive-It collections for better understanding the characteristics of the collections we intend to summarize. We then filter off-topic pages from the collections the using different methods to detect when an archived page in a collection has gone off-topic. We created a gold standard dataset from three Archive-It collections to evaluate the proposed methods at different thresholds. From the gold standard dataset, we identified five behaviors for the TimeMaps (a list of archived copies of a page) based on the page's aboutness. Based on a dynamic slicing algorithm, we divide the collection and cluster the pages in each slice. We then select the best representative page from each cluster based on different quality metrics (e.g., the replay quality, and the quality of the generated snippet

from the page). At the end, we put the selected pages in chronological order and visualize them using Storify.

For evaluating the DSA framework, we obtained a ground truth dataset of hand-crafted stories from Archive-It collections generated by expert archivists. We used Amazon's Mechanical Turk to evaluate the automatically generated stories against the stories that were created by domain experts. The results show that the automatically generated stories by the DSA are indistinguishable from those created by human subject domain experts, while at the same time both kinds of stories (automatic and human) are easily distinguished from randomly generated stories.

To my husband Ahmed, my son Yousof, and the martyrs who sacrificed their life for freedom, justice and dignity in Egypt.

ACKNOWLEDGMENTS

First and above all, praise be to God (Alhamdulelah), the almighty for giving me the opportunity to finish what I started. I hope the science introduced in this dissertation glorifies him and benefits the people.

I would like to sincerely thank my advisor Dr. Michael L. Nelson, not only for his superb guidance, but also for his time, patience, and support during my Ph.D. journey. Discussions with Dr. Nelson always sparked interesting and great ideas and allowed me to increase my knowledge in various aspects. Not only my presentation skills and performance in research improved because of Dr. Nelson, but I also became interested in old cars and learned that cats can live with dogs peacefully!

I am grateful to my dissertation committee, Dr. Michele C. Weigle, Dr. Hussein Abdel-Wahab, and Dr. M'hammad Abdous for their support and for their input to enhance the dissertation. I cannot thank enough Dr. Michele C. Weigle whose support, guidance, and encouragement were always there throughout my time at Old Dominion University. Dr. Weigle's valuable comments helped in achieving high-quality standards in my research and her support strengthened me in hard times when I was down and wanted to quit. A special appreciation to Dr. Hussein Abdel-Wahab for his friendship, encouragement, and support in the whole journey. Dr. Abdel-Wahab is always willing to listen, help, and answer questions.

I would like to acknowledge Kristine Hanna, Jefferson Bailey and the Archive-It team and partners for supporting my research by providing us with the datasets and also for helping in the evaluation phases. I'm grateful to the former and current members of the Web Science and Digital Library group. Special appreciations for Lulwah Alkwai, my best friend and teammate, for her support, encouragement, and help. I would like to thank Justin F. Brunelle for answering my questions about the Ph.D. process and for being a supportive friend. I would like also to thank him for allowing me to meet the youngest member of Brunelle's family, Brayden. I would like to acknowledge Sawood, Mohamed, and Alex for their help and Mat, Shawn, Chuck, Scott, Martin, and Hany for the useful discussions in the lab.

I'm indebted to have a supportive family (my mom, dad, and my brothers). Without their prayers, I would not accomplish this. I am grateful for Yousof, my son and the joy of my life. Currently, Yousof is 7 years old. I will not forget how much he was gentle and understanding at this young age and also how much he prayed for me to finish, especially in the last year. I would like to thank Ahmed Hesham and Sarra Elgammal, our second family, for their help and support during this journey. I'm blessed to have many thoughtful and supportive friends from all over the world who supported me these past several years:

Nermeen, Tanlaya, Tasneem, Marwa, Lamia, Manar, Hana, Hend, Amira, Soad, Dalia, Eman, Ghada, Dalia, Doaa, Maha, Ayat, Omnia, Azza, Ahmed, May, Mona, Doha, Heba, Reem, Amr, Ethar, Walaa, Adrian, Yasmin, Ingie, Taghrid, Khadija, Waliyya. Their uphold and encouragement cheered me on through the hard times.

Finally, words are not enough to thank the most patient, caring, and encouraging man ever, my husband, Ahmed. I owe Ahmed a lot, he is my everything.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Since it was invented approximately 25 years ago, the Web has developed significantly and new research methodologies have evolved. Moreover, the beginning of Web 2.0 in early 2000 allowed users to contribute born-digital materials to the Web including images, videos, geo-locations, and text. With the emergence of Web 2.0, digital materials have become part of our cultural heritage and preserving the resources of the Web has become essential to facilitate research in history, sociology, political science, media, literature, and other related disciplines. For many, social media has become the primary resource when an important event occurs [195], and it also can provide the initial spark for important stories (for example, the initial events of the Egyptian Revolution occurred on Facebook [308, 172, 128]).

With the extensive growth of the Web, multiple Web archiving initiatives have been started to archive different aspects of the Web [37]. This was followed by emerging practices and technologies from the archiving communities. For example, the Internet Archive[1] (IA), which has been archiving the Web since 1996, generated standards, tools, and technologies to capture Web pages and replay them (e.g., the Wayback Machine [238]). Several universities built their own Web archives for research purposes (e.g., the Stanford WebBase Archive) [73].

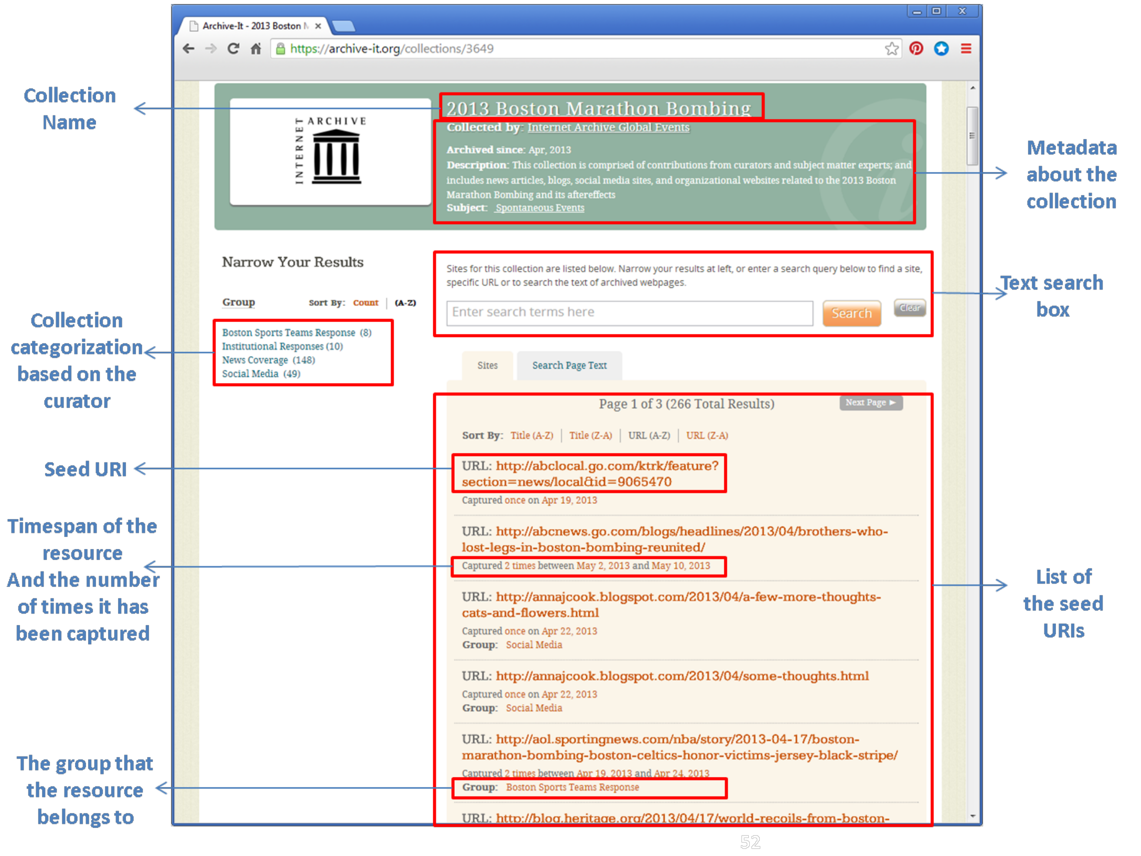
Additionally, multiple archiving initiatives exist to allow people to archive Web resources into themed collections to ensure these resources are available for posterity. For example, Archive-It[2], a subscription service from the IA, allows institutions to develop, curate, and preserve topic-oriented collections of Web resources by specifying a set of seeds, Uniform Resource Identifiers (URIs), that should be crawled periodically. Archive-It provides a listing of all seeds in the collection along with the number of times and dates over which each page was archived, as well as a full-text search of archived pages (Figure 1). Although Archive-It provides users with tools for searching and browsing collections, it is not easy for users to understand the essence of these collections [252].

With the user in the loop, we develop the Dark and Stormy Archive (DSA) framework, which automatically extracts summary stories[3] from Archive-It collections to help the user

[1] http://archive.org/

[2] http://www.archive-it.org/

[3] We use “story” in its current, loose context of social media, which is sometimes missing elements from the more formal literary tradition of dramatic structure, morality, humor, improvisation, etc.


Archive-It - 2013 Boston M
https://archive-it.org/collections/3649
INTERNET ARCHIVE
Collection Name
Metadata about the collection
Text search box
Collection categorization based on the curator
Seed URI
Timespan of the resource And the number of times it has been captured
List of the seed URIs
The group that the resource belongs to
Search
Clear
Sites
Search Page Text
Next Page ►

2013 Boston Marathon Bombing

Collected by: Internet Archive Global Events

Archived since: Apr, 2013

Description: This collection is comprised of contributions from curators and subject matter experts; and includes news articles, blogs, social media sites, and organizational websites related to the 2013 Boston Marathon Bombing and its aftereffects

Subject: Spontaneous Events

Narrow Your Results

Group Sort By: Count | (A-Z)

Boston Sports Teams Response (8)
Institutional Responses (10)
News Coverage (148)
Social Media (49)

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here

Page 1 of 3 (266 Total Results)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

URL: http://abclocal.go.com/ktrk/feature?section=news/local&id=9065470
Captured once on Apr 19, 2013

URL: http://abcnews.go.com/blogs/headlines/2013/04/brothers-who-lost-legs-in-boston-bombing-reunited/
Captured 2 times between May 2, 2013 and May 10, 2013

URL: http://annajcook.blogspot.com/2013/04/a-few-more-thoughts-cats-and-flowers.html
Captured once on Apr 22, 2013
Group: Social Media

URL: http://annajcook.blogspot.com/2013/04/some-thoughts.html
Captured once on Apr 22, 2013
Group: Social Media

URL: http://aol.sportingnews.com/nba/story/2013-04-17/boston-marathon-bombing-boston-celtics-honor-victims-jersey-black-stripe/
Captured 2 times between Apr 19, 2013 and Apr 24, 2013
Group: Boston Sports Teams Response

URL: http://blog.heritage.org/2013/04/17/world-recoils-from-boston-

FIG. 1: The Archive-It interface of the 2013 Boston Marathon Bombing collection is a list of URIs that are ordered alphabetically.

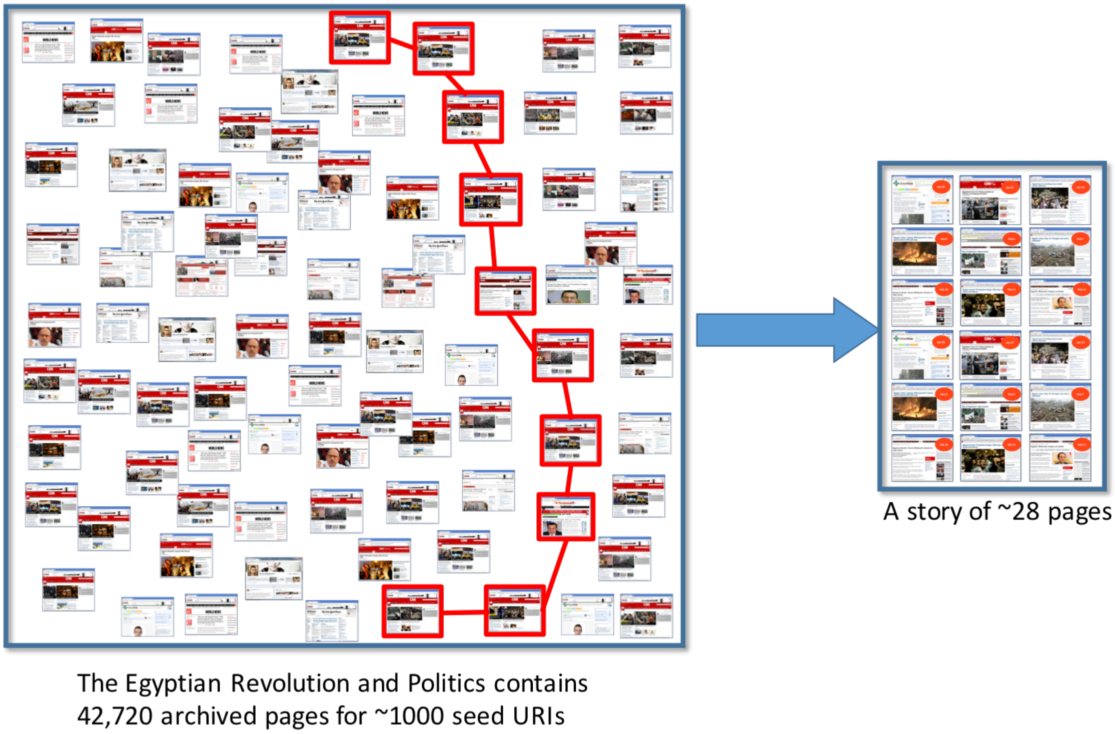

The Egyptian Revolution and Politics contains
42,720 archived pages for ~1000 seed URIs
A story of ~28 pages

FIG. 2: We derive a story S from the collection C ($C \rightarrow S$). For example, the "Egyptian Revolution and Politics" collection at Archive-It contains more than 1000 URIs in which they have 42,720 archived copies. We will automatically generate a story of $k \approx 28$ archived pages that best represent the collection.

to understand the collections. Figure 2 shows an example of deriving a story S from a collection C, represented as $C \rightarrow S$. From all of the pages in a collection, we choose different sets of k archived pages to create summary stories that give different perspectives about the collection. Then we push those chosen pages to Storify. We also help to improve the collection's quality by detecting the non-relevant pages to the topic of a collection.

We start by discussing the importance of the archived collections and their societal impact for understanding world events (Section 1.1). Then, we explain the problem of collection understanding and issues with seed URIs using examples from Archive-It collections (Section 1.2). We demonstrate the importance of replaying stories from the past through multiple stories about two important events: the Egyptian Revolution and the Boston Marathon Bombing (Section 1.3). The research questions we address in this dissertation are overviewed in Section 1.4 and the roadmap of the dissertation is detailed in Section 1.5.

1.1 ARCHIVED COLLECTIONS ARE IMPORTANT FOR POSTERITY

> Most societies place importance on preserving artifacts of their culture and heritage. Without such artifacts, civilization has no memory and no mechanism to learn from its successes and failures. Our culture now produces more and more artifacts in digital form. The Archive's mission is to help preserve those artifacts and create an Internet library for researchers, historians, and scholars.
> — The Internet Archive's mission statement [9].

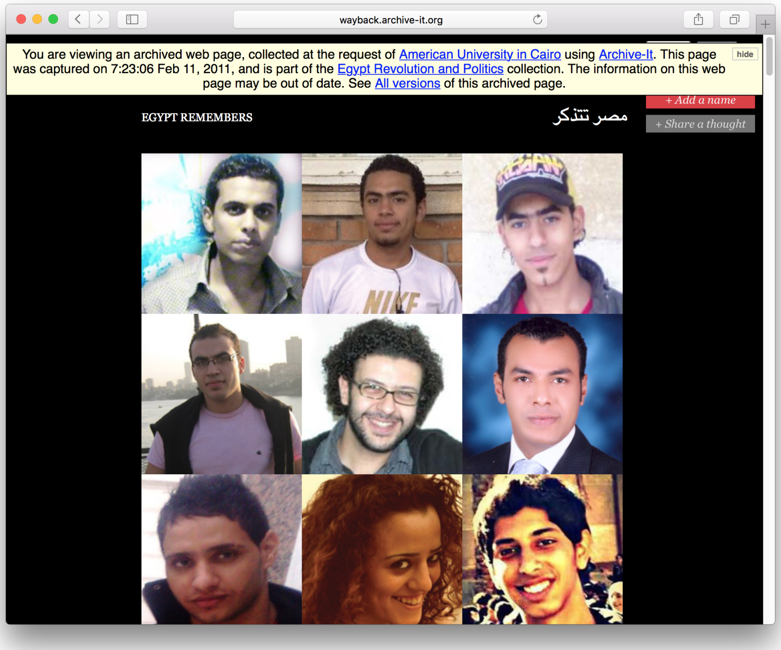
Because the Web is a dynamic information space, the resources may change, disappear, and frequently move from one location to another [183, 177]. Many studies have shown that the expected lifetime of a Web page is short (between 44 to 190 days) [222, 201, 162, 49] and that Web resources disappear quickly [277, 180, 182]. This could be for various reasons such as service discontinuance, deliberate deletion by authors or system administrators, death, removing information that was publicly known at a certain time and preventing third parties to access this information, etc. An example is the "the right to be forgotten" movement by the European Court of Justice forcing search engines like Google to remove links of specific Web sites [269]. Jones et al. [158] claimed that Google received 239,337 requests to eliminate 867,930 URLs from search results and has removed 305,095 URLs as of April 2015.

Much of our cultural discourse occurs primarily on the Web and its preservation is a fundamental precondition for research in history, sociology, political science, media, literature, and other related disciplines [240]. There are multiple examples where Web sites dedicated to documenting important events have been lost. However, some of these Web sites were partially archived. For example `sonicmemorial.com` was constructed to be an archive of digital memorials and shared media from 9/11 [77], but that site itself has since been lost and is only partially archived[4].

The conversation around major revolutionary events, such as the Arab Spring, started on the Web, specifically Twitter and Facebook [308, 172, 128]. Several Web sites, blogs, Storify entries, and channels on YouTube were created by the public, not historians, utilizing the tools of Web 2.0 to document the Jan. 25 Egyptian Revolution. Several digital libraries of thousands of articles, posts, images, videos, etc. resulted from collecting these resources to save the history of the revolution for the future generations. For example, `1000memories.com/egypt` was an online memorial for the martyrs of the 18 January protest to free Egypt. An example of an archived page[5] of `1000memories.com/egypt` is

[4] It became spam in 2006: `http://wayback.archive.org/web/*/http://www.sonicmemorial.com/`

[5] `http://wayback.archive-it.org/2358/20110211072306/http://1000memories.com/egypt`



You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 7:23:06 Feb 11, 2011, and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page.

(a) The people who were killed during the Jan. 25 Egyptian Revolution.

```
curl -I http://1000memories.com/egypt
HTTP/1.1 404 Not Found
Content-Length: 177
Content-Type: text/html
Date: Mon, 09 March 2016 17:11:10 GMT
Server: nginx/1.4.6 (Ubuntu)
Connection: keep-alive
```

(b) The HTTP response headers for the 1000memories.com/egypt as of March 2016.

FIG. 3: 1000memories.com is not available now on the live Web as of March 2016.

shown in Figure 3(a). The 1000memories site contained a digital collection of around 403 photos with information about the lives of the martyrs [275]. The entire Web site is unavailable now (see Figure 3(b)). Fortunately, `1000memories.com/egypt` has multiple copies in the "Egypt Revolution and Politics"[6] collection in Archive-It (Figure 3(a)).

Other examples are `iamtahrir.com`, which contained the artwork produced during the Egyptian Revolution, and `iamjan25.com`, which contained about 3,525 images and 2,387 videos posted by people about the January demonstrations [276]. Both sites were created for collecting content related to the Egyptian Revolution. The two repositories were lost from the live Web[7], but luckily there are multiple copies of the two repositories in Archive-It[8,9].

Additionally, storytelling services such as Storify have been used widely during the Egyptian Revolution to craft digital narratives in real time by curating social media content (e.g., Facebook, Twitter, and other Web resources). For example, a Storify story related to the Egyptian revolution is shown in Figure 4(a). Figure 4(b) shows an embedded link in the story in Figure 4(a) that is no longer available on the live Web. The reader of this story will not be able to get an idea about the content of the missing resources, especially if text around the link does not provide enough context.

The Jan. 25 Egyptian Revolution is one of the most important events that has happened in recent history. Several books and initiatives have been published for documenting the 18 days of the Egyptian Revolution [271, 185, 95, 288]. Furthermore, an enormous number of studies [256, 110, 344, 136] have been conducted for studying the Arab Spring and, specifically, the Egyptian Revolution. These books and studies cited the digital collections that we mentioned earlier in this section and other sites that were dedicated to document the Egyptian Revolution (e.g., `25Leaks.com`). Unfortunately, the links to many of these Web sites are now broken and there is no way (without the archive) to know what they contained.

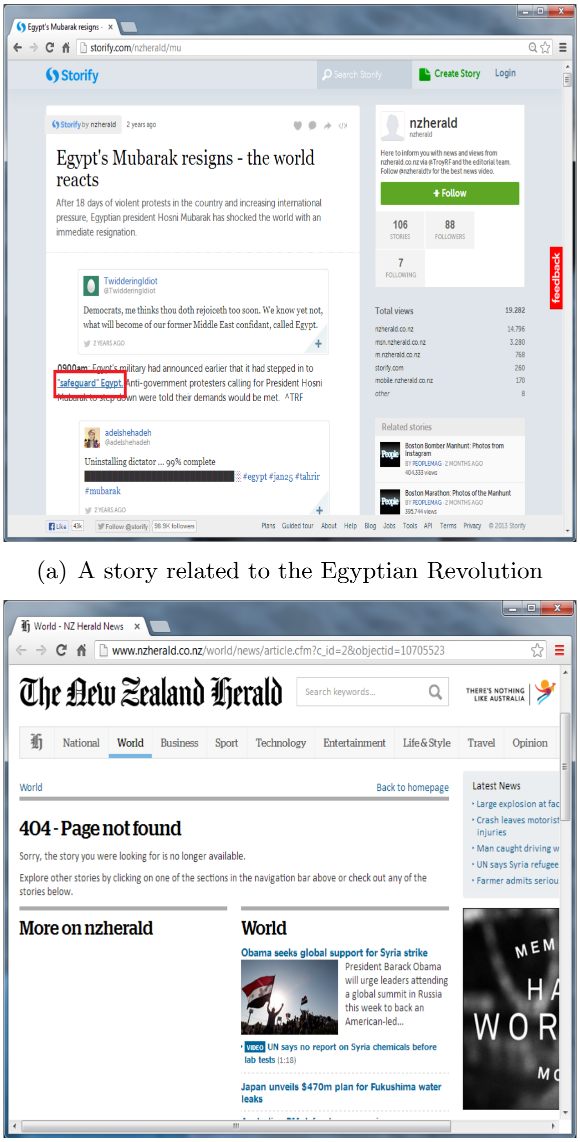
Today's ordinary information will be tomorrow's resources for historical research. The content captured and published on the Web narrating the incidents and giving unfiltered insights for future generations and historians is important to clarify the exact turning points in history. Therefore, archiving Web pages into themed collections is a method for ensuring these resources are available for posterity. Happily, the Web sites mentioned in the previous examples were captured as a part of the Egyptian Revolution collection in Archive-It.

[6] `https://archive-it.org/collections/2358/`

[7] As of March 2016, they are 404 Not Found

[8] `http://wayback.archive-it.org/2358/*/http://www.iamtahrir.com/`

[9] `http://wayback.archive-it.org/2358/*/http://iamjan25.com/`


Egypt's Mubarak resigns -
storify.com/nzherald/mu
Storify
Search Storify
Create Story
Login
Storify by nzherald
2 years ago
+ Follow
feedback
Like 43k
Follow @storify 98.9K followers
Plans Guided tour About Help Blog Jobs Tools API Terms Privacy © 2013 Storify

Egypt's Mubarak resigns - the world reacts

After 18 days of violent protests in the country and increasing international pressure, Egyptian president Hosni Mubarak has shocked the world with an immediate resignation.

TwidderingIdiot
@TwidderingIdiot

Democrats, me thinks thou doth rejoiceth too soon. We know yet not, what will become of our former Middle East confidant, called Egypt.

2 YEARS AGO

0900am: Egypt's military had announced earlier that it had stepped in to "safeguard" Egypt. Anti-government protesters calling for President Hosni Mubarak to step down were told their demands would be met. ^TRF

adelshehadeh
@adelshehadeh

Uninstalling dictator ... 99% complete #egypt #jan25 #tahrir #mubarak

2 YEARS AGO

nzherald
nzherald

Here to inform you with news and views from nzherald.co.nz via @TroyRF and the editorial team. Follow @nzheraldtv for the best news video.

106	88
STORIES	FOLLOWERS

7
FOLLOWING

Total views	19.282
nzherald.co.nz	14.796
msn.nzherald.co.nz	3.280
m.nzherald.co.nz	768
storify.com	260
mobile.nzherald.co.nz	170
other	8

Related stories

Boston Bomber Manhunt: Photos from Instagram
BY PEOPLEMAG · 2 MONTHS AGO
404.333 views

Boston Marathon: Photos of the Manhunt
BY PEOPLEMAG · 2 MONTHS AGO
395.744 views

(a) A story related to the Egyptian Revolution

404 - Page not found

Sorry, the story you were looking for is no longer available.

Explore other stories by clicking on one of the sections in the navigation bar above or check out any of the stories below.

More on nzherald

World

Obama seeks global support for Syria strike

President Barack Obama will urge leaders attending a global summit in Russia this week to back an American-led...

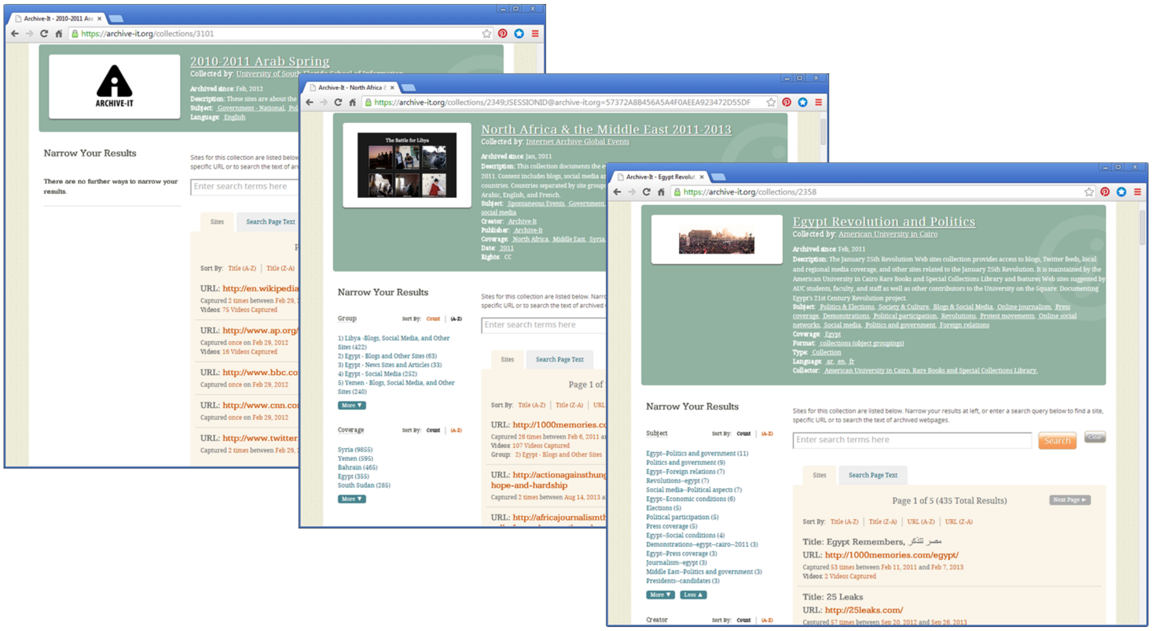
VIDEO UN says no report on Syria chemicals before lab tests (1:18)

Japan unveils $470m plan for Fukushima water leaks

Latest News
- Large explosion at fac
- Crash leaves motorist injuries
- Man caught driving w
- UN says Syria refugee
- Farmer admits seriou

(b) The bookmarked link is broken

FIG. 4: Storify is for bookmarking, not for preserving. When the annotated link (on the top) is requested, it results in a 404 (on the bottom).


Archive-It - 2010-2011 Ara
https://archive-it.org/collections/3101

2010-2011 Arab Spring
Collected by: University of South
Archived since: Feb, 2012
Description: These sites are about the
Subject: Government - National, Pol
Language: English

Narrow Your Results

There are no further ways to narrow your results.

Sites for this collection are listed below
specific URL or to search the text of ar

Enter search terms here

Sites | Search Page Text

Sort By: Title (A-Z) | Title (Z-A)

URL: http://en.wikipedia
Captured 2 times between Feb 29,
Videos: 75 Videos Captured

URL: http://www.ap.org/
Captured once on Feb 29, 2012
Videos: 16 Videos Captured

URL: http://www.bbc.co
Captured once on Feb 29, 2012

URL: http://www.cnn.co
Captured once on Feb 29, 2012

URL: http://www.twitter.
Captured 2 times between Feb 29,

Archive-It - North Africa &
https://archive-it.org/collections/2349;JSESSIONID@archive-it.org=57372A8B456A5A4F0AEEA923472D55DF

North Africa & the Middle East 2011-2013
Collected by: Internet Archive Global Events
Archived since: Jan, 2011
Description: This collection documents the e
2011. Content includes blogs, social media ar
countries. Countries separated by site group
Arabic, English, and French.
Subject: Spontaneous Events, Government,
social media
Creator: Archive-It
Publisher: Archive-It
Coverage: North Africa, Middle East, Syria,
Date: 2011
Rights: CC

Narrow Your Results

Group — Sort By: Count | (A-Z)

1) Libya -Blogs, Social Media, and Other Sites (422)
2) Egypt - Blogs and Other Sites (63)
3) Egypt - News Sites and Articles (33)
4) Egypt - Social Media (252)
5) Yemen - Blogs, Social Media, and Other Sites (240)

More ▼

Coverage — Sort By: Count | (A-Z)

Syria (9855)
Yemen (595)
Bahrain (465)
Egypt (355)
South Sudan (285)

More ▼

Sites for this collection are listed below. Narr
specific URL or to search the text of archived

Enter search terms here

Sites | Search Page Text

Page 1 of

Sort By: Title (A-Z) | Title (Z-A) | URL

URL: http://1000memories.co
Captured 26 times between Feb 6, 2011 a
Videos: 107 Videos Captured
Group: 2) Egypt - Blogs and Other Sites

URL: http://actionagainsthung
hope-and-hardship
Captured 2 times between Aug 14, 2013 a

URL: http://africajournalismth

Archive-It - Egypt Revolut
https://archive-it.org/collections/2358

Egypt Revolution and Politics
Collected by: American University in Cairo
Archived since: Feb, 2011
Description: The January 25th Revolution Web sites collection provides access to blogs, Twitter feeds, local and regional media coverage, and other sites related to the January 25th Revolution. It is maintained by the American University in Cairo Rare Books and Special Collections Library and features Web sites suggested by AUC students, faculty, and staff as well as other contributors to the University on the Square: Documenting Egypt's 21st Century Revolution project.
Subject: Politics & Elections, Society & Culture, Blogs & Social Media, Online journalism, Press coverage, Demonstrations, Political participation, Revolutions, Protest movements, Online social networks, Social media, Politics and government, Foreign relations
Coverage: Egypt
Format: collections (object groupings)
Type: Collection
Language: ar, en, fr
Collector: American University in Cairo, Rare Books and Special Collections Library.

Narrow Your Results

Subject — Sort By: Count | (A-Z)

Egypt--Politics and government (11)
Politics and government (9)
Egypt--Foreign relations (7)
Revolutions--egypt (7)
Social media--Political aspects (7)
Egypt--Economic conditions (6)
Elections (5)
Political participation (5)
Press coverage (5)
Egypt--Social conditions (4)
Demonstrations--egypt--cairo--2011 (3)
Egypt--Press coverage (3)
Journalism--egypt (3)
Middle East--Politics and government (3)
Presidents--candidates (3)

More ▼ Less ▲

Creator — Sort By: Count | (A-Z)

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here [Search] [Clear]

Sites | Search Page Text

Page 1 of 5 (435 Total Results) [Next Page ►]

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: Egypt Remembers, مصر تتذكر
URL: http://1000memories.com/egypt/
Captured 53 times between Feb 11, 2011 and Feb 7, 2013
Videos: 2 Videos Captured

Title: 25 Leaks
URL: http://25leaks.com/
Captured 57 times between Sep 20, 2012 and Sep 26, 2013

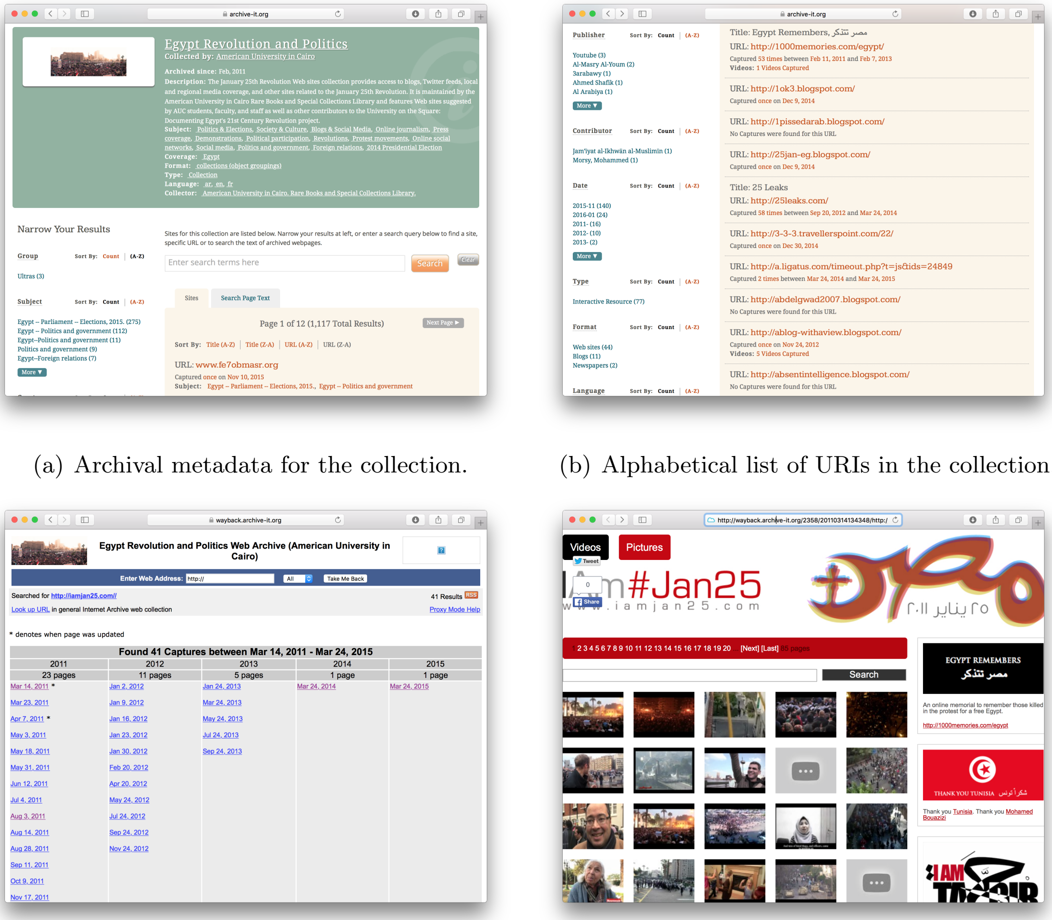
FIG. 5: There are multiple collections in Archive-It about the Jan. 25 Egyptian Revolution.

1.2 PROBLEM STATEMENT

In this section, we demonstrate the limitations of understanding archived collections and the current issues with seed URIs.

1.2.1 COLLECTION UNDERSTANDING

I want my son, who is now 7 years old, to know what happened during the Jan. 25 Egyptian Revolution as I saw it happening on the Web. Let us assume that he knows about the archived collections that are devoted of archiving important events, such as those at Archive-It. He will use the current browsing interface of `archive-it.org` to look for collections related to the Egypt Revolution. If he uses the searching and browsing tools that Archive-It provides, he will find about four or five collections containing information about the Jan. 25 Egyptian Revolution (Figure 5). Aside from the brief metadata about the collection (Figure 6(a)), the interface mainly consists of a list of seed URIs in alphabetical order (Figure 6(b)), and for each of these URIs a list of the times when the page was archived (Figure 7(c)). It is not feasible for him to figure out what is inside the collection without going through all the URIs in the collection and their relative archived copies. Understanding the essence of the collection from the current interface of Archive-It is not easy.



archive-it.org

Egypt Revolution and Politics
Collected by: American University in Cairo

Archived since: Feb, 2011
Description: The January 25th Revolution Web sites collection provides access to blogs, Twitter feeds, local and regional media coverage, and other sites related to the January 25th Revolution. It is maintained by the American University in Cairo Rare Books and Special Collections Library and features Web sites suggested by AUC students, faculty, and staff as well as other contributors to the University on the Square: Documenting Egypt's 21st Century Revolution project.
Subject: Politics & Elections, Society & Culture, Blogs & Social Media, Online journalism, Press coverage, Demonstrations, Political participation, Revolutions, Protest movements, Online social networks, Social media, Politics and government, Foreign relations, 2014 Presidential Election
Coverage: Egypt
Format: collections (object groupings)
Type: Collection
Language: ar, en, fr
Collector: American University in Cairo. Rare Books and Special Collections Library.

Narrow Your Results

Group Sort By: Count | (A-Z)
Ultras (3)

Subject Sort By: Count | (A-Z)
Egypt -- Parliament -- Elections, 2015. (275)
Egypt -- Politics and government (112)
Egypt--Politics and government (11)
Politics and government (9)
Egypt--Foreign relations (7)

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Page 1 of 12 (1,117 Total Results)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

URL: www.fe7obmasr.org
Captured once on Nov 10, 2015
Subject: Egypt -- Parliament -- Elections, 2015., Egypt -- Politics and government

(a) Archival metadata for the collection.

archive-it.org

Publisher Sort By: Count | (A-Z)
Youtube (3)
Al-Masry Al-Youm (2)
3arabawy (1)
Ahmed Shafik (1)
Al Arabiya (1)

Contributor Sort By: Count | (A-Z)
Jamʻīyat al-Ikhwān al-Muslimīn (1)
Morsy, Mohammed (1)

Date Sort By: Count | (A-Z)
2015-11 (140)
2016-01 (24)
2011- (16)
2012- (10)
2013- (2)

Type Sort By: Count | (A-Z)
Interactive Resource (77)

Format Sort By: Count | (A-Z)
Web sites (44)
Blogs (11)
Newspapers (2)

Language Sort By: Count | (A-Z)

Title: Egypt Remembers, مصر تتذكر
URL: http://1000memories.com/egypt/
Captured 53 times between Feb 11, 2011 and Feb 7, 2013
Videos: 1 Videos Captured

URL: http://1ok3.blogspot.com/
Captured once on Dec 9, 2014

URL: http://1pissedarab.blogspot.com/
No Captures were found for this URL

URL: http://25jan-eg.blogspot.com/
Captured once on Dec 9, 2014

Title: 25 Leaks
URL: http://25leaks.com/
Captured 58 times between Sep 20, 2012 and Mar 24, 2014

URL: http://3-3-3.travellerspoint.com/22/
Captured once on Dec 30, 2014

URL: http://a.ligatus.com/timeout.php?t=js&ids=24849
Captured 2 times between Mar 24, 2014 and Mar 24, 2015

URL: http://abdelgwad2007.blogspot.com/
No Captures were found for this URL

URL: http://ablog-withaview.blogspot.com/
Captured once on Nov 24, 2012
Videos: 5 Videos Captured

URL: http://absentintelligence.blogspot.com/
No Captures were found for this URL

(b) Alphabetical list of URIs in the collection.

wayback.archive-it.org

Egypt Revolution and Politics Web Archive (American University in Cairo)

Enter Web Address: http:// All Take Me Back

Searched for http://iamjan25.com// 41 Results RSS

Look up URL in general Internet Archive web collection Proxy Mode Help

* denotes when page was updated

Found 41 Captures between Mar 14, 2011 - Mar 24, 2015

2011	2012	2013	2014	2015
23 pages	11 pages	5 pages	1 page	1 page
Mar 14, 2011 *	Jan 2, 2012	Jan 24, 2013	Mar 24, 2014	Mar 24, 2015
Mar 23, 2011	Jan 9, 2012	Mar 24, 2013		
Apr 7, 2011 *	Jan 16, 2012	May 24, 2013		
May 3, 2011	Jan 23, 2012	Jul 24, 2013		
May 18, 2011	Jan 30, 2012	Sep 24, 2013		
May 31, 2011	Feb 20, 2012			
Jun 12, 2011	Apr 20, 2012			
Jul 4, 2011	May 24, 2012			
Aug 3, 2011	Jul 24, 2012			
Aug 14, 2011	Sep 24, 2012			
Aug 28, 2011	Nov 24, 2012			
Sep 11, 2011				
Oct 9, 2011				
Nov 17, 2011				

(c) Archived copies of a URI in the collection.

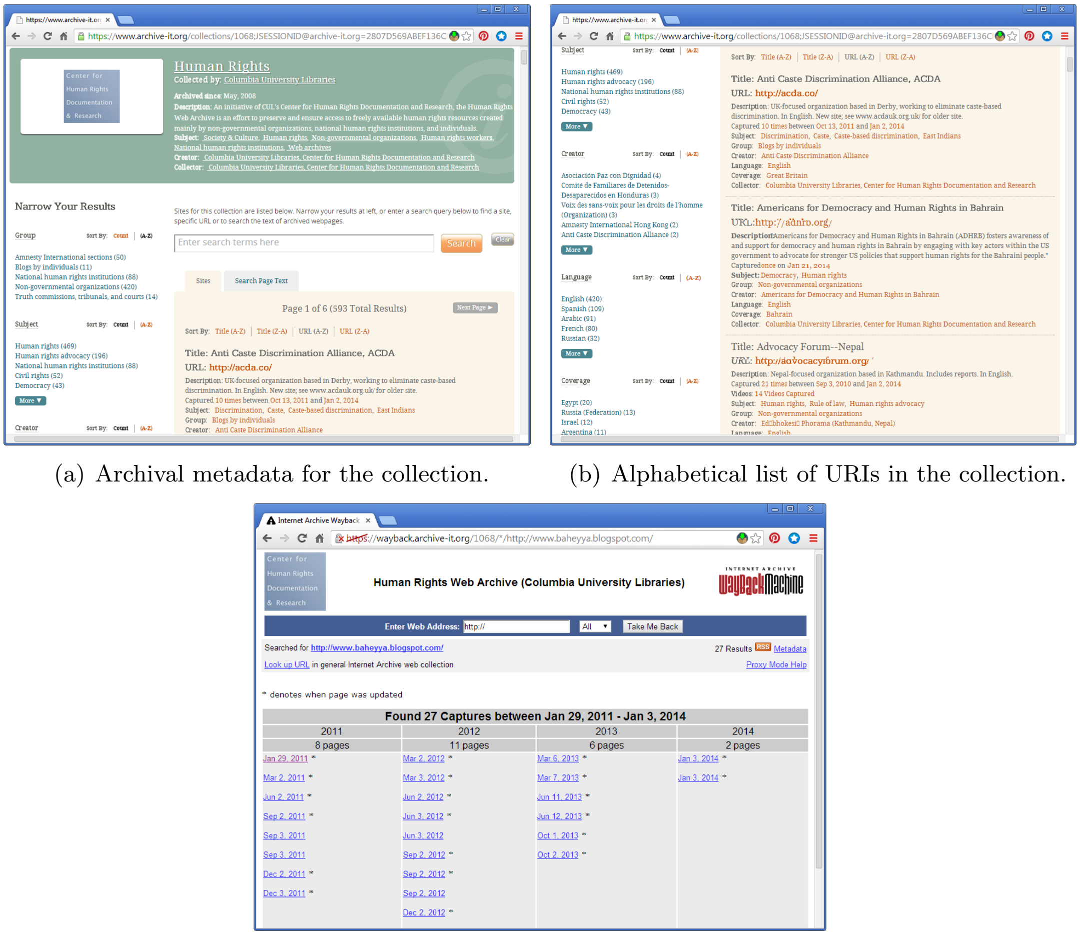
http://wayback.archive-it.org/2358/20110314134348/http:/

An online memorial to remember those killed in the protest for a free Egypt.
http://1000memories.com/egypt

Thank you Tunisia. Thank you Mohamed Bouazizi

(d) A copy of "Iam25Jan"

FIG. 6: Current browsing and searching services for the "Egypt Revolution and Politics" collection in Archive-It.



https://www.archive-it.org/collections/1068;JSESSIONID@archive-it.org=2807D569ABEF136C

Center for Human Rights Documentation & Research

Human Rights
Collected by: Columbia University Libraries

Archived since: May, 2008
Description: An initiative of CUL's Center for Human Rights Documentation and Research, the Human Rights Web Archive is an effort to preserve and ensure access to freely available human rights resources created mainly by non-governmental organizations, national human rights institutions, and individuals.
Subject: Society & Culture, Human rights, Non-governmental organizations, Human rights workers, National human rights institutions, Web archives
Creator: Columbia University Libraries, Center for Human Rights Documentation and Research
Collector: Columbia University Libraries, Center for Human Rights Documentation and Research

Narrow Your Results

Group Sort By: Count | (A-Z)

Amnesty International sections (50)
Blogs by individuals (11)
National human rights institutions (88)
Non-governmental organizations (420)
Truth commissions, tribunals, and courts (14)

Subject Sort By: Count | (A-Z)

Human rights (469)
Human rights advocacy (196)
National human rights institutions (88)
Civil rights (52)
Democracy (43)
More ▼

Creator Sort By: Count | (A-Z)

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here Search Clear

Sites | Search Page Text

Page 1 of 6 (593 Total Results) Next Page ►

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: Anti Caste Discrimination Alliance, ACDA
URL: http://acda.co/
Description: UK-focused organization based in Derby, working to eliminate caste-based discrimination. In English. New site; see www.acdauk.org.uk/ for older site.
Captured 10 times between Oct 13, 2011 and Jan 2, 2014
Subject: Discrimination, Caste, Caste-based discrimination, East Indians
Group: Blogs by individuals
Creator: Anti Caste Discrimination Alliance

(a) Archival metadata for the collection.

https://www.archive-it.org/collections/1068;JSESSIONID@archive-it.org=2807D569ABEF136C

Subject Sort By: Count | (A-Z)

Human rights (469)
Human rights advocacy (196)
National human rights institutions (88)
Civil rights (52)
Democracy (43)
More ▼

Creator Sort By: Count | (A-Z)

Asociacion Paz con Dignidad (4)
Comite de Familiares de Detenidos-Desaparecidos en Honduras (3)
Voix des sans-voix pour les droits de l'homme (Organization) (3)
Amnesty International Hong Kong (2)
Anti Caste Discrimination Alliance (2)
More ▼

Language Sort By: Count | (A-Z)

English (420)
Spanish (109)
Arabic (91)
French (80)
Russian (32)
More ▼

Coverage Sort By: Count | (A-Z)

Egypt (20)
Russia (Federation) (13)
Israel (12)
Argentina (11)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: Anti Caste Discrimination Alliance, ACDA
URL: http://acda.co/
Description: UK-focused organization based in Derby, working to eliminate caste-based discrimination. In English. New site; see www.acdauk.org.uk/ for older site.
Captured 10 times between Oct 13, 2011 and Jan 2, 2014
Subject: Discrimination, Caste, Caste-based discrimination, East Indians
Group: Blogs by individuals
Creator: Anti Caste Discrimination Alliance
Language: English
Coverage: Great Britain
Collector: Columbia University Libraries, Center for Human Rights Documentation and Research

Title: Americans for Democracy and Human Rights in Bahrain
URL:http://adhrb.org/
Description:Americans for Democracy and Human Rights in Bahrain (ADHRB) fosters awareness of and support for democracy and human rights in Bahrain by engaging with key actors within the US government to advocate for stronger US policies that support human rights for the Bahraini people."
Capturedonce on Jan 21, 2014
Subject: Democracy, Human rights
Group: Non-governmental organizations
Creator: Americans for Democracy and Human Rights in Bahrain
Language: English
Coverage: Bahrain
Collector: Columbia University Libraries, Center for Human Rights Documentation and Research

Title: Advocacy Forum--Nepal
URL: http://advocacyforum.org/
Description: Nepal-focused organization based in Kathmandu. Includes reports. In English.
Captured 21 times between Sep 3, 2010 and Jan 2, 2014
Videos: 14 Videos Captured
Subject: Human rights, Rule of law, Human rights advocacy
Group: Non-governmental organizations
Creator: Ed□bhokesi□ Phorama (Kathmandu, Nepal)
Language: English

(b) Alphabetical list of URIs in the collection.

Internet Archive Wayback

https://wayback.archive-it.org/1068/*/http://www.baheyya.blogspot.com/

Center for Human Rights Documentation & Research

Human Rights Web Archive (Columbia University Libraries)

INTERNET ARCHIVE WaybackMachine

Enter Web Address: http:// All ▾ Take Me Back

Searched for http://www.baheyya.blogspot.com/ 27 Results RSS Metadata

Look up URL in general Internet Archive web collection Proxy Mode Help

* denotes when page was updated

Found 27 Captures between Jan 29, 2011 - Jan 3, 2014

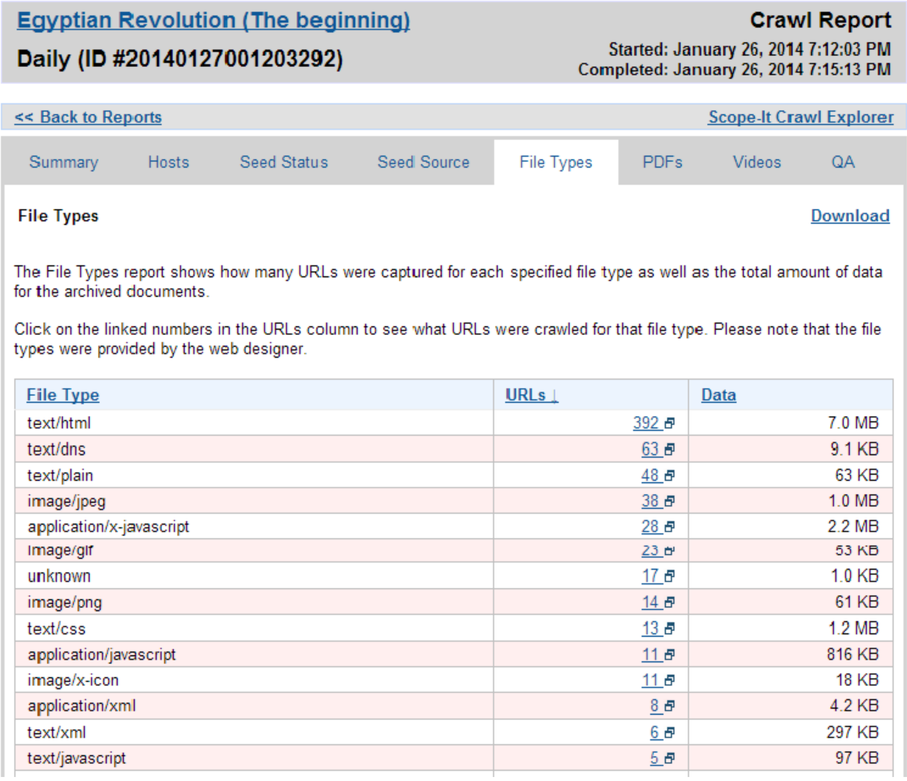
2011	2012	2013	2014
8 pages	11 pages	6 pages	2 pages
Jan 29, 2011 *	Mar 2, 2012 *	Mar 6, 2013 *	Jan 3, 2014 *
Mar 2, 2011 *	Mar 3, 2012 *	Mar 7, 2013 *	Jan 3, 2014 *
Jun 2, 2011 *	Jun 2, 2012 *	Jun 11, 2013 *	
Sep 2, 2011 *	Jun 3, 2012 *	Jun 12, 2013 *	
Sep 3, 2011	Jun 3, 2012	Oct 1, 2013 *	
Sep 3, 2011	Sep 2, 2012 *	Oct 2, 2013 *	
Dec 2, 2011 *	Sep 2, 2012 *		
Dec 3, 2011 *	Sep 2, 2012		
	Dec 2, 2012 *		

(c) Archived copies of the first URI in the collection.

FIG. 7: Current browsing and searching services for the "Human Rights" collection in Archive-It.

Collection understanding is defined as gaining a comprehensive view of a collection [70]. When an archivist creates a collection, it can include 1000s of seed URIs (see Chapter 7). Over time, each of these URIs can be crawled 100s or 1000s of times, resulting in a collection having thousands to millions of archived Web pages. Understanding the contents and boundaries of a collection is then difficult for most people, resulting in the paradox of the larger the collection, the harder it is to use.

Figure 7 is another example that shows the current browsing interface for a collection about human rights. It is difficult for users arriving at the page shown in Figure 7(a) to understand what is in this collection and how it differs from the approximately 17 other collections in Archive-It that are also about human rights, albeit each with their own specialization.



Crawl Report

Daily (ID #20140127001203292)

Started: January 26, 2014 7:12:03 PM
Completed: January 26, 2014 7:15:13 PM

File Types

The File Types report shows how many URLs were captured for each specified file type as well as the total amount of data for the archived documents.

Click on the linked numbers in the URLs column to see what URLs were crawled for that file type. Please note that the file types were provided by the web designer.

File Type	URLs ↓	Data
text/html	392	7.0 MB
text/dns	63	9.1 KB
text/plain	48	63 KB
image/jpeg	38	1.0 MB
application/x-javascript	28	2.2 MB
image/gif	23	53 KB
unknown	17	1.0 KB
image/png	14	61 KB
text/css	13	1.2 MB
application/javascript	11	816 KB
image/x-icon	11	18 KB
application/xml	8	4.2 KB
text/xml	6	297 KB
text/javascript	5	97 KB

FIG. 8: Archive-It provides the collection curators with information about their crawls. Retrieved in January 2016.

Providing a summary of the content of archived collections is a challenge because there are two dimensions that should be summarized: the URIs that comprise the collection (e.g., `cnn.com`) and the archived copies (called "mementos") of those URIs at different times (e.g., `cnn.com`@t_1, `cnn.com`@t_2,.., `bbc.co.uk`@t_n). Either dimension by itself is difficult, but combined they present a number of challenges, and are hard to adapt to most conventional visualization techniques.

We have explored applying well-known, advanced visual interfaces (e.g., timelines, wordles, bubble charts, image plots with histogram) to Archive-It collections and the results are sufficient for those already with an understanding of what is in the collection, but they do not facilitate an understanding to those who are unfamiliar with collection [252]. One problem with the above approaches is there is not an emphasis on ignoring content: there is often an implicit assumption that everything in a collection is equally valuable and should be visualized. Some of the web pages change frequently and some are near-duplicates. Some go off-topic and no longer contribute to the collection. Furthermore, collections grow quickly: the Human Rights collection in Figure 7(a) has nearly 1000 seed URIs, and each

URI has between one and 60 archived pages. Visualization techniques with an emphasis on recall (i.e., “here’s everything in the collection”) do not scale.

1.2.2 ISSUES WITH SEED URIS

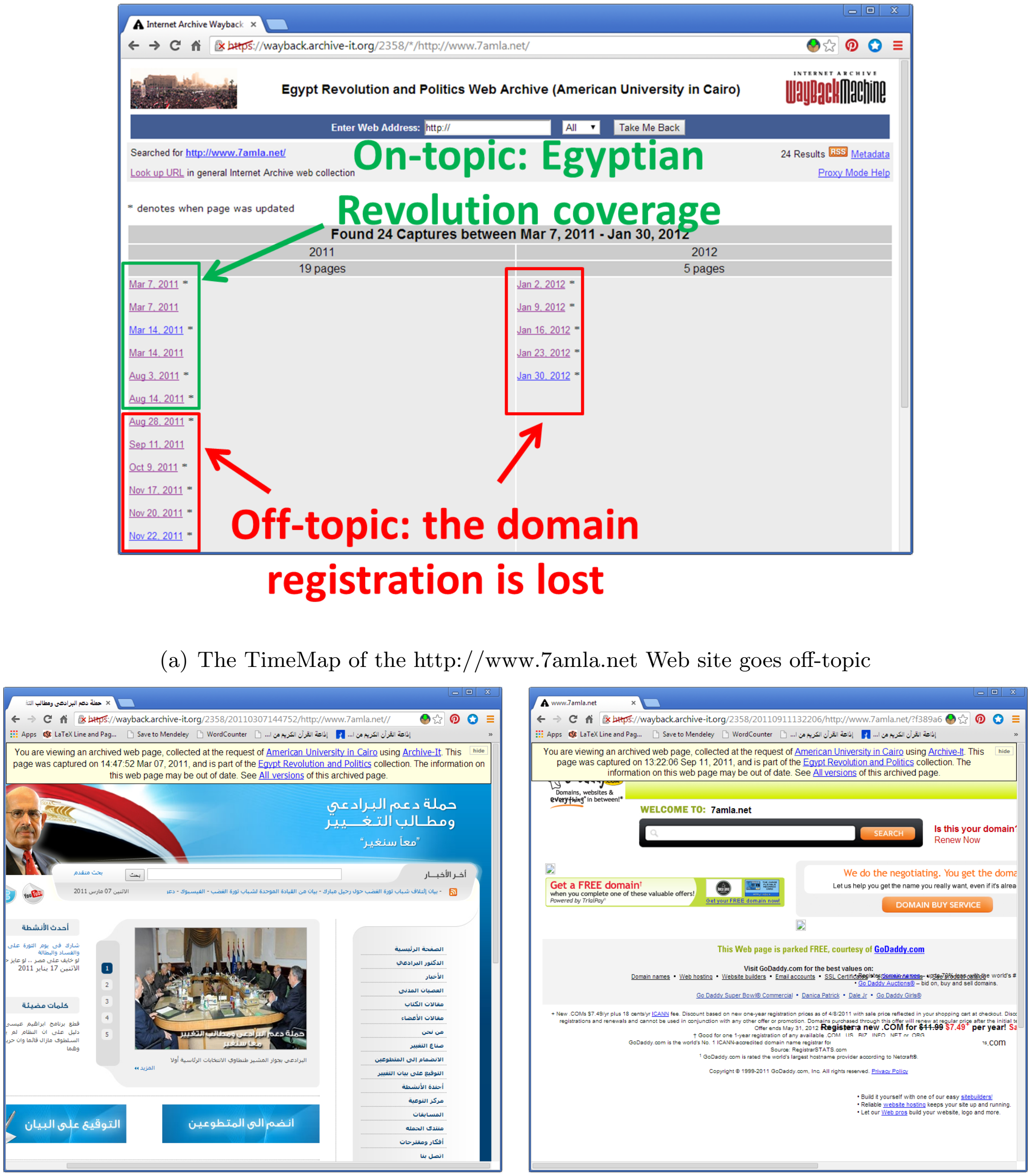
Archive-It provides tools that allow collection curators to perform collection management as well as quality control for the crawls. However, the tools are currently focused on crawl issues such as the mechanics of HTTP (e.g., how many HTML files and other file types, how many 404 missing URIs). For example, Figure 8 shows a report of the file types in the created collection. Currently, there are no content-based tools that allow curators to detect when seed URIs (and other crawled pages) are off-topic to discover candidate seed URIs that are not currently included. Off-topic here means pages that are not relevant to the topic of the collection.

Figures 9 and 10 are different scenarios for pages in the “Egypt Revolution and Politics” collection that go off-topic. Figure 9(a) shows the TimeMap of a relevant Web page (Figure 9(b)) that goes off-topic after losing the domain registration (Figure 9(c)). Figure 10(a) is the TimeMap of the homepage of Middle East news on BBC[10], which has multiple archived pages over many years. The page goes off-topic (Figure 10(c)) and on-topic (Figure 10(b)) many times. For example, some archived copies of this page contains news about Syria, Iraq, etc. This is an example for the frequency of change of the “aboutness” of the page in terms of relevancy to the collection. The relevancy can be for the topic of the page or the topic of the collection. For example, the topic of the page in Figure 10 should be relevant to the countries in the Middle East, not only Egypt. So, in terms of the page topic, the archived copies of this page are all on-topic. However, the collection that contains this page is about Egypt Revolution and Politics only. So the pages go off-topic relative to the collection when their content has nothing about Egypt. There are different cases for changing the “aboutness” of a page through time. We will explain these cases in detail in Chapter 8.

1.3 WITNESSING/LIVING THE PAST

There are multiple stories that can be generated from an archived collection with different perspectives about the collection. For example, a user may want to see a story that is composed of the key events from a specific Web site, a story that is composed of the key events of the story regardless of the sources, or how a specific event at a specific point in time was covered by different Web sites, etc. We will explore many different types of stories. The story types will be defined and explained in Chapter 5.

[10] `http://www.bbc.co.uk/news/world/middle_east/`



Internet Archive Wayback

~~https~~://wayback.archive-it.org/2358/*/http://www.7amla.net/

Egypt Revolution and Politics Web Archive (American University in Cairo)

Enter Web Address: http:// All Take Me Back

Searched for http://www.7amla.net/ 24 Results RSS Metadata

Look up URL in general Internet Archive web collection Proxy Mode Help

* denotes when page was updated

Found 24 Captures between Mar 7, 2011 - Jan 30, 2012

2011 19 pages	2012 5 pages
Mar 7, 2011 *	Jan 2, 2012 *
Mar 7, 2011	Jan 9, 2012 *
Mar 14, 2011 *	Jan 16, 2012 *
Mar 14, 2011	Jan 23, 2012 *
Aug 3, 2011 *	Jan 30, 2012 *
Aug 14, 2011 *	
Aug 28, 2011 *	
Sep 11, 2011	
Oct 9, 2011 *	
Nov 17, 2011 *	
Nov 20, 2011 *	
Nov 22, 2011 *	

On-topic: Egyptian Revolution coverage
Off-topic: the domain registration is lost

(a) The TimeMap of the http://www.7amla.net Web site goes off-topic

~~https~~://wayback.archive-it.org/2358/20110307144752/http://www.7amla.net/

You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 14:47:52 Mar 07, 2011, and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page.

(b) March 07, 2011: on-topic

~~https~~://wayback.archive-it.org/2358/20110911132206/http://www.7amla.net/?f389a6

You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 13:22:06 Sep 11, 2011, and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page.

WELCOME TO: 7amla.net

Is this your domain? Renew Now

Get a FREE domain!

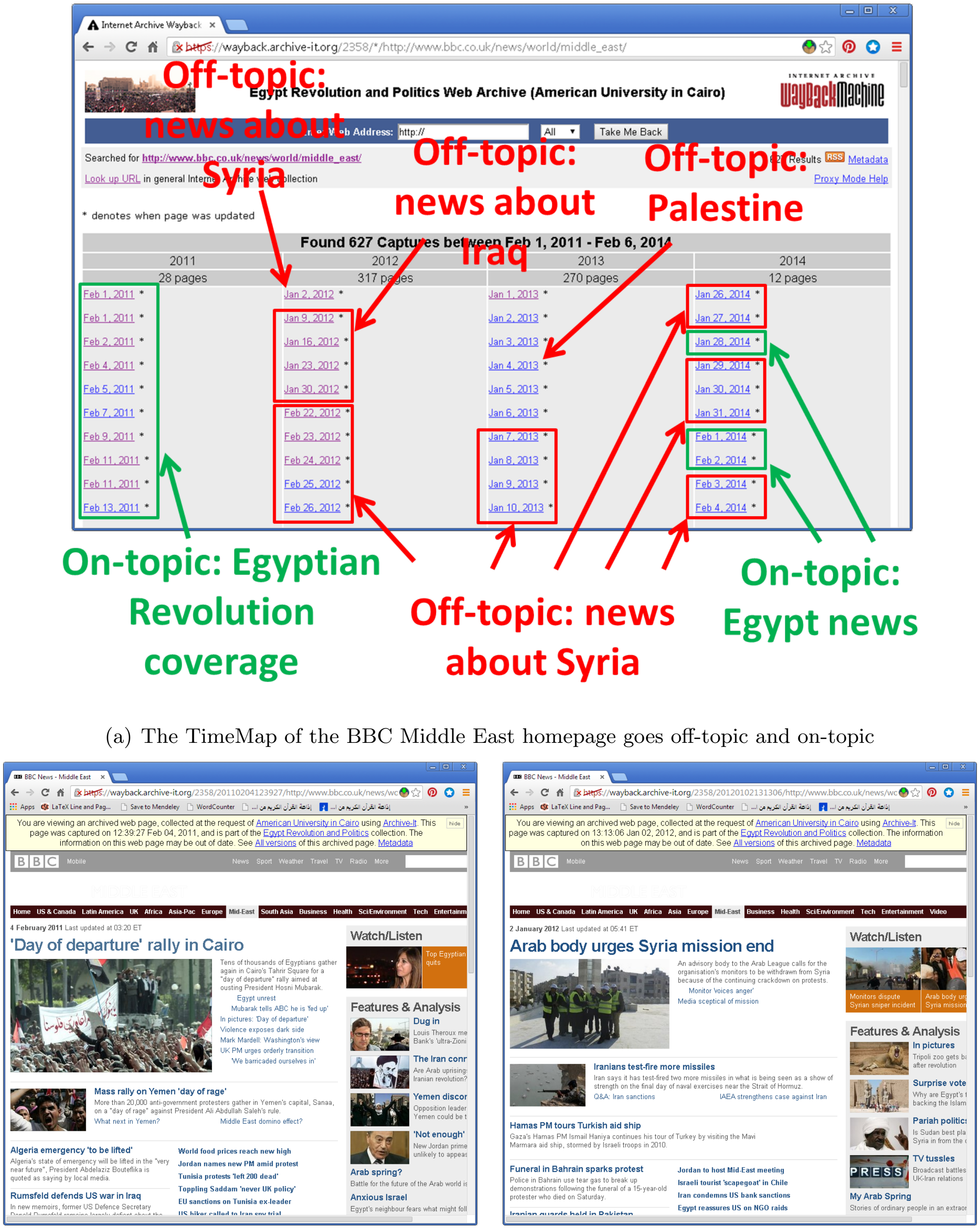
We do the negotiating. You get the doma

DOMAIN BUY SERVICE

This Web page is parked FREE, courtesy of GoDaddy.com

(c) Sept. 11, 2011: off-topic the domain registration is lost

FIG. 9: Most of the archived pages of 7amla.com are off-topic, but are still included in the "Egypt Revolution and Politics" collection.



Internet Archive Wayback

https://wayback.archive-it.org/2358/*/http://www.bbc.co.uk/news/world/middle_east/

Egypt Revolution and Politics Web Archive (American University in Cairo)

Web Address: http:// All Take Me Back

Searched for http://www.bbc.co.uk/news/world/middle_east/ 627 Results RSS Metadata

Look up URL in general Internet Archive Wayback collection Proxy Mode Help

* denotes when page was updated

Found 627 Captures between Feb 1, 2011 - Feb 6, 2014

2011	2012	2013	2014
28 pages	317 pages	270 pages	12 pages
Feb 1, 2011 *	Jan 2, 2012 *	Jan 1, 2013 *	Jan 26, 2014 *
Feb 1, 2011 *	Jan 9, 2012 *	Jan 2, 2013 *	Jan 27, 2014 *
Feb 2, 2011 *	Jan 16, 2012 *	Jan 3, 2013 *	Jan 28, 2014 *
Feb 4, 2011 *	Jan 23, 2012 *	Jan 4, 2013 *	Jan 29, 2014 *
Feb 5, 2011 *	Jan 30, 2012 *	Jan 5, 2013 *	Jan 30, 2014 *
Feb 7, 2011 *	Feb 22, 2012 *	Jan 6, 2013 *	Jan 31, 2014 *
Feb 9, 2011 *	Feb 23, 2012 *	Jan 7, 2013 *	Feb 1, 2014 *
Feb 11, 2011 *	Feb 24, 2012 *	Jan 8, 2013 *	Feb 2, 2014 *
Feb 11, 2011 *	Feb 25, 2012 *	Jan 9, 2013 *	Feb 3, 2014 *
Feb 13, 2011 *	Feb 26, 2012 *	Jan 10, 2013 *	Feb 4, 2014 *

Off-topic: news about Syria

Off-topic: news about Iraq

Off-topic: Palestine

On-topic: Egyptian Revolution coverage

Off-topic: news about Syria

On-topic: Egypt news

(a) The TimeMap of the BBC Middle East homepage goes off-topic and on-topic

BBC News - Middle East

https://wayback.archive-it.org/2358/20110204123927/http://www.bbc.co.uk/news/wc

You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 12:39:27 Feb 04, 2011, and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page. Metadata

BBC Mobile News Sport Weather Travel TV Radio More

Home US & Canada Latin America UK Africa Asia-Pac Europe Mid-East South Asia Business Health Sci/Environment Tech Entertainm

4 February 2011 Last updated at 03:20 ET

'Day of departure' rally in Cairo

Tens of thousands of Egyptians gather again in Cairo's Tahrir Square for a "day of departure" rally aimed at ousting President Hosni Mubarak.

Egypt unrest
Mubarak tells ABC he is 'fed up'
In pictures: 'Day of departure'
Violence exposes dark side
Mark Mardell: Washington's view
UK PM urges orderly transition
'We barricaded ourselves in'

Mass rally on Yemen 'day of rage'

More than 20,000 anti-government protesters gather in Yemen's capital, Sanaa, on a "day of rage" against President Ali Abdullah Saleh's rule.

What next in Yemen? Middle East domino effect?

Algeria emergency 'to be lifted'

Algeria's state of emergency will be lifted in the "very near future", President Abdelaziz Bouteflika is quoted as saying by local media.

Rumsfeld defends US war in Iraq

In new memoirs, former US Defence Secretary

World food prices reach new high
Jordan names new PM amid protest
Tunisia protests 'left 200 dead'
Toppling Saddam 'never UK policy'
EU sanctions on Tunisia ex-leader

Watch/Listen

Top Egyptian quits

Features & Analysis

Dug in
Louis Theroux me Bank's 'ultra-Zioni

The Iran conr
Are Arab uprising Iranian revolution?

Yemen discor
Opposition leader Yemen could be t

'Not enough'
New Jordan prime unlikely to appeas

Arab spring?
Battle for the future of the Arab world is

Anxious Israel
Egypt's neighbour fears what might foll

(b) Feb. 04, 2011: on-topic

BBC News - Middle East

https://wayback.archive-it.org/2358/20120102131306/http://www.bbc.co.uk/news/wc

You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 13:13:06 Jan 02, 2012, and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page. Metadata

BBC Mobile News Sport Weather Travel TV Radio More

Home US & Canada Latin America UK Africa Asia Europe Mid-East Business Health Sci/Environment Tech Entertainment Video

2 January 2012 Last updated at 05:41 ET

Arab body urges Syria mission end

An advisory body to the Arab League calls for the organisation's monitors to be withdrawn from Syria because of the continuing crackdown on protests.

Monitor 'voices anger'
Media sceptical of mission

Iranians test-fire more missiles

Iran says it has test-fired two more missiles in what is being seen as a show of strength on the final day of naval exercises near the Strait of Hormuz.

Q&A: Iran sanctions IAEA strengthens case against Iran

Hamas PM tours Turkish aid ship

Gaza's Hamas PM Ismail Haniya continues his tour of Turkey by visiting the Mavi Marmara aid ship, stormed by Israeli troops in 2010.

Funeral in Bahrain sparks protest

Police in Bahrain use tear gas to break up demonstrations following the funeral of a 15-year-old protester who died on Saturday.

Jordan to host Mid-East meeting
Israeli tourist 'scapegoat' in Chile
Iran condemns US bank sanctions
Egypt reassures US on NGO raids

Watch/Listen

Monitors dispute Syrian sniper incident

Arab body urg Syria mission

Features & Analysis

In pictures
Tripoli zoo gets ba after revolution

Surprise vote
Why are Egypt's t backing the Islam

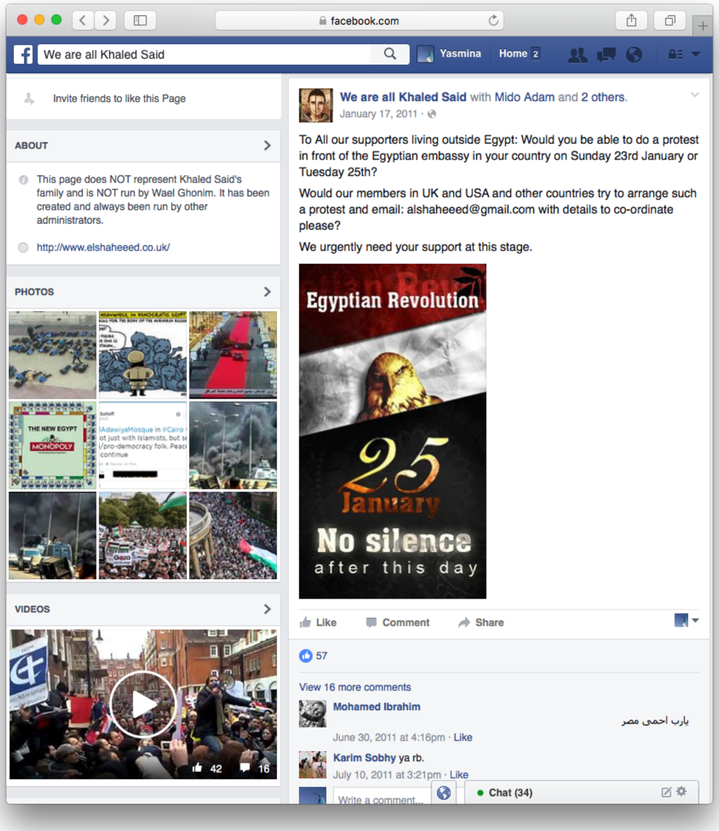
Pariah politics
Is Sudan best pla Syria in from the c

TV tussles
Broadcast battles UK-Iran relations

My Arab Spring
Stories of ordinary people in an extraor

(c) Jan. 02, 2012: not relevant to the Egyptian Revolution

FIG. 10: An example of a URI that oscillates between on-topic and off-topic in the “Egypt Revolution and Politics” collection.



This page does NOT represent Khaled Said's family and is NOT run by Wael Ghonim. It has been created and always been run by other administrators.

http://www.elshaheeed.co.uk/

We are all Khaled Said with Mido Adam and 2 others.
January 17, 2011 ·

To All our supporters living outside Egypt: Would you be able to do a protest in front of the Egyptian embassy in your country on Sunday 23rd January or Tuesday 25th?

Would our members in UK and USA and other countries try to arrange such a protest and email: alshaheeed@gmail.com with details to co-ordinate please?

We urgently need your support at this stage.

Mohamed Ibrahim يارب احمى مصر
June 30, 2011 at 4:16pm · Like

Karim Sobhy ya rb.
July 10, 2011 at 3:21pm · Like

FIG. 11: The beginning of the events for the Jan. 25 Egyptian Revolution started on "We are All Khaled Saeed" Facebook page, which formed in the aftermath of Saeed's beating and death. This post is from Jan. 17, 2011 before the start of the Revolution.

In the following scenarios, we show manually created stories that bring insight into two Archive-It collections, the "Egypt Revolution and Politics"[11] and the "2013 Boston Marathon Bombing"[12], using different sets of archived pages from the collections.

1.3.1 THE JAN. 25 EGYPTIAN REVOLUTION

I was in Norfolk, Virginia when the uprisings of the Jan. 25 Egyptian Revolution started. I remember my feeling at that time and how I badly wanted to go back to Egypt and do something for freedom and dignity. I could not do something during the 18 days except watch all the news and social media channels, witnessing the events. It started with a group of brave young Egyptians calling for demonstrations on Facebook and Twitter (Figure 11).

[11] `https://archive-it.org/collections/2358/`

[12] `https://archive-it.org/collections/3649/`

Millions of people took to the streets in a nationwide protest against President Hosni Mubarak. They aimed to battle injustice, corruption, and poverty. Street demonstrations quickly grew into a national revolutionary movement that in 18 days removed Mubarak and his National Democratic Party (NDP). In the following subsections, we will go back in time and see the 18 days and other perspectives about the Egyptian Revolution as they appeared on the Web.

How did the Jan. 25 Egyptian Revolution evolve over time in 18 days?

- 2011-01-25: Tens of thousands of young people gathered in Tahrir Square on January 25, 2011 protesting against the government (Figure 12(a)).

- 2011-01-27: Newspapers started full coverage of the protests with increasing number of protesters because of violent clashes between security forces and protesters (Figures 12(b)).

- 2011-01-31: The Egyptian military took to the streets, but vowed not to use force against protesters (Figure 12(c)).

- 2011-01-31 to 2011-02-02: With the increasing anger and the number of protests all over Egypt, the government used multiple ways to stop the protests, such as shutting down access to the Internet [33, 337] and suppressing the media to close communications as these were the main methods that gathered and connected the people. During this period, there was also a speech from Mubarak promising not to seek re-elections, police brutality against the protesters, deadly attacks and clashes from the pro-Mubarak protesters, etc. (Figures 12(d) - 12(f) and Figure 13(a)).

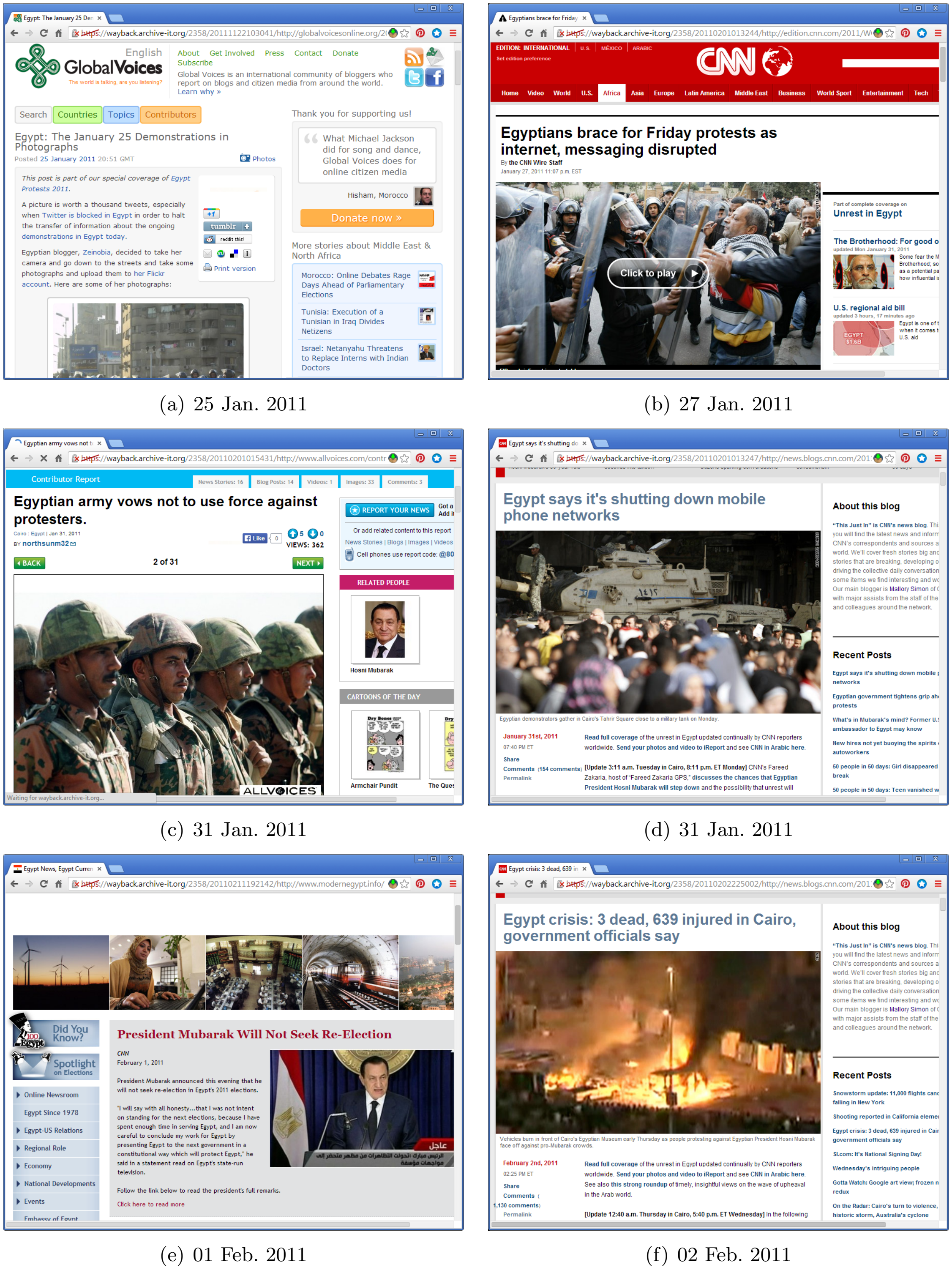
- 2011-02-04: After the number of martyrs increased, the people's anger grew and the numbers of protesters increased significantly all over the country (Figure 13(b)).

- 2011-02-07: During the protests, Google executive Wael Ghonim revealed that he was behind the account of "We are All Khaled Saeed"[13] Facebook page, which started the anti-government protests that began on Jan. 25. He was arrested during the protests, then he was released (Figure 13(c)).

- 2011-02-10: Mubarak re-appeared on television in Feb. 10, 2011 and struck a defiant tone (Figure 13(d)).

- 2011-02-11: The crowd raised their shoes in a response to his speech and insisted that they will not leave until he leaves (Figure 13(e)).

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



Egypt: The January 25 Dem
https://wayback.archive-it.org/2358/20111122103041/http://globalvoicesonline.org/2
English
GlobalVoices
The world is talking, are you listening?
About Get Involved Press Contact Donate
Subscribe
Search
Countries
Topics
Contributors
Photos
tumblr
reddit this!
Print version
Donate now »

Global Voices is an international community of bloggers who report on blogs and citizen media from around the world. Learn why »

Egypt: The January 25 Demonstrations in Photographs

Posted 25 January 2011 20:51 GMT

This post is part of our special coverage of Egypt Protests 2011.

A picture is worth a thousand tweets, especially when Twitter is blocked in Egypt in order to halt the transfer of information about the ongoing demonstrations in Egypt today.

Egyptian blogger, Zeinobia, decided to take her camera and go down to the streets and take some photographs and upload them to her Flickr account. Here are some of her photographs:

Thank you for supporting us!

“What Michael Jackson did for song and dance, Global Voices does for online citizen media

Hisham, Morocco

More stories about Middle East & North Africa

Morocco: Online Debates Rage Days Ahead of Parliamentary Elections

Tunisia: Execution of a Tunisian in Iraq Divides Netizens

Israel: Netanyahu Threatens to Replace Interns with Indian Doctors

(a) 25 Jan. 2011

Egyptians brace for Friday protests as internet, messaging disrupted

By the CNN Wire Staff

January 27, 2011 11:07 p.m. EST

Part of complete coverage on

Unrest in Egypt

The Brotherhood: For good o

updated Mon January 31, 2011

Some fear the M Brotherhood; so as a potential pa how influential is

U.S. regional aid bill

updated 3 hours, 17 minutes ago

Egypt is one of t when it comes t U.S. aid

(b) 27 Jan. 2011

Egyptian army vows not to use force against protesters.

Cairo : Egypt | Jan 31, 2011

BY northsunm32

Or add related content to this report

News Stories | Blogs | Images | Videos

Cell phones use report code: @80

(c) 31 Jan. 2011

Egypt says it's shutting do
https://wayback.archive-it.org/2358/20110201013247/http://news.blogs.cnn.com/201
AFP/GETTY IMAGES
Share
Comments
Permalink

Egypt says it's shutting down mobile phone networks

Egyptian demonstrators gather in Cairo's Tahrir Square close to a military tank on Monday.

January 31st, 2011

07:40 PM ET

(154 comments)

Read full coverage of the unrest in Egypt updated continually by CNN reporters worldwide. Send your photos and video to iReport and see CNN in Arabic here.

[Update 3:11 a.m. Tuesday in Cairo, 8:11 p.m. ET Monday] CNN's Fareed Zakaria, host of “Fareed Zakaria GPS,” discusses the chances that Egyptian President Hosni Mubarak will step down and the possibility that unrest will

About this blog

“This Just In” is CNN's news blog. Thi you will find the latest news and inform CNN's correspondents and sources a world. We'll cover fresh stories big and stories that are breaking, developing o driving the collective daily conversation some items we find interesting and wo Our main blogger is Mallory Simon of C with major assists from the staff of the and colleagues around the network.

Recent Posts

Egypt says it's shutting down mobile p networks

Egyptian government tightens grip ah protests

What's in Mubarak's mind? Former U.S ambassador to Egypt may know

New hires not yet buoying the spirits autoworkers

50 people in 50 days: Girl disappeared break

50 people in 50 days: Teen vanished w

(d) 31 Jan. 2011

President Mubarak Will Not Seek Re-Election

CNN

February 1, 2011

President Mubarak announced this evening that he will not seek re-election in Egypt's 2011 elections.

"I will say with all honesty...that I was not intent on standing for the next elections, because I have spent enough time in serving Egypt, and I am now careful to conclude my work for Egypt by presenting Egypt to the next government in a constitutional way which will protect Egypt," he said in a statement read on Egypt's state-run television.

Follow the link below to read the president's full remarks.

Click here to read more

(e) 01 Feb. 2011

Egypt crisis: 3 dead, 639 in
https://wayback.archive-it.org/2358/20110202225002/http://news.blogs.cnn.com/201
CNN
Share
Comments
Permalink

Egypt crisis: 3 dead, 639 injured in Cairo, government officials say

Vehicles burn in front of Cairo's Egyptian Museum early Thursday as people protesting against Egyptian President Hosni Mubarak face off against pro-Mubarak crowds.

February 2nd, 2011

02:25 PM ET

(1,130 comments)

Read full coverage of the unrest in Egypt updated continually by CNN reporters worldwide. Send your photos and video to iReport and see CNN in Arabic here. See also this strong roundup of timely, insightful views on the wave of upheaval in the Arab world.

[Update 12:40 a.m. Thursday in Cairo, 5:40 p.m. ET Wednesday] In the following

About this blog

“This Just In” is CNN's news blog. Thi you will find the latest news and inform CNN's correspondents and sources a world. We'll cover fresh stories big and stories that are breaking, developing o driving the collective daily conversation some items we find interesting and wo Our main blogger is Mallory Simon of C with major assists from the staff of the and colleagues around the network.

Recent Posts

Snowstorm update: 11,000 flights canc falling in New York

Shooting reported in California eleme

Egypt crisis: 3 dead, 639 injured in Cair government officials say

SI.com: It's National Signing Day!

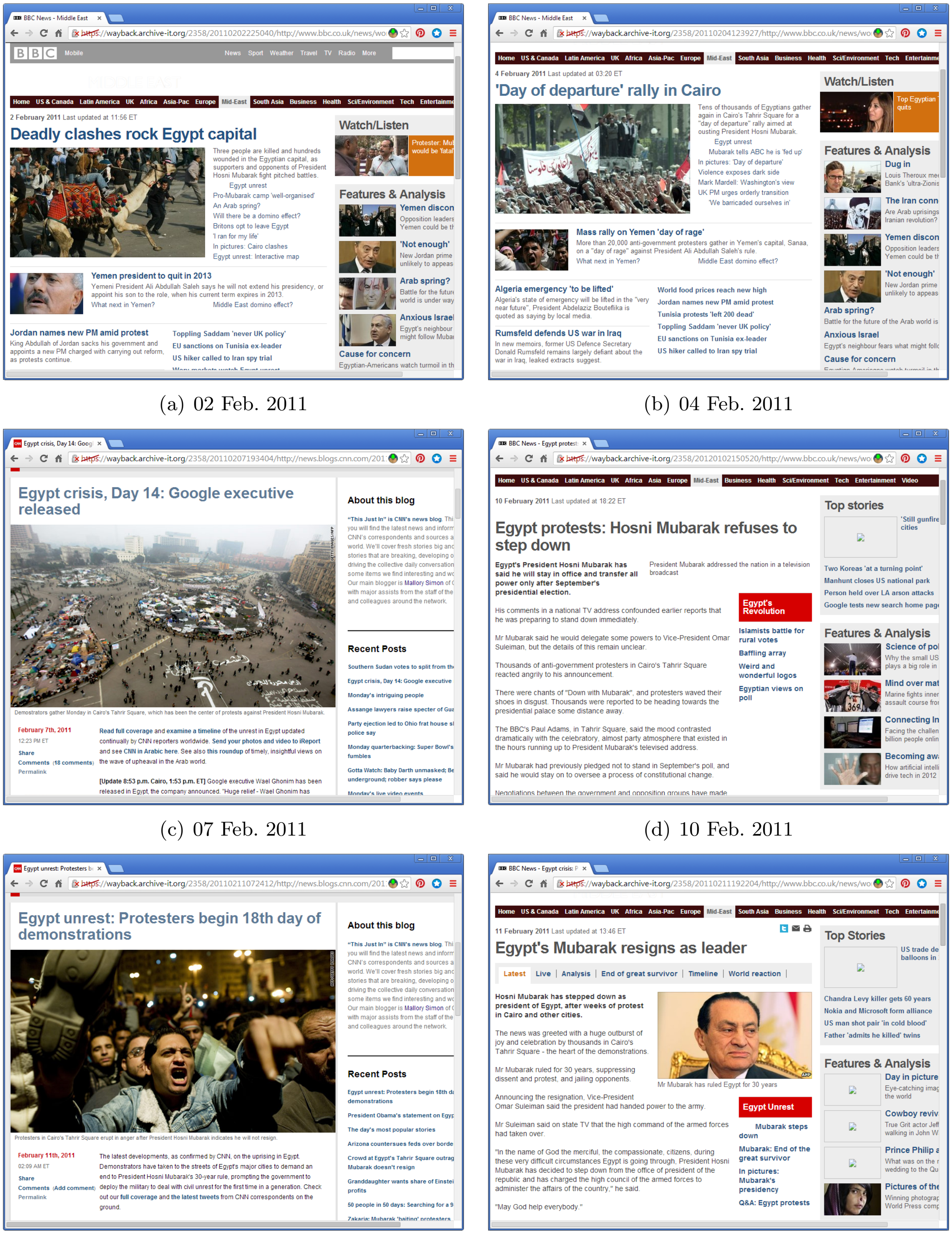
Wednesday's intriguing people

Gotta Watch: Google art view; frozen n redux

On the Radar: Cairo's turn to violence, historic storm, Australia's cyclone

(f) 02 Feb. 2011

FIG. 12: Coverage of the Egyptian Revolution from different Web sites at different times.



BBC News - Middle East

https://wayback.archive-it.org/2358/20110202225040/http://www.bbc.co.uk/news/wo

Home | US & Canada | Latin America | UK | Africa | Asia-Pac | Europe | Mid-East | South Asia | Business | Health | Sci/Environment | Tech | Entertainme

2 February 2011 Last updated at 11:56 ET

Deadly clashes rock Egypt capital

Three people are killed and hundreds wounded in the Egyptian capital, as supporters and opponents of President Hosni Mubarak fight pitched battles.

Egypt unrest
Pro-Mubarak camp 'well-organised'
An Arab spring?
Will there be a domino effect?
Britons opt to leave Egypt
'I ran for my life'
In pictures: Cairo clashes
Egypt unrest: Interactive map

Yemen president to quit in 2013

Yemeni President Ali Abdullah Saleh says he will not extend his presidency, or appoint his son to the role, when his current term expires in 2013.

What next in Yemen? Middle East domino effect?

Jordan names new PM amid protest

King Abdullah of Jordan sacks his government and appoints a new PM charged with carrying out reform, as protests continue.

Toppling Saddam 'never UK policy'
EU sanctions on Tunisia ex-leader
US hiker called to Iran spy trial

Watch/Listen

Protester: Mub would be 'fatal'

Features & Analysis

Yemen discon
Opposition leaders Yemen could be th

'Not enough'
New Jordan prime unlikely to appeas

Arab spring?
Battle for the future world is under way

Anxious Israel
Egypt's neighbour might follow Mubar

Cause for concern
Egyptian-Americans watch turmoil in th

(a) 02 Feb. 2011

BBC News - Middle East

https://wayback.archive-it.org/2358/20110204123927/http://www.bbc.co.uk/news/wo

Home | US & Canada | Latin America | UK | Africa | Asia-Pac | Europe | Mid-East | South Asia | Business | Health | Sci/Environment | Tech | Entertainme

4 February 2011 Last updated at 03:20 ET

'Day of departure' rally in Cairo

Tens of thousands of Egyptians gather again in Cairo's Tahrir Square for a "day of departure" rally aimed at ousting President Hosni Mubarak.

Egypt unrest
Mubarak tells ABC he is 'fed up'
In pictures: 'Day of departure'
Violence exposes dark side
Mark Mardell: Washington's view
UK PM urges orderly transition
'We barricaded ourselves in'

Mass rally on Yemen 'day of rage'

More than 20,000 anti-government protesters gather in Yemen's capital, Sanaa, on a "day of rage" against President Ali Abdullah Saleh's rule.

What next in Yemen? Middle East domino effect?

Algeria emergency 'to be lifted'

Algeria's state of emergency will be lifted in the "very near future", President Abdelaziz Bouteflika is quoted as saying by local media.

Rumsfeld defends US war in Iraq

In new memoirs, former US Defence Secretary Donald Rumsfeld remains largely defiant about the war in Iraq, leaked extracts suggest.

World food prices reach new high
Jordan names new PM amid protest
Tunisia protests 'left 200 dead'
Toppling Saddam 'never UK policy'
EU sanctions on Tunisia ex-leader
US hiker called to Iran spy trial

Watch/Listen

Top Egyptian quits

Features & Analysis

Dug in
Louis Theroux mee Bank's 'ultra-Zionis

The Iran conn
Are Arab uprisings Iranian revolution?

Yemen discon
Opposition leaders Yemen could be th

'Not enough'
New Jordan prime unlikely to appeas

Arab spring?
Battle for the future of the Arab world is

Anxious Israel
Egypt's neighbour fears what might follo

Cause for concern

(b) 04 Feb. 2011

Egypt crisis, Day 14: Googl

https://wayback.archive-it.org/2358/20110207193404/http://news.blogs.cnn.com/201

Egypt crisis, Day 14: Google executive released

Demostrators gather Monday in Cairo's Tahrir Square, which has been the center of protests against President Hosni Mubarak.

February 7th, 2011
12:23 PM ET
Share
Comments (18 comments)
Permalink

Read full coverage and examine a timeline of the unrest in Egypt updated continually by CNN reporters worldwide. Send your photos and video to iReport and see CNN in Arabic here. See also this roundup of timely, insightful views on the wave of upheaval in the Arab world.

[Update 8:53 p.m. Cairo, 1:53 p.m. ET] Google executive Wael Ghonim has been released in Egypt, the company announced. "Huge relief - Wael Ghonim has

About this blog

"This Just In" is CNN's news blog. Thi you will find the latest news and inform CNN's correspondents and sources a world. We'll cover fresh stories big and stories that are breaking, developing o driving the collective daily conversation some items we find interesting and wo Our main blogger is Mallory Simon of C with major assists from the staff of the and colleagues around the network.

Recent Posts

Southern Sudan votes to split from th
Egypt crisis, Day 14: Google executive
Monday's intriguing people
Assange lawyers raise specter of Gua
Party ejection led to Ohio frat house sl police say
Monday quarterbacking: Super Bowl's fumbles
Gotta Watch: Baby Darth unmasked; Be underground; robber says please
Monday's live video events

(c) 07 Feb. 2011

BBC News - Egypt protest

https://wayback.archive-it.org/2358/20120102150520/http://www.bbc.co.uk/news/wo

Home | US & Canada | Latin America | UK | Africa | Asia | Europe | Mid-East | Business | Health | Sci/Environment | Tech | Entertainment | Video

10 February 2011 Last updated at 18:22 ET

Egypt protests: Hosni Mubarak refuses to step down

Egypt's President Hosni Mubarak has said he will stay in office and transfer all power only after September's presidential election.

President Mubarak addressed the nation in a television broadcast

His comments in a national TV address confounded earlier reports that he was preparing to stand down immediately.

Mr Mubarak said he would delegate some powers to Vice-President Omar Suleiman, but the details of this remain unclear.

Thousands of anti-government protesters in Cairo's Tahrir Square reacted angrily to his announcement.

There were chants of "Down with Mubarak", and protesters waved their shoes in disgust. Thousands were reported to be heading towards the presidential palace some distance away.

The BBC's Paul Adams, in Tahrir Square, said the mood contrasted dramatically with the celebratory, almost party atmosphere that existed in the hours running up to President Mubarak's televised address.

Mr Mubarak had previously pledged not to stand in September's poll, and said he would stay on to oversee a process of constitutional change.

Negotiations between the government and opposition groups have made

Egypt's Revolution
Islamists battle for rural votes
Baffling array
Weird and wonderful logos
Egyptian views on poll

Top stories

'Still gunfire cities

Two Koreas 'at a turning point'
Manhunt closes US national park
Person held over LA arson attacks
Google tests new search home page

Features & Analysis

Science of pol
Why the small US plays a big role in

Mind over mat
Marine fights inner assault course fro

Connecting In
Facing the challen billion people onlin

Becoming awa
How artificial intelli drive tech in 2012

(d) 10 Feb. 2011

Egypt unrest: Protesters b

https://wayback.archive-it.org/2358/20110211072412/http://news.blogs.cnn.com/201

Egypt unrest: Protesters begin 18th day of demonstrations

Protesters in Cairo's Tahrir Square erupt in anger after President Hosni Mubarak indicates he will not resign.

February 11th, 2011
02:09 AM ET
Share
Comments (Add comment)
Permalink

The latest developments, as confirmed by CNN, on the uprising in Egypt. Demonstrators have taken to the streets of Egypt's major cities to demand an end to President Hosni Mubarak's 30-year rule, prompting the government to deploy the military to deal with civil unrest for the first time in a generation. Check out our full coverage and the latest tweets from CNN correspondents on the ground.

About this blog

"This Just In" is CNN's news blog. Thi you will find the latest news and inform CNN's correspondents and sources a world. We'll cover fresh stories big and stories that are breaking, developing o driving the collective daily conversation some items we find interesting and wo Our main blogger is Mallory Simon of C with major assists from the staff of the and colleagues around the network.

Recent Posts

Egypt unrest: Protesters begin 18th da demonstrations
President Obama's statement on Egyp
The day's most popular stories
Arizona countersues feds over borde
Crowd at Egypt's Tahrir Square outrag Mubarak doesn't resign
Granddaughter wants share of Einstei profits
50 people in 50 days: Searching for a 9
Zakaria: Mubarak 'baiting' protesters

(e) 11 Feb. 2011

BBC News - Egypt crisis: P

https://wayback.archive-it.org/2358/20110211192204/http://www.bbc.co.uk/news/wo

Home | US & Canada | Latin America | UK | Africa | Asia-Pac | Europe | Mid-East | South Asia | Business | Health | Sci/Environment | Tech | Entertainme

11 February 2011 Last updated at 13:46 ET

Egypt's Mubarak resigns as leader

Latest | Live | Analysis | End of great survivor | Timeline | World reaction |

Hosni Mubarak has stepped down as president of Egypt, after weeks of protest in Cairo and other cities.

The news was greeted with a huge outburst of joy and celebration by thousands in Cairo's Tahrir Square - the heart of the demonstrations.

Mr Mubarak ruled for 30 years, suppressing dissent and protest, and jailing opponents.

Mr Mubarak has ruled Egypt for 30 years

Announcing the resignation, Vice-President Omar Suleiman said the president had handed power to the army.

Mr Suleiman said on state TV that the high command of the armed forces had taken over.

"In the name of God the merciful, the compassionate, citizens, during these very difficult circumstances Egypt is going through, President Hosni Mubarak has decided to step down from the office of president of the republic and has charged the high council of the armed forces to administer the affairs of the country," he said.

"May God help everybody."

Egypt Unrest
Mubarak steps down
Mubarak: End of the great survivor
In pictures: Mubarak's presidency
Q&A: Egypt protests

Top Stories

US trade de balloons in

Chandra Levy killer gets 60 years
Nokia and Microsoft form alliance
US man shot pair 'in cold blood'
Father 'admits he killed' twins

Features & Analysis

Day in picture
Eye-catching imag the world

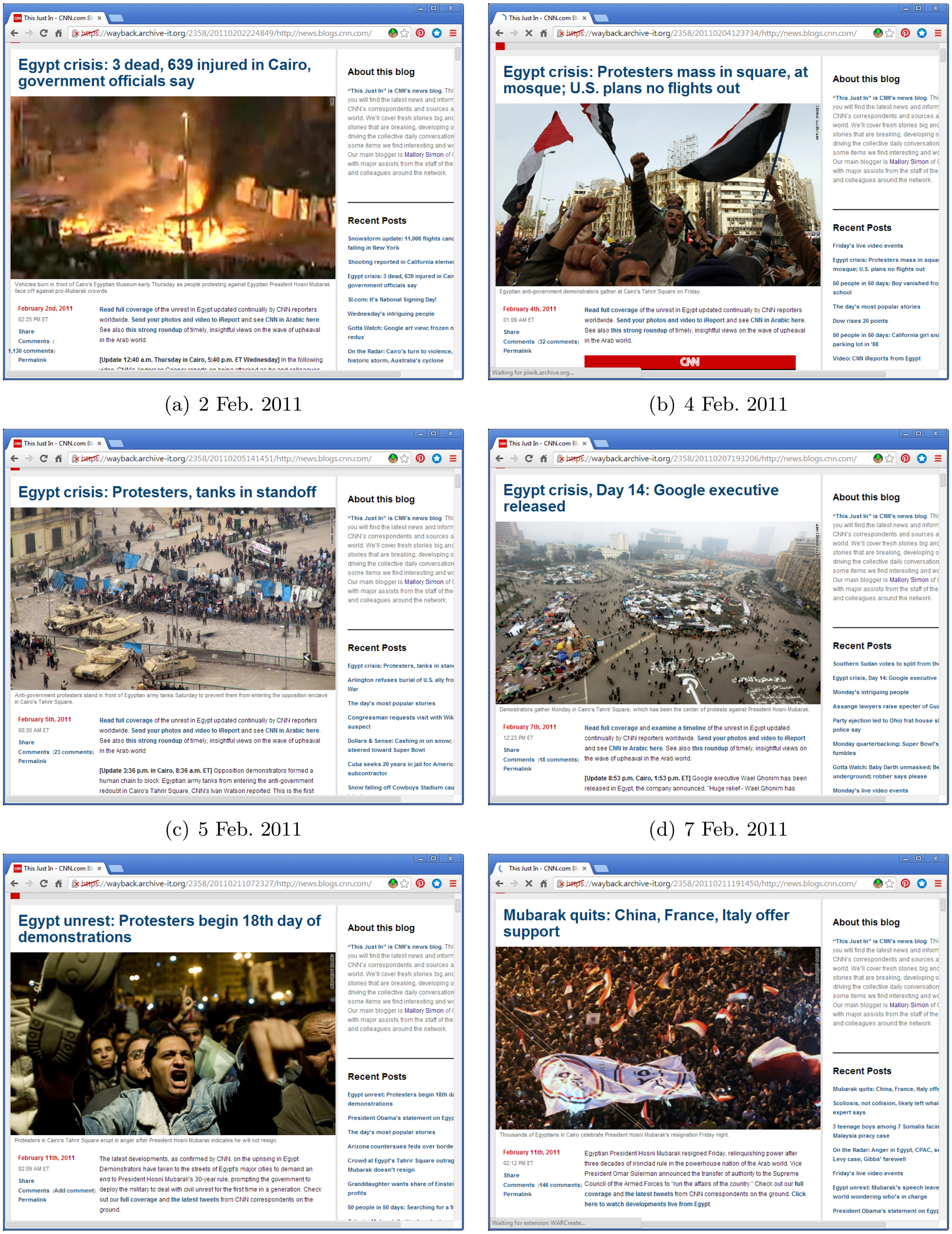
Cowboy reviv
True Grit actor Jeff walking in John W

Prince Philip a
What was on the r wedding to the Qu

Pictures of the
Winning photograp World Press comp

(f) 11 Feb. 2011

FIG. 13: Coverage of the Egyptian Revolution from different Web sites at different times (continued).



This Just In - CNN.com Bl
~~https~~://wayback.archive-it.org/2358/20110202224849/http://news.blogs.cnn.com/

Egypt crisis: 3 dead, 639 injured in Cairo, government officials say

Vehicles burn in front of Cairo's Egyptian Museum early Thursday as people protesting against Egyptian President Hosni Mubarak face off against pro-Mubarak crowds.

February 2nd, 2011
02:25 PM ET
Share
Comments (1,130 comments)
Permalink

Read full coverage of the unrest in Egypt updated continually by CNN reporters worldwide. Send your photos and video to iReport and see CNN in Arabic here. See also this strong roundup of timely, insightful views on the wave of upheaval in the Arab world.

[Update 12:40 a.m. Thursday in Cairo, 5:40 p.m. ET Wednesday] In the following

About this blog

"This Just In" is CNN's news blog. Thi
you will find the latest news and inform
CNN's correspondents and sources a
world. We'll cover fresh stories big and
stories that are breaking, developing o
driving the collective daily conversation
some items we find interesting and wo
Our main blogger is Mallory Simon of C
with major assists from the staff of the
and colleagues around the network.

Recent Posts

Snowstorm update: 11,000 flights canc falling in New York

Shooting reported in California eleme

Egypt crisis: 3 dead, 639 injured in Cair government officials say

SI.com: It's National Signing Day!

Wednesday's intriguing people

Gotta Watch: Google art view; frozen n redux

On the Radar: Cairo's turn to violence, historic storm, Australia's cyclone

(a) 2 Feb. 2011

This Just In - CNN.com Bl
~~https~~://wayback.archive-it.org/2358/20110204123734/http://news.blogs.cnn.com/

Egypt crisis: Protesters mass in square, at mosque; U.S. plans no flights out

AFP/GETTY IMAGES

Egyptian anti-government demonstrators gather at Cairo's Tahrir Square on Friday.

February 4th, 2011
01:09 AM ET
Share
Comments (32 comments)
Permalink

Read full coverage of the unrest in Egypt updated continually by CNN reporters worldwide. Send your photos and video to iReport and see CNN in Arabic here. See also this strong roundup of timely, insightful views on the wave of upheaval in the Arab world.

CNN

About this blog

"This Just In" is CNN's news blog. Thi
you will find the latest news and inform
CNN's correspondents and sources a
world. We'll cover fresh stories big and
stories that are breaking, developing o
driving the collective daily conversation
some items we find interesting and wo
Our main blogger is Mallory Simon of C
with major assists from the staff of the
and colleagues around the network.

Recent Posts

Friday's live video events

Egypt crisis: Protesters mass in squa mosque; U.S. plans no flights out

50 people in 50 days: Boy vanished fro school

The day's most popular stories

Dow rises 20 points

50 people in 50 days: California girl sna parking lot in '88

Video: CNN iReports from Egypt

Waiting for piwik.archive.org...

(b) 4 Feb. 2011

This Just In - CNN.com Bl
~~https~~://wayback.archive-it.org/2358/20110205141451/http://news.blogs.cnn.com/

Egypt crisis: Protesters, tanks in standoff

IVAN WATSON/CNN

Anti-government protesters stand in front of Egyptian army tanks Saturday to prevent them from entering the opposition enclave in Cairo's Tahrir Square.

February 5th, 2011
08:30 AM ET
Share
Comments (23 comments)
Permalink

Read full coverage of the unrest in Egypt updated continually by CNN reporters worldwide. Send your photos and video to iReport and see CNN in Arabic here. See also this strong roundup of timely, insightful views on the wave of upheaval in the Arab world

[Update 3:36 p.m. in Cairo, 8:36 a.m. ET] Opposition demonstrators formed a human chain to block Egyptian army tanks from entering the anti-government redoubt in Cairo's Tahrir Square, CNN's Ivan Watson reported. This is the first

About this blog

"This Just In" is CNN's news blog. Thi
you will find the latest news and inform
CNN's correspondents and sources a
world. We'll cover fresh stories big and
stories that are breaking, developing o
driving the collective daily conversation
some items we find interesting and wo
Our main blogger is Mallory Simon of C
with major assists from the staff of the
and colleagues around the network.

Recent Posts

Egypt crisis: Protesters, tanks in stan

Arlington refuses burial of U.S. ally fro War

The day's most popular stories

Congressman requests visit with Wik suspect

Dollars & Sense: Cashing in on snow; steered toward Super Bowl

Cuba seeks 20 years in jail for Americ subcontractor

Snow falling off Cowboys Stadium cau

(c) 5 Feb. 2011

This Just In - CNN.com Bl
~~https~~://wayback.archive-it.org/2358/20110207193206/http://news.blogs.cnn.com/

Egypt crisis, Day 14: Google executive released

GETTY IMAGES/AFP

Demostrators gather Monday in Cairo's Tahrir Square, which has been the center of protests against President Hosni Mubarak.

February 7th, 2011
12:23 PM ET
Share
Comments (18 comments)
Permalink

Read full coverage and examine a timeline of the unrest in Egypt updated continually by CNN reporters worldwide. Send your photos and video to iReport and see CNN in Arabic here. See also this roundup of timely, insightful views on the wave of upheaval in the Arab world.

[Update 8:53 p.m. Cairo, 1:53 p.m. ET] Google executive Wael Ghonim has been released in Egypt, the company announced. "Huge relief - Wael Ghonim has

About this blog

"This Just In" is CNN's news blog. Thi
you will find the latest news and inform
CNN's correspondents and sources a
world. We'll cover fresh stories big and
stories that are breaking, developing o
driving the collective daily conversation
some items we find interesting and wo
Our main blogger is Mallory Simon of C
with major assists from the staff of the
and colleagues around the network.

Recent Posts

Southern Sudan votes to split from th

Egypt crisis, Day 14: Google executive

Monday's intriguing people

Assange lawyers raise specter of Gua

Party ejection led to Ohio frat house sl police say

Monday quarterbacking: Super Bowl's fumbles

Gotta Watch: Baby Darth unmasked; Be underground; robber says please

Monday's live video events

(d) 7 Feb. 2011

This Just In - CNN.com Bl
~~https~~://wayback.archive-it.org/2358/20110211072327/http://news.blogs.cnn.com/

Egypt unrest: Protesters begin 18th day of demonstrations

AFP/GETTY IMAGES

Protesters in Cairo's Tahrir Square erupt in anger after President Hosni Mubarak indicates he will not resign.

February 11th, 2011
02:09 AM ET
Share
Comments (Add comment)
Permalink

The latest developments, as confirmed by CNN, on the uprising in Egypt. Demonstrators have taken to the streets of Egypt's major cities to demand an end to President Hosni Mubarak's 30-year rule, prompting the government to deploy the military to deal with civil unrest for the first time in a generation. Check out our full coverage and the latest tweets from CNN correspondents on the ground.

About this blog

"This Just In" is CNN's news blog. Thi
you will find the latest news and inform
CNN's correspondents and sources a
world. We'll cover fresh stories big and
stories that are breaking, developing o
driving the collective daily conversation
some items we find interesting and wo
Our main blogger is Mallory Simon of C
with major assists from the staff of the
and colleagues around the network.

Recent Posts

Egypt unrest: Protesters begin 18th da demonstrations

President Obama's statement on Egyp

The day's most popular stories

Arizona countersues feds over borde

Crowd at Egypt's Tahrir Square outrag Mubarak doesn't resign

Granddaughter wants share of Einstei profits

50 people in 50 days: Searching for a 9

(e) 11 Feb. 2011

This Just In - CNN.com Bl
~~https~~://wayback.archive-it.org/2358/20110211191450/http://news.blogs.cnn.com/

Mubarak quits: China, France, Italy offer support

AFP/GETTY IMAGES

Thousands of Egyptians in Cairo celebrate President Hosni Mubarak's resignation Friday night.

February 11th, 2011
02:12 PM ET
Share
Comments (146 comments)
Permalink

Egyptian President Hosni Mubarak resigned Friday, relinquishing power after three decades of ironclad rule in the powerhouse nation of the Arab world. Vice President Omar Suleiman announced the transfer of authority to the Supreme Council of the Armed Forces to "run the affairs of the country." Check out our full coverage and the latest tweets from CNN correspondents on the ground. Click here to watch developments live from Egypt.

About this blog

"This Just In" is CNN's news blog. Thi
you will find the latest news and inform
CNN's correspondents and sources a
world. We'll cover fresh stories big and
stories that are breaking, developing o
driving the collective daily conversation
some items we find interesting and wo
Our main blogger is Mallory Simon of C
with major assists from the staff of the
and colleagues around the network.

Recent Posts

Mubarak quits: China, France, Italy offe

Scoliosis, not collision, likely left whal expert says

3 teenage boys among 7 Somalis facin Malaysia piracy case

On the Radar: Anger in Egypt, CPAC, se Levy case, Gibbs' farewell

Friday's live video events

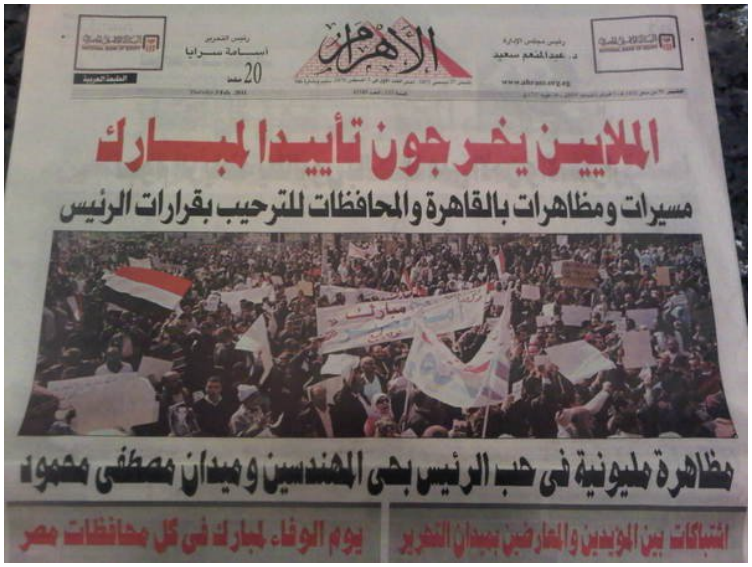
Egypt unrest: Mubarak's speech leave world wondering who's in charge

President Obama's statement on Egyp

Waiting for extension WARCreate...

(f) 11 Feb. 2011

FIG. 14: Coverage of the Egyptian Revolution from CNN's "This Just In" blog at different times.



رئيس مجلس الإدارة
د. عبدالمنعم سعيد

الأهرام

رئيس التحرير
أسامة سرايا

20 صفحة

www.ahram.org.eg

الملايين يخرجون تأييدا لمبارك

مسيرات ومظاهرات بالقاهرة والمحافظات للترحيب بقرارات الرئيس

مظاهرة مليونية فى حب الرئيس بحى المهندسين وميدان مصطفى محمود

اشتباكات بين المؤيدين والمعارضين بميدان التحرير

يوم الوفاء لمبارك فى كل محافظات مصر

FIG. 15: Egyptian State Newspaper, *Al Ahram*: "Millions go out in support of Mubarak: Demonstrations in Cairo and surrounding areas to welcome - Mubarak's latest decisions, Millions demonstrate for their love of the president in Muhandiseen and Mustafa Mahmood Square". Source: `http://imgur.com/DbtK1`

- 2011-02-11: On the Friday of departure (as it was called by the protesters), Egypt's Vice President Omar Suleiman announced that Mubarak would step down after 30 years of rule in an address on state television (Figure 13(f)).

By looking at the Web pages in the previous example, the user can get an idea about the Egyptian Revolution's main events from the start of the protests on Jan. 25, 2011 until Mubarak resigned on Feb. 11, 2011. The story in this section is composed of different Web sites at different times.

How did CNN cover the 18 days of the 25 Jan. Egyptian Revolution?

Figure 14 contains different snapshots of the timeline of the Egyptian Revolution as it appeared on `http://news.blogs.cnn.com/`. We notice that the start date of the crawl for the URIs in the Egyptian Revolution collection is Feb. 1, 2011, which is seven days after the start date of the Egyptian Revolution (Jan. 25, 2011).

This scenario shows the evolution of a single page through time. There are several cases where the user might want to see the evolution of a single Web page through time [150]. For example, a user might be interested in the main changes of a popular Web site, or key events from specific Web sites during given period.

How did different newspapers cover Mubarak resigning?

Egypt is the largest Arabic country and has played a central role in Middle Eastern politics. Therefore, there were widely varying reactions toward the Egyptian Revolution, nationally and internationally. This is how Pasha described the coverage of the Egyptian media during the 18 days of demonstrations [256]:

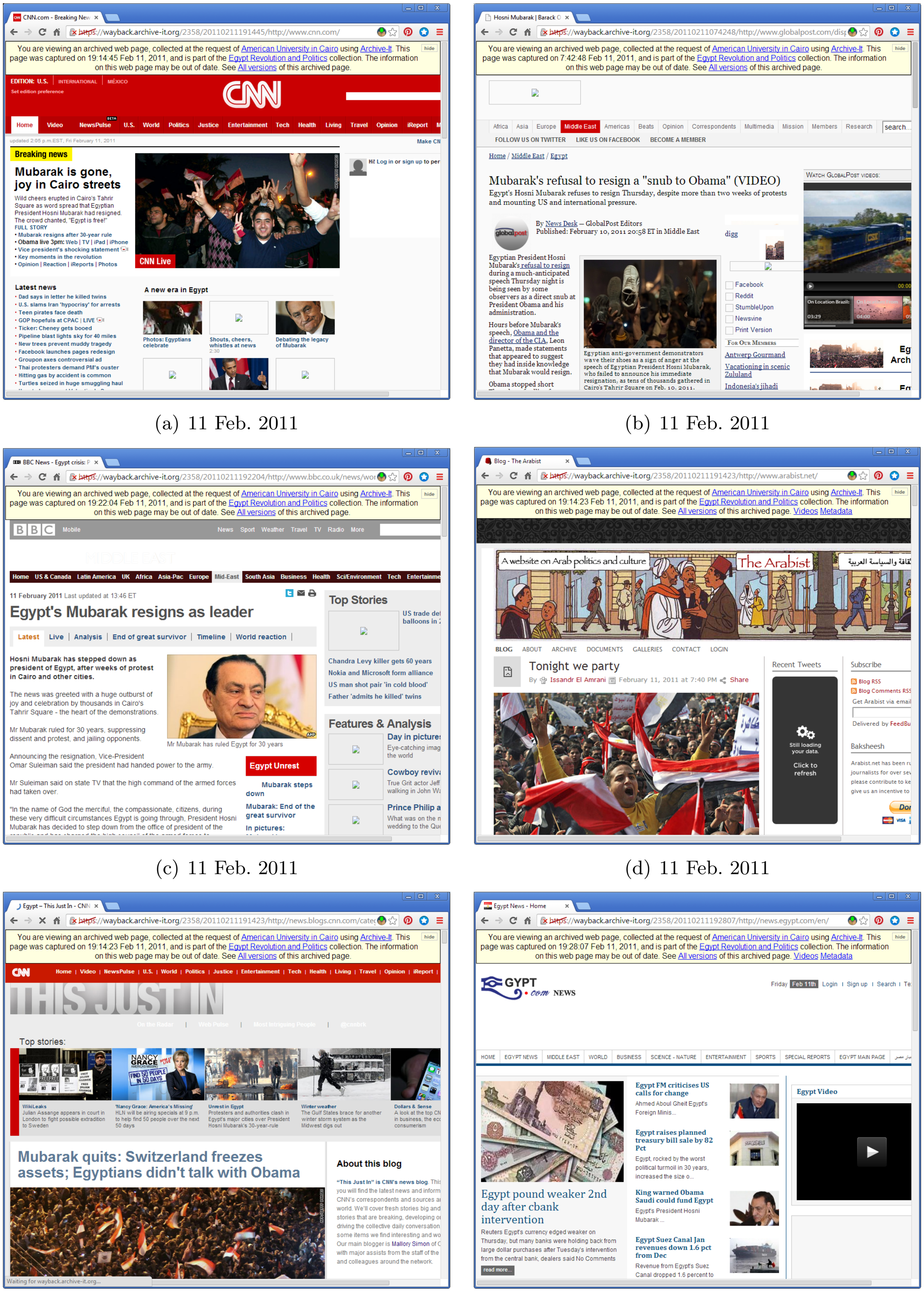
> Egyptian media, including *Al-Ahram*, falls under the authoritarian type, where the ruling regime and the elites monopolize media outlets. The authoritarian type indicates that journalism is subservient to the interests of the state in maintaining social order and achieving political goals. Saying that *Al-Ahram* is under the authoritarian type implies it avoids criticism to the President, the government policies or officials, and it censors publishing any material that challenges the established order.

Inside Egypt, the official newspapers did not cover the protests as they were happening. They were biased against the protests and supported Mubarak until he stepped down [218]. An example shown in Figure 15 contains the headline from Feb. 3, 2011 on the cover page of *Al-Ahram*, the most widely circulating state-owned daily newspaper and the second oldest newspaper in Egypt, founded in 1875. It reads “Millions go out in support of Mubarak” and has no news about the protests against Mubarak at that time.

A wide range of research has been conducted to study the media’s coverage of the Egyptian Revolution [110, 344, 256]. These studies discovered that the coverage by the governmental newspapers of the Egyptian demonstration differed from the international newspapers. Youssef Ahmed presented many examples for how *Al-Ahram* was prone to accentuate protesters’ acts of violence and published many articles to affect people’s opinions against the protests [344].

The pages shown in Figure 16 cover through multiple sites the reactions to Mubarak stepping down. Figure 16(f) shows the coverage of one of the national Egyptian newspaper on Feb. 11, 2011, the day when Mubarak stepped down. Although the page shows the reaction of Saudi Arabia on the Revolution and their support of Mubarak, it does not have any coverage for Mubarak stepping down.

To gain insight about a specific event, there is a need to know the date of the event. If a user wants to browse all the pages that were crawled on Feb. 11, 2011 (e.g., the pages in Figure 16), there is no way to do this with the current Archive-It interface. Our proposed DSA framework will provide the ability to create a story about a specific event from different resources with multiple perspectives at nearly the same time. This type of story will be important to social science and humanities researchers.



CNN.com - Breaking New

~~https~~://wayback.archive-it.org/2358/20110211191445/http://www.cnn.com/

You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 19:14:45 Feb 11, 2011, and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page.

EDITION: U.S. | INTERNATIONAL | MÉXICO

Set edition preference

Home | Video | NewsPulse | U.S. | World | Politics | Justice | Entertainment | Tech | Health | Living | Travel | Opinion | iReport

updated 2:05 p.m.EST, Fri February 11, 2011

Breaking news

Mubarak is gone, joy in Cairo streets

Wild cheers erupted in Cairo's Tahrir Square as word spread that Egyptian President Hosni Mubarak had resigned. The crowd chanted, "Egypt is free!" FULL STORY

- Mubarak resigns after 30-year rule
- Obama live 3pm: Web | TV | iPad | iPhone
- Vice president's shocking statement
- Key moments in the revolution
- Opinion | Reaction | iReports | Photos

CNN Live

Hi! Log in or sign up to per

Latest news

- Dad says in letter he killed twins
- U.S. slams Iran 'hypocrisy' for arrests
- Teen pirates face death
- GOP hopefuls at CPAC | LIVE
- Ticker: Cheney gets booed
- Pipeline blast lights sky for 40 miles
- New trees prevent muddy tragedy
- Facebook launches pages redesign
- Groupon axes controversial ad
- Thai protesters demand PM's ouster
- Hitting gas by accident is common
- Turtles seized in huge smuggling haul

A new era in Egypt

Photos: Egyptians celebrate

Shouts, cheers, whistles at news 2:30

Debating the legacy of Mubarak

(a) 11 Feb. 2011

Hosni Mubarak | Barack O

~~https~~://wayback.archive-it.org/2358/20110211074248/http://www.globalpost.com/disp

You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 7:42:48 Feb 11, 2011, and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page.

Africa | Asia | Europe | Middle East | Americas | Beats | Opinion | Correspondents | Multimedia | Mission | Members | Research

FOLLOW US ON TWITTER | LIKE US ON FACEBOOK | BECOME A MEMBER

Home / Middle East / Egypt

Mubarak's refusal to resign a "snub to Obama" (VIDEO)

Egypt's Hosni Mubarak refuses to resign Thursday, despite more than two weeks of protests and mounting US and international pressure.

By News Desk — GlobalPost Editors

Published: February 10, 2011 20:58 ET in Middle East

Egyptian President Hosni Mubarak's refusal to resign during a much-anticipated speech Thursday night is being seen by some observers as a direct snub at President Obama and his administration.

Hours before Mubarak's speech, Obama and the director of the CIA, Leon Panetta, made statements that appeared to suggest they had inside knowledge that Mubarak would resign.

Obama stopped short

Egyptian anti-government demonstrators wave their shoes as a sign of anger at the speech of Egyptian President Hosni Mubarak, who failed to announce his immediate resignation, as tens of thousands gathered in Cairo's Tahrir Square on Feb. 10, 2011.

digg

☐ Facebook
☐ Reddit
☐ StumbleUpon
☐ Newsvine
☐ Print Version

For Our Members

Antwerp Gourmand

Vacationing in scenic Zululand

Indonesia's jihadi

Watch GlobalPost videos:

On Location Brazil: 03:29

(b) 11 Feb. 2011

BBC News - Egypt crisis: P

~~https~~://wayback.archive-it.org/2358/20110211192204/http://www.bbc.co.uk/news/wor

You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 19:22:04 Feb 11, 2011, and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page.

BBC Mobile | News | Sport | Weather | Travel | TV | Radio | More

Home | US & Canada | Latin America | UK | Africa | Asia-Pac | Europe | Mid-East | South Asia | Business | Health | Sci/Environment | Tech | Entertainme

11 February 2011 Last updated at 13:46 ET

Egypt's Mubarak resigns as leader

Latest | Live | Analysis | End of great survivor | Timeline | World reaction |

Hosni Mubarak has stepped down as president of Egypt, after weeks of protest in Cairo and other cities.

The news was greeted with a huge outburst of joy and celebration by thousands in Cairo's Tahrir Square - the heart of the demonstrations.

Mr Mubarak ruled for 30 years, suppressing dissent and protest, and jailing opponents.

Mr Mubarak has ruled Egypt for 30 years

Announcing the resignation, Vice-President Omar Suleiman said the president had handed power to the army.

Mr Suleiman said on state TV that the high command of the armed forces had taken over.

"In the name of God the merciful, the compassionate, citizens, during these very difficult circumstances Egypt is going through, President Hosni Mubarak has decided to step down from the office of president of the

Egypt Unrest

Mubarak steps down

Mubarak: End of the great survivor

In pictures:

Top Stories

US trade def balloons in

Chandra Levy killer gets 60 years

Nokia and Microsoft form alliance

US man shot pair 'in cold blood'

Father 'admits he killed' twins

Features & Analysis

Day in picture Eye-catching imag the world

Cowboy reviva True Grit actor Jeff walking in John W

Prince Philip a What was on the m wedding to the Que

(c) 11 Feb. 2011

Blog - The Arabist

~~https~~://wayback.archive-it.org/2358/20110211191423/http://www.arabist.net/

You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 19:14:23 Feb 11, 2011, and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page. Videos Metadata

A website on Arab politics and culture | The Arabist | الثقافة والسياسة العربية

BLOG | ABOUT | ARCHIVE | DOCUMENTS | GALLERIES | CONTACT | LOGIN

Tonight we party

By Issandr El Amrani February 11, 2011 at 7:40 PM Share

Recent Tweets

Still loading your data. Click to refresh

Subscribe

Blog RSS

Blog Comments RSS

Get Arabist via email

Delivered by FeedBu

Baksheesh

Arabist.net has been ru journalists for over sev please contribute to ke give us an incentive to

(d) 11 Feb. 2011

Egypt – This Just In - CNN

~~https~~://wayback.archive-it.org/2358/20110211191423/http://news.blogs.cnn.com/cate

You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 19:14:23 Feb 11, 2011, and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page.

Home | Video | NewsPulse | U.S. | World | Politics | Justice | Entertainment | Tech | Health | Living | Travel | Opinion | iReport |

THIS JUST IN

On the Radar | Web Pulse | Most Intriguing People | @cnnbrk

Top stories:

WikiLeaks: Julian Assange appears in court in London to fight possible extradition to Sweden

'Nancy Grace: America's Missing': HLN will be airing specials at 9 p.m. to help find 50 people over the next 50 days

Unrest in Egypt: Protesters and authorities clash in Egypt's major cities over President Hosni Mubarak's 30-year-rule

Winter weather: The Gulf States brace for another winter storm system as the Midwest digs out

Dollars & Sense: A look at the top CN in business, the eco consumerism

Mubarak quits: Switzerland freezes assets; Egyptians didn't talk with Obama

About this blog

"This Just In" is CNN's news blog. This you will find the latest news and inform CNN's correspondents and sources a world. We'll cover fresh stories big and stories that are breaking, developing o driving the collective daily conversation some items we find interesting and wo Our main blogger is Mallory Simon of C with major assists from the staff of the and colleagues around the network.

Waiting for wayback.archive-it.org...

(e) 11 Feb. 2011

Egypt News - Home

~~https~~://wayback.archive-it.org/2358/20110211192807/http://news.egypt.com/en/

You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 19:28:07 Feb 11, 2011, and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page. Videos Metadata

GYPT .com NEWS

Friday Feb 11th Login | Sign up | Search | Te

HOME | EGYPT NEWS | MIDDLE EAST | WORLD | BUSINESS | SCIENCE - NATURE | ENTERTAINMENT | SPORTS | SPECIAL REPORTS | EGYPT MAIN PAGE | أخبار مصر

Egypt pound weaker 2nd day after cbank intervention

Reuters Egypt's currency edged weaker on Thursday, but many banks were holding back from large dollar purchases after Tuesday's intervention from the central bank, dealers said No Comments

read more...

Egypt FM criticises US calls for change

Ahmed Aboul Gheit Egypt's Foreign Minis...

Egypt raises planned treasury bill sale by 82 Pct

Egypt, rocked by the worst political turmoil in 30 years, increased the size o...

King warned Obama Saudi could fund Egypt

Egypt's President Hosni Mubarak ...

Egypt Suez Canal Jan revenues down 1.6 pct from Dec

Revenue from Egypt's Suez Canal dropped 1.6 percent to

Egypt Video

(f) 11 Feb. 2011

FIG. 16: Coverage of the Egyptian Revolution from different sites at a specific time (Feb. 11, 2011).

1.3.2 THE BOSTON MARATHON BOMBING

The Boston Marathon Bombing was a terrorist attack that occurred when two pressure cooker bombs exploded during the Boston Marathon on April 15, 2013 near the marathon's finish line on Boylston Street.

Figures 17 - 20 show different types of stories in which the events unfolded according to media at that time. The pages of the events are archived pages from the "2013 Boston Marathon Bombing" collection in Archive-It.

The Timeline of the Boston Marathon Bombing

- 2013-04-15: Two bombs exploded near the finish line of the Boston Marathon, killing three spectators and wounding over 100 people (Figure 17(a)).
- 2013-04-15: The investigation started directly by looking at explosive devices at the place of the incident (Figure 17(b)).
- 2013-04-16 and 2013-04-17: There were reactions to the attack such as President Obama's remarks about the explosions and how Muslims reacted (Figure 17(c) - 17(e)).
- 2013-04-18: The terrorists were identified and an intense manhunt occurred that shut down the Boston area ((Figure 17(f)).
- 2013-04-18: President Obama came to Boston for a memorial for the victims (Figure 18(a)).
- 2013-04-19: More information about the incident was released (Figure 18(b)).
- 2013-04-19: An emergency declaration for Massachusetts was issued by the governor and then an intensive manhunt followed (Figure 18(c)).
- 2013-04-19: Gunfire was heard in Watertown between the suspects and authorities who had tracked them, resulting in the death of one of the suspects and an injury for a police officer (Figure 18(d)).
- 2013-04-20 and 2013-04-21: The other suspect escaped, but was later found in a boat and was captured (Figure 18(e) - 19(f)).

The story in Figures 17 and 18 shows the coverage of the Boston Marathon Bombing from different Web sites over a broad range of times. This story has the broadest coverage because the diversity of the sources over a different times. The user gets an idea about collection and main events of the Boston Marathon Bombing from browsing the Web pages of Figures 17 and 18.

The Story of the Boston Marathon Bombing from *The Guardian*

Figure 19 shows two key events in the story of the Boston Marathon Bombing from the same Web site, *The Guardian*. Note that this illustrates a slightly different approach from the Egyptian Revolution story from CNN (Figure 14). The story of the Egyptian Revolution that was covered on CNN was created from the same URI (`http://news.blogs.cnn.com/category/world/egypt-world-latest-news/`) at different times. However, the Boston Marathon Bombing story of Figure 19 was created from different URIs but from the same domain.

The Coverage of the Boston Marathon Bombing on April 15

Figure 20 shows the coverage of Boston Marathon Bombing on April 15, 2013 from different Web sites. It shows how the newspapers covered this event and the reactions of different sites. We see that most of the reactions of the newspapers in Figure 20 are similar in covering the event. Some of the newspapers focused on reaction of Boston area baseball players and fans (Figure 20(a)), and others just focused on the incident and the numbers of victims (Figure 20(c) and 20(d)). Note that April 15, 2013 is the creation date of the Web pages that were used to create this story.

1.4 RESEARCH QUESTIONS

We will combine two existing tools in an innovative way. The goal of Archive-It is not necessarily crafting a story, but preserving content. The goal of Storify is not necessarily preserving content, but crafting a story. By combining Archive-It and Storify we can do both. The focus of this research is to explore information retrieval techniques to automatically generate stories summarizing a collection that will approximate what a knowledgeable human would generate. In other words, we will develop techniques to automatically (with optional human review and "steering") sample pages from a collection that summarize and describe the collection. For example, given a collection of 1000s of pages, our tool will automatically select approximately 28 representative pages that will then be linked in a storytelling web applications, such as Storify. Although page selection is not dependent on tools such as Storify, we are committed to the approach of using existing tools instead of developing new ones. It will also provide tools for collection curators to help them detecting when the seed URIs go off-topic.

The research in this dissertation addresses the following questions to understand the current challenges and to better construct a framework to solve them.

RQ1. How do people browse the past Web? One of the concerns in Web archiving world is how to generate more interest in using Web archives. To better understand the current use of Web archives, we investigate how users access Web archives based on Web access logs from the IA's Wayback Machine (Chapter 4). We check what users are looking for, why they come to the IA, where they come from, and how pages link to the IA. We also investigate the differences between human and robot accesses of the Wayback Machine, identify four major Web archive access patterns along with the browsing sessions' length, and uncover the temporal preference for Web archive access.

RQ2. Can we automatically generate stories that convey different perspectives of the collection? Different types of stories that give different perspectives of the topic of the collection can be generated from the collection. For example, the user may want to see a broadly defined story that samples from different URIs and different times, a story from different URIs at approximately the same time, or a story from the same URI at different times, and the same URI at the same time. We provide the definitions of the different types of stories that can be extracted from a collection (Chapter 5). We also provide definitions of an archived collection.

RQ3. How do we build quantitative, descriptive models of human-generated stories and collections in Archive-It? We need to understand the measurables of both stories and collections, as generated by humans, before we can automatically generate stories from archived collections. By sampling stories from Storify, we determine the characteristics of the human-generated stories such as the mean and median length of resources in the stories, the nature of the resources, how quickly the resources linked to from stories become unavailable, and the popularity of the resources linked to from stories (e.g., popular like `cnn.com` or little-known outlets, blogs, and other sites). We establish structural features for what differentiates popular stories from unpopular stories for building a baseline for the stories we will automatically generate from the archives (Chapter 6).

We also determine the characteristics of Archive-It collections through measurements of all Archive-It collections such as the number of URIs, the number of mementos, the most used resources in these collections, the average timespan of the collections, etc. (Chapter 7). In summarizing a collection, we can only choose from what is archived. For example, if there are no tweets in the collection, `twitter.com` will not appear in the generated stories.

We compare the descriptive models of the created stories on social media and the collections in Archive-It to understand the similarities and the differences between the human-generated stories in social media and the human-curated collections in archives.

RQ4. How to detect the off-topic Web pages in the archives? In our DSA framework, we do not consider off-topic pages for selection when creating a story. We propose different methods (cosine similarity, Jaccard similarity, intersection of the 20 most frequent terms, Web-based kernel function, and the change in size using number of words and content length) to detect when the page has gone off-topic through subsequent captures (Chapter 8). Those predicted off-topic pages will be presented to the collection's curator for possible elimination from the collection or cessation of crawling.

RQ5. How do we identify, evaluate, and select candidate (archived) Web pages for the story? Choosing the best representative pages for a story is a challenge. There are multiple dimensions of quality when we have multiple candidates for the same event in a story, such as quality of the replayed archived page or quality of the visual snippet that will be generated. We propose quality metrics for evaluating Web pages and then select the page with the highest weight (Chapter 9).

There are also important factors that affect the quality of the generated summary story. For example, for the broad story, the selection should cover the time range of the collection equally. We propose a slicing algorithm that allows equally covering all the parts of the collections through time. We leverage a storytelling service, Storify, to visualize the set of selected pages.

1.5 DISSERTATION ROADMAP

This dissertation is organized as follows:

Chapter 2: Background - Before discussing our specific work toward establishing the DSA framework, we explore the evolution of Web archives and existing content curation tools. We also explain the related terminology that is adopted in the following chapters, such as the Memento terminology.

Chapter 3: Related Work - The work performed by prior researchers that provided the foundation for our work is presented in this chapter. We first present how the archival community tended to solve the collection understanding problem through the development of a new standard for archival description (Section 3.1). Then we overview many related fields such as summarizing text and image collections, video summarization, time series visualizations, information retrieval techniques, Web usage mining technique, and the different notions of time of a Web page.

Chapter 4: How People Use Web Archives - Our work toward understanding how people use Web archives, why do they come to Web archives, how people link to Web archives, how many of links that users access disappear from the live Web, and many other questions related to the usage of Web archives is presented in this chapter.

Chapter 5: Generating Stories From Archived Collections - This chapter contains the contextualization of our work along with the terminology and definitions that represent the basics of the DSA framework. It also discusses the research methodology for achieving the proposed framework. We also present the abstract idea of the DSA framework along with the framework components that will be detailed in later chapters.

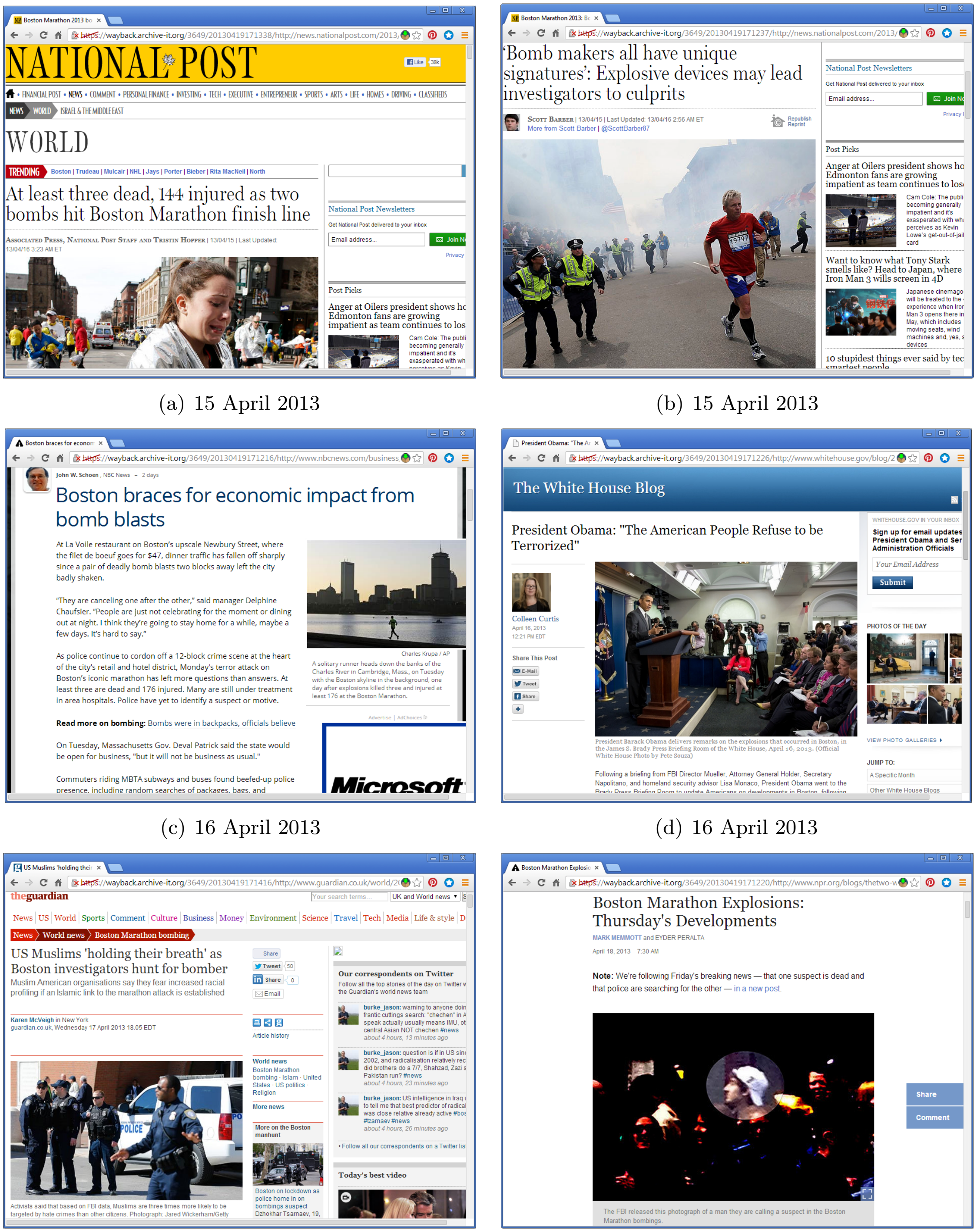
Chapter 6: Characteristics of Social Media Stories - To inform our work of generating stories from archived collections automatically, we studied 14,568 stories from Storify, comprising 1,251,160 individual resources. We modeled the structural characteristics of these stories, with particular emphasis on "popular" stories (i.e., the top 25% of views, normalized by time available on the Web). We checked the domain used in the stories, the types of elements, and the number of elements. We also investigated how many resources in the stories are missing from the live Web and how many are available in public Web archives.

Chapter 7: Characteristics of Archive-It Collections - We present a baseline for the characteristics of the archived collections using the whole population of Archive-It collections in terms of the number of seed URIs, the average number of the mementos per seed, and the timespan, which is the range of time period over which the Web pages have been archived. Furthermore, we contrast the general characteristics of human-generated stories from Storify that were presented in Chapter 6 and human-curated collections from Archive-It.

Chapter 8: Detecting Off-Topic Pages in Web Archives - We introduce different approaches for detecting off-topic pages in Web archives. The approaches depend on comparing the versions of the pages through time. Three methods depend on the textual content (cosine similarity, intersection of the most frequent terms, and Jaccard coefficient), one method uses the semantics of the text (Web-based kernel function using the search engine (SE)), and two methods use the size of pages (the change in number of words and the content length). We also investigate how the page's aboutness changes through time based on a dataset from Archive-It. We evaluate the proposed methods at different thresholds at the end of the chapter.

Chapter 9: Selecting Representative Pages for Generating Stories - This chapter shows the steps of selecting representative archived pages for the stories. It starts with eliminating the duplicates of each page (Algorithm 1), then slicing the collection dynamically based on the collection size (Algorithm 2). We then cluster the archived pages in each slice based on their content. The chapter also includes the quality metrics we used for selecting pages that best represent the story. Furthermore, we show how the extraction of page's metadata is done and how we visualize the selected pages using Storify. At the end of the chapter, we present an evaluation inspired by the Turing Test for the automatically generated stories from archived collections. We obtained a set of human-generated stories by domain experts of the collections. We used human evaluation (e.g., Mechanical Turk) to see if the resulting stories are distinguishable from human generated stories.

Chapter 10: Contributions, Future Work, and Conclusions - We revisit the research questions we introduced in Chapter 1 and summarize how we addressed each question, including the contributions of this dissertation. Directions for future work are presented. We conclude with a summary of our findings.



Boston Marathon 2013 bo
https://wayback.archive-it.org/3649/20130419171338/http://news.nationalpost.com/2013/
NATIONAL POST
Like 38k
FINANCIAL POST • NEWS • COMMENT • PERSONAL FINANCE • INVESTING • TECH • EXECUTIVE • ENTREPRENEUR • SPORTS • ARTS • LIFE • HOMES • DRIVING • CLASSIFIEDS
NEWS WORLD ISRAEL & THE MIDDLE EAST
WORLD
TRENDING Boston | Trudeau | Mulcair | NHL | Jays | Porter | Bieber | Rita MacNeil | North
National Post Newsletters
Get National Post delivered to your inbox
Email address...
Join N
Privacy
Post Picks

At least three dead, 144 injured as two bombs hit Boston Marathon finish line

Associated Press, National Post Staff and Tristin Hopper | 13/04/15 | Last Updated: 13/04/16 3:23 AM ET

Anger at Oilers president shows ho Edmonton fans are growing impatient as team continues to los

Cam Cole: The publi becoming generally impatient and it's exasperated with wh

(a) 15 April 2013

Boston Marathon 2013: Bo
https://wayback.archive-it.org/3649/20130419171237/http://news.nationalpost.com/2013/
Republish Reprint
National Post Newsletters
Get National Post delivered to your inbox
Email address...
Join N
Privacy
Post Picks

'Bomb makers all have unique signatures': Explosive devices may lead investigators to culprits

Scott Barber | 13/04/15 | Last Updated: 13/04/16 2:56 AM ET
More from Scott Barber | @ScottBarber87

Anger at Oilers president shows ho Edmonton fans are growing impatient as team continues to los

Cam Cole: The publi becoming generally impatient and it's exasperated with wh perceives as Kevin Lowe's get-out-of-jail card

Want to know what Tony Stark smells like? Head to Japan, where Iron Man 3 wills screen in 4D

Japanese cinemago will be treated to the experience when Iror Man 3 opens there in May, which includes moving seats, wind machines and, yes, devices

10 stupidest things ever said by tec smartest people

(b) 15 April 2013

Boston braces for econon
https://wayback.archive-it.org/3649/20130419171216/http://www.nbcnews.com/business
Advertise | AdChoices
Microsoft

John W. Schoen, NBC News – 2 days

Boston braces for economic impact from bomb blasts

At La Voile restaurant on Boston's upscale Newbury Street, where the filet de boeuf goes for $47, dinner traffic has fallen off sharply since a pair of deadly bomb blasts two blocks away left the city badly shaken.

"They are canceling one after the other," said manager Delphine Chaufsier. "People are just not celebrating for the moment or dining out at night. I think they're going to stay home for a while, maybe a few days. It's hard to say."

As police continue to cordon off a 12-block crime scene at the heart of the city's retail and hotel district, Monday's terror attack on Boston's iconic marathon has left more questions than answers. At least three are dead and 176 injured. Many are still under treatment in area hospitals. Police have yet to identify a suspect or motive.

Read more on bombing: Bombs were in backpacks, officials believe

On Tuesday, Massachusetts Gov. Deval Patrick said the state would be open for business, "but it will not be business as usual."

Commuters riding MBTA subways and buses found beefed-up police presence, including random searches of packages, bags, and

Charles Krupa / AP
A solitary runner heads down the banks of the Charles River in Cambridge, Mass., on Tuesday with the Boston skyline in the background, one day after explosions killed three and injured at least 176 at the Boston Marathon.

(c) 16 April 2013

President Obama: "The Ar
https://wayback.archive-it.org/3649/20130419171226/http://www.whitehouse.gov/blog/2
The White House Blog
Share This Post
E-Mail
Tweet
Share
WHITEHOUSE.GOV IN YOUR INBOX
Sign up for email updates President Obama and Ser Administration Officials
Your Email Address
Submit
PHOTOS OF THE DAY
VIEW PHOTO GALLERIES
JUMP TO:
A Specific Month
Other White House Blogs

President Obama: "The American People Refuse to be Terrorized"

Colleen Curtis
April 16, 2013
12:21 PM EDT

President Barack Obama delivers remarks on the explosions that occurred in Boston, in the James S. Brady Press Briefing Room of the White House, April 16, 2013. (Official White House Photo by Pete Souza)

Following a briefing from FBI Director Mueller, Attorney General Holder, Secretary Napolitano, and homeland security advisor Lisa Monaco, President Obama went to the Brady Press Briefing Room to update Americans on developments in Boston, following

(d) 16 April 2013

US Muslims 'holding their
https://wayback.archive-it.org/3649/20130419171416/http://www.guardian.co.uk/world/2
theguardian
Your search terms...
UK and World news
News | US | World | Sports | Comment | Culture | Business | Money | Environment | Science | Travel | Tech | Media | Life & style | D
News World news Boston Marathon bombing
Share
Tweet 50
Share 0
Email
Article history
World news
Boston Marathon bombing · Islam · United States · US politics · Religion
More news
More on the Boston manhunt
Today's best video

US Muslims 'holding their breath' as Boston investigators hunt for bomber

Muslim American organisations say they fear increased racial profiling if an Islamic link to the marathon attack is established

Karen McVeigh in New York
guardian.co.uk, Wednesday 17 April 2013 18.05 EDT

Activists said that based on FBI data, Muslims are three times more likely to be targeted by hate crimes than other citizens. Photograph: Jared Wickerham/Getty

Boston on lockdown as police home in on bombings suspect Dzhokhar Tsarnaev, 19,

Our correspondents on Twitter
Follow all the top stories of the day on Twitter w the Guardian's world news team

burke_jason: warning to anyone doin frantic cuttings search: "chechen" in A speak actually usually means IMU, ot central Asian NOT chechen #news
about 4 hours, 13 minutes ago

burke_jason: question is if in US sinc 2002, and radicalisation relatively rec did brothers do a 7/7, Shahzad, Zazi s Pakistan run? #news
about 4 hours, 23 minutes ago

burke_jason: US intelligence in Iraq u to tell me that best predictor of radical was close relative already active #bos #tzarnaev #news
about 4 hours, 26 minutes ago

• Follow all our correspondents on a Twitter lis

(e) 17 April 2013

Boston Marathon Explosio
https://wayback.archive-it.org/3649/20130419171220/http://www.npr.org/blogs/thetwo-w
Share
Comment

Boston Marathon Explosions: Thursday's Developments

MARK MEMMOTT and EYDER PERALTA

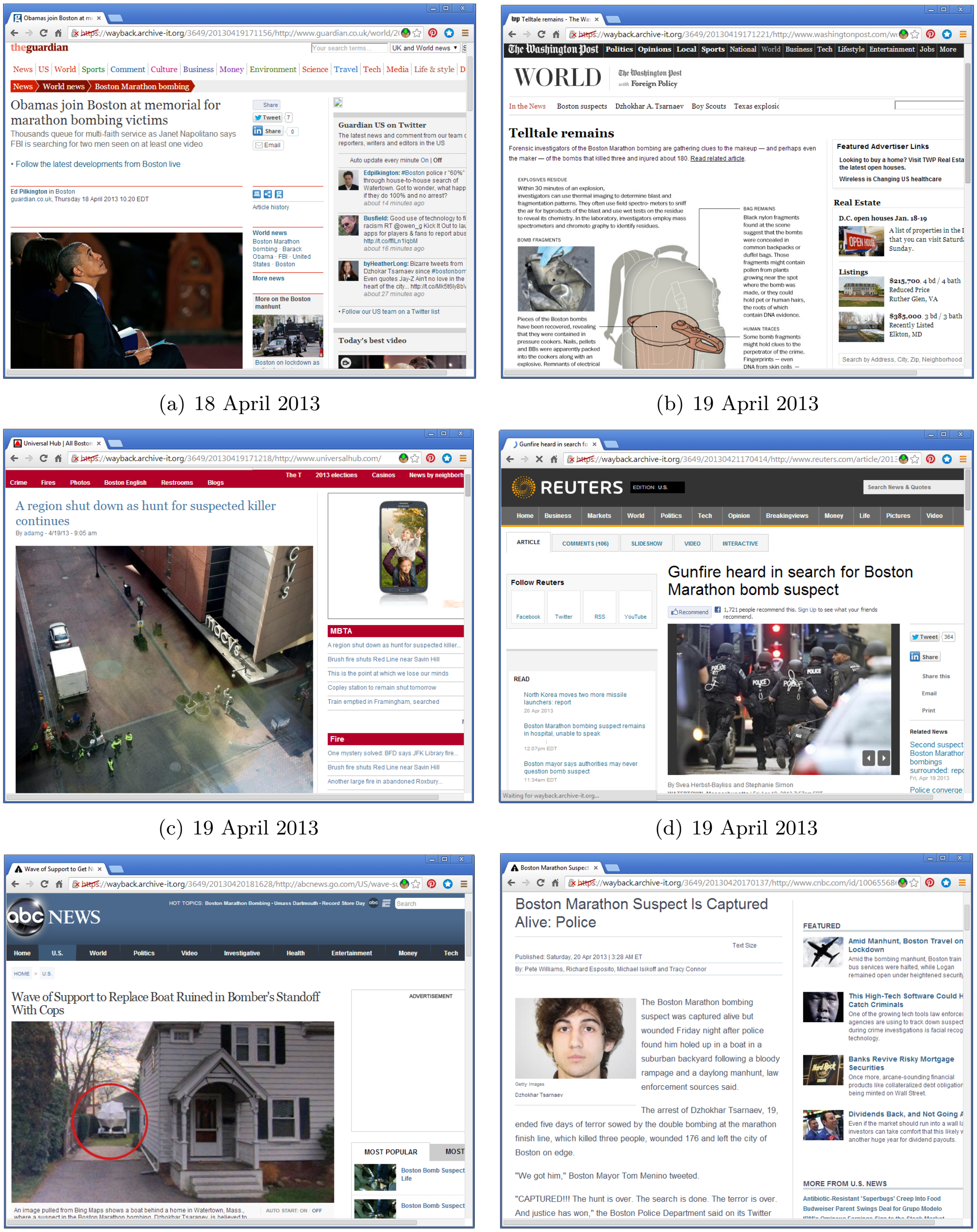
April 18, 2013 7:30 AM

Note: We're following Friday's breaking news — that one suspect is dead and that police are searching for the other — in a new post.

The FBI released this photograph of a man they are calling a suspect in the Boston Marathon bombings.

(f) 18 April 2013

FIG. 17: Coverage of the Boston Marathon Bombing from different Web sites at different times.



Obamas join Boston at m
~~https~~://wayback.archive-it.org/3649/20130419171156/http://www.guardian.co.uk/world/2

theguardian

News | US | World | Sports | Comment | Culture | Business | Money | Environment | Science | Travel | Tech | Media | Life & style | D

News › World news › Boston Marathon bombing

Obamas join Boston at memorial for marathon bombing victims

Thousands queue for multi-faith service as Janet Napolitano says FBI is searching for two men seen on at least one video

• Follow the latest developments from Boston live

Ed Pilkington in Boston
guardian.co.uk, Thursday 18 April 2013 10.20 EDT

Article history

World news
Boston Marathon bombing · Barack Obama · FBI · United States · Boston

More news

More on the Boston manhunt

Boston on lockdown as

Guardian US on Twitter

The latest news and comment from our team c reporters, writers and editors in the US

Auto update every minute On | Off

Edpilkington: #Boston police r "60%" through house-to-house search of Watertown. Got to wonder, what happ if they do 100% and no arrest?
about 14 minutes ago

Busfield: Good use of technology to fi racism RT @owen_g Kick It Out to lau apps for players & fans to report abus http://t.co/ffILn1IqbM
about 16 minutes ago

byHeatherLong: Bizarre tweets from Dzhokar Tsarnaev since #bostonbom Even quotes Jay-Z Ain't no love in the heart of the city... http://t.co/Mk5t6ly8bV
about 27 minutes ago

• Follow our US team on a Twitter list

Today's best video

(a) 18 April 2013

Telltale remains - The Wa
~~https~~://wayback.archive-it.org/3649/20130419171221/http://www.washingtonpost.com/w

The Washington Post | Politics | Opinions | Local | Sports | National | World | Business | Tech | Lifestyle | Entertainment | Jobs | More

WORLD | The Washington Post with Foreign Policy

In the News | Boston suspects | Dzhokhar A. Tsarnaev | Boy Scouts | Texas explosic

Telltale remains

Forensic investigators of the Boston Marathon bombing are gathering clues to the makeup — and perhaps even the maker — of the bombs that killed three and injured about 180. Read related article.

EXPLOSIVES RESIDUE
Within 30 minutes of an explosion, investigators can use thermal imaging to determine blast and fragmentation patterns. They often use field spectro- meters to sniff the air for byproducts of the blast and use wet tests on the residue to reveal its chemistry. In the laboratory, investigators employ mass spectrometers and chromoto graphy to identify residues.

BOMB FRAGMENTS
Pieces of the Boston bombs have been recovered, revealing that they were contained in pressure cookers. Nails, pellets and BBs were apparently packed into the cookers along with an explosive. Remnants of electrical

BAG REMAINS
Black nylon fragments found at the scene suggest that the bombs were concealed in common backpacks or duffel bags. Those fragments might contain pollen from plants growing near the spot where the bomb was made, or they could hold pet or human hairs, the roots of which contain DNA evidence.

HUMAN TRACES
Some bomb fragments might hold clues to the perpetrator of the crime. Fingerprints — even DNA from skin cells —

Featured Advertiser Links
Looking to buy a home? Visit TWP Real Esta the latest open houses.
Wireless is Changing US healthcare

Real Estate
D.C. open houses Jan. 18-19
A list of properties in the I that you can visit Saturd Sunday.

Listings
$215,700, 4 bd / 4 bath
Reduced Price
Ruther Glen, VA

$385,000, 3 bd / 3 bath
Recently Listed
Elkton, MD

Search by Address, City, Zip, Neighborhood

(b) 19 April 2013

Universal Hub | All Boston
~~https~~://wayback.archive-it.org/3649/20130419171218/http://www.universalhub.com/

Crime | Fires | Photos | Boston English | Restrooms | Blogs | The T | 2013 elections | Casinos | News by neighborh

A region shut down as hunt for suspected killer continues

By adamg - 4/19/13 - 9:05 am

MBTA
A region shut down as hunt for suspected killer...
Brush fire shuts Red Line near Savin Hill
This is the point at which we lose our minds
Copley station to remain shut tomorrow
Train emptied in Framingham, searched

Fire
One mystery solved: BFD says JFK Library fire...
Brush fire shuts Red Line near Savin Hill
Another large fire in abandoned Roxbury...

(c) 19 April 2013

Gunfire heard in search fo
~~https~~://wayback.archive-it.org/3649/20130421170414/http://www.reuters.com/article/2013

REUTERS | EDITION: U.S. | Search News & Quotes

Home | Business | Markets | World | Politics | Tech | Opinion | Breakingviews | Money | Life | Pictures | Video

ARTICLE | COMMENTS (106) | SLIDESHOW | VIDEO | INTERACTIVE

Follow Reuters
Facebook | Twitter | RSS | YouTube

READ
North Korea moves two more missile launchers: report
20 Apr 2013
Boston Marathon bombing suspect remains in hospital, unable to speak
12:07pm EDT
Boston mayor says authorities may never question bomb suspect
11:34am EDT

Gunfire heard in search for Boston Marathon bomb suspect

Recommend 1,721 people recommend this. Sign Up to see what your friends recommend.

By Svea Herbst-Bayliss and Stephanie Simon

Tweet 364
Share
Share this
Email
Print

Related News
Second suspect Boston Marathor bombings surrounded: repo
Fri, Apr 19 2013
Police converge

Waiting for wayback.archive-it.org...

(d) 19 April 2013

Wave of Support to Get N
~~https~~://wayback.archive-it.org/3649/20130420181628/http://abcnews.go.com/US/wave-su

abc NEWS
HOT TOPICS: Boston Marathon Bombing · Umass Dartmouth · Record Store Day | Search

Home | U.S. | World | Politics | Video | Investigative | Health | Entertainment | Money | Tech

HOME > U.S.

Wave of Support to Replace Boat Ruined in Bomber's Standoff With Cops

An image pulled from Bing Maps shows a boat behind a home in Watertown, Mass., where a suspect in the Boston Marathon bombing, Dzhokhar Tsaranev, is believed to

AUTO START: ON | OFF

ADVERTISEMENT

MOST POPULAR | MOST
Boston Bomb Suspect Life
Boston Bomb Suspect

(e) 20 April 2013

Boston Marathon Suspect
~~https~~://wayback.archive-it.org/3649/20130420170137/http://www.cnbc.com/id/10065568

Boston Marathon Suspect Is Captured Alive: Police

Text Size

Published: Saturday, 20 Apr 2013 | 3:28 AM ET

By: Pete Williams, Richard Esposito, Michael Isikoff and Tracy Connor

Getty Images
Dzhokhar Tsarnaev

The Boston Marathon bombing suspect was captured alive but wounded Friday night after police found him holed up in a boat in a suburban backyard following a bloody rampage and a daylong manhunt, law enforcement sources said.

The arrest of Dzhokhar Tsarnaev, 19, ended five days of terror sowed by the double bombing at the marathon finish line, which killed three people, wounded 176 and left the city of Boston on edge.

"We got him," Boston Mayor Tom Menino tweeted.

"CAPTURED!!! The hunt is over. The search is done. The terror is over. And justice has won," the Boston Police Department said on its Twitter

FEATURED

Amid Manhunt, Boston Travel on Lockdown
Amid the bombing manhunt, Boston train bus services were halted, while Logan remained open under heightened security

This High-Tech Software Could H Catch Criminals
One of the growing tech tools law enforce agencies are using to track down suspect during crime investigations is facial recog technology.

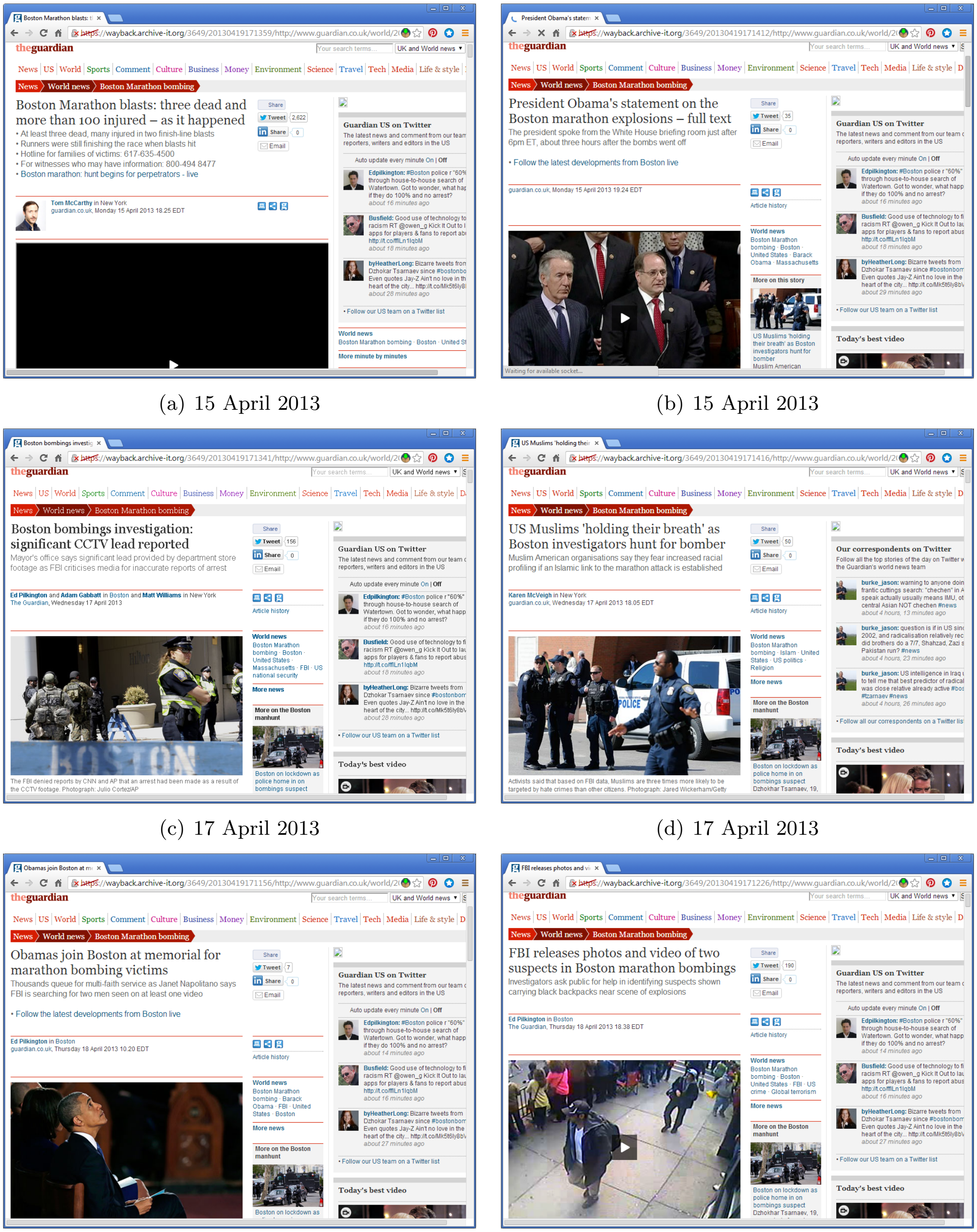
Banks Revive Risky Mortgage Securities
Once more, arcane-sounding financial products like collateralized debt obligation being minted on Wall Street.

Dividends Back, and Not Going A
Even if the market should run into a wall la investors can take comfort that this likely w another huge year for dividend payouts.

MORE FROM U.S. NEWS
Antibiotic-Resistant 'Superbugs' Creep Into Food
Budweiser Parent Swings Deal for Grupo Modelo

(f) 21 April 2013

FIG. 18: Coverage of the Boston Marathon Bombing from different Web sites at different times (continued).


theguardian
Your search terms...
UK and World news
News | US | World | Sports | Comment | Culture | Business | Money | Environment | Science | Travel | Tech | Media | Life & style
News › World news › Boston Marathon bombing
Share
Tweet
Share
Email
Article history
Today's best video

Boston Marathon blasts: three dead and more than 100 injured – as it happened

- At least three dead, many injured in two finish-line blasts
- Runners were still finishing the race when blasts hit
- Hotline for families of victims: 617-635-4500
- For witnesses who may have information: 800-494 8477
- Boston marathon: hunt begins for perpetrators - live

Tom McCarthy in New York
guardian.co.uk, Monday 15 April 2013 18.25 EDT

Guardian US on Twitter
The latest news and comment from our team of reporters, writers and editors in the US

Auto update every minute On | Off

Edpilkington: #Boston police r "60%" through house-to-house search of Watertown. Got to wonder, what hap if they do 100% and no arrest?
about 16 minutes ago

Busfield: Good use of technology to racism RT @owen_g Kick It Out to l apps for players & fans to report ab http://t.co/fflLn1IqbM
about 18 minutes ago

byHeatherLong: Bizarre tweets from Dzhokar Tsarnaev since #bostonbo Even quotes Jay-Z Ain't no love in th heart of the city... http://t.co/Mk5t6ly8
about 28 minutes ago

• Follow our US team on a Twitter list

World news
Boston Marathon bombing · Boston · United St
More minute by minutes

(a) 15 April 2013

President Obama's statement on the Boston marathon explosions – full text

The president spoke from the White House briefing room just after 6pm ET, about three hours after the bombs went off

• Follow the latest developments from Boston live

guardian.co.uk, Monday 15 April 2013 19.24 EDT

World news
Boston Marathon bombing · Boston · United States · Barack Obama · Massachusetts

More on this story
US Muslims 'holding their breath' as Boston investigators hunt for bomber
Muslim American

Guardian US on Twitter
The latest news and comment from our team of reporters, writers and editors in the US

Auto update every minute On | Off

Edpilkington: #Boston police r "60%" through house-to-house search of Watertown. Got to wonder, what happ if they do 100% and no arrest?
about 16 minutes ago

Busfield: Good use of technology to fi racism RT @owen_g Kick It Out to lau apps for players & fans to report abus http://t.co/fflLn1IqbM
about 18 minutes ago

byHeatherLong: Bizarre tweets from Dzhokar Tsarnaev since #bostonbom Even quotes Jay-Z Ain't no love in the heart of the city... http://t.co/Mk5t6ly8b
about 29 minutes ago

• Follow our US team on a Twitter list

Waiting for available socket...

(b) 15 April 2013

Boston bombings investigation: significant CCTV lead reported

Mayor's office says significant lead provided by department store footage as FBI criticises media for inaccurate reports of arrest

Ed Pilkington and Adam Gabbatt in Boston and Matt Williams in New York
The Guardian, Wednesday 17 April 2013

The FBI denied reports by CNN and AP that an arrest had been made as a result of the CCTV footage. Photograph: Julio Cortez/AP

World news
Boston Marathon bombing · Boston · United States · Massachusetts · FBI · US national security
More news

More on the Boston manhunt
Boston on lockdown as police home in on bombings suspect

Guardian US on Twitter
The latest news and comment from our team of reporters, writers and editors in the US

Auto update every minute On | Off

Edpilkington: #Boston police r "60%" through house-to-house search of Watertown. Got to wonder, what happ if they do 100% and no arrest?
about 16 minutes ago

Busfield: Good use of technology to fi racism RT @owen_g Kick It Out to lau apps for players & fans to report abus http://t.co/fflLn1IqbM
about 18 minutes ago

byHeatherLong: Bizarre tweets from Dzhokar Tsarnaev since #bostonbom Even quotes Jay-Z Ain't no love in the heart of the city... http://t.co/Mk5t6ly8b
about 28 minutes ago

• Follow our US team on a Twitter list

(c) 17 April 2013

US Muslims 'holding their breath' as Boston investigators hunt for bomber

Muslim American organisations say they fear increased racial profiling if an Islamic link to the marathon attack is established

Karen McVeigh in New York
guardian.co.uk, Wednesday 17 April 2013 18.05 EDT

Activists said that based on FBI data, Muslims are three times more likely to be targeted by hate crimes than other citizens. Photograph: Jared Wickerham/Getty

World news
Boston Marathon bombing · Islam · United States · US politics · Religion
More news

More on the Boston manhunt
Boston on lockdown as police home in on bombings suspect
Dzhokhar Tsarnaev, 19,

Our correspondents on Twitter
Follow all the top stories of the day on Twitter w the Guardian's world news team

burke_jason: warning to anyone doin frantic cuttings search: "chechen" in A speak actually usually means IMU, ot central Asian NOT chechen #news
about 4 hours, 13 minutes ago

burke_jason: question is if in US sinc 2002, and radicalisation relatively rec did brothers do a 7/7, Shahzad, Zazi s Pakistan run? #news
about 4 hours, 23 minutes ago

burke_jason: US intelligence in Iraq u to tell me that best predictor of radical was close relative already active #bos #tzarnaev #news
about 4 hours, 26 minutes ago

• Follow all our correspondents on a Twitter lis

(d) 17 April 2013

Obamas join Boston at memorial for marathon bombing victims

Thousands queue for multi-faith service as Janet Napolitano says FBI is searching for two men seen on at least one video

• Follow the latest developments from Boston live

Ed Pilkington in Boston
guardian.co.uk, Thursday 18 April 2013 10.20 EDT

World news
Boston Marathon bombing · Barack Obama · FBI · United States · Boston
More news

More on the Boston manhunt
Boston on lockdown as

Guardian US on Twitter
The latest news and comment from our team of reporters, writers and editors in the US

Auto update every minute On | Off

Edpilkington: #Boston police r "60%" through house-to-house search of Watertown. Got to wonder, what happ if they do 100% and no arrest?
about 14 minutes ago

Busfield: Good use of technology to fi racism RT @owen_g Kick It Out to lau apps for players & fans to report abus http://t.co/fflLn1IqbM
about 16 minutes ago

byHeatherLong: Bizarre tweets from Dzhokar Tsarnaev since #bostonbom Even quotes Jay-Z Ain't no love in the heart of the city... http://t.co/Mk5t6ly8b
about 27 minutes ago

• Follow our US team on a Twitter list

(e) 18 April 2013

FBI releases photos and video of two suspects in Boston marathon bombings

Investigators ask public for help in identifying suspects shown carrying black backpacks near scene of explosions

Ed Pilkington in Boston
The Guardian, Thursday 18 April 2013 18.38 EDT

World news
Boston Marathon bombing · Boston · United States · FBI · US crime · Global terrorism
More news

More on the Boston manhunt
Boston on lockdown as police home in on bombings suspect
Dzhokhar Tsarnaev, 19,

Guardian US on Twitter
The latest news and comment from our team of reporters, writers and editors in the US

Auto update every minute On | Off

Edpilkington: #Boston police r "60%" through house-to-house search of Watertown. Got to wonder, what happ if they do 100% and no arrest?
about 14 minutes ago

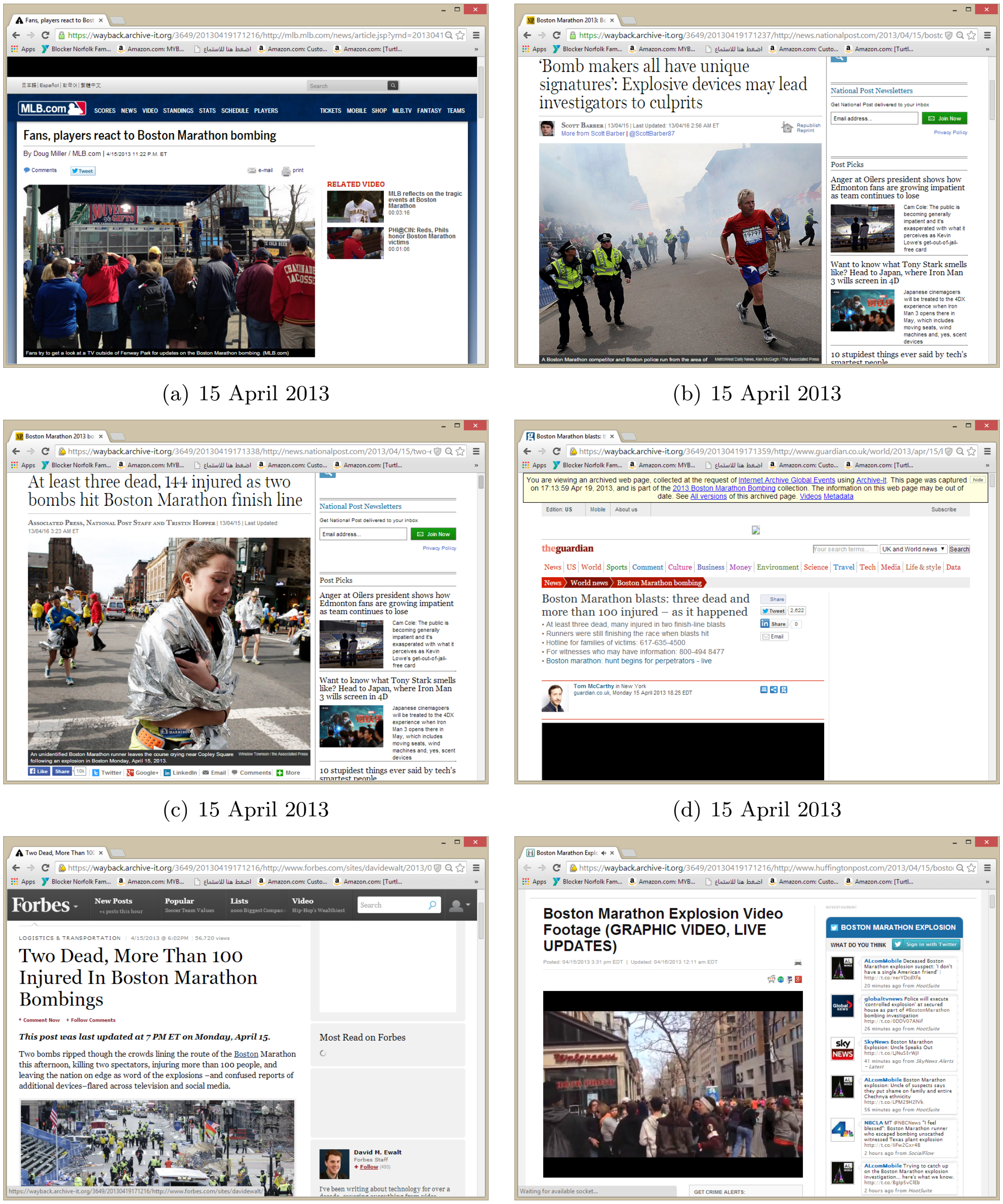
Busfield: Good use of technology to fi racism RT @owen_g Kick It Out to lau apps for players & fans to report abus http://t.co/fflLn1IqbM
about 16 minutes ago

byHeatherLong: Bizarre tweets from Dzhokar Tsarnaev since #bostonbom Even quotes Jay-Z Ain't no love in the heart of the city... http://t.co/Mk5t6ly8b
about 27 minutes ago

• Follow our US team on a Twitter list

(f) 18 April 2013

FIG. 19: Coverage of the Boston Marathon Bombing from *the Guardian* at different times.



Fans, players react to Boston Marathon bombing

By Doug Miller / MLB.com | 4/15/2013 11:22 P.M. ET

Fans try to get a look at a TV outside of Fenway Park for updates on the Boston Marathon bombing. (MLB.com)

RELATED VIDEO

MLB reflects on the tragic events at Boston Marathon 00:03:16

PHI@CIN: Reds, Phils honor Boston Marathon victims 00:01:06

(a) 15 April 2013

'Bomb makers all have unique signatures': Explosive devices may lead investigators to culprits

Scott Barber | 13/04/15 | Last Updated: 13/04/16 2:56 AM ET
More from Scott Barber | @ScottBarber87

A Boston Marathon competitor and Boston police run from the area of MetroWest Daily News, Ken McGagh / The Associated Press

National Post Newsletters

Get National Post delivered to your inbox

Post Picks

Anger at Oilers president shows how Edmonton fans are growing impatient as team continues to lose

Cam Cole: The public is becoming generally impatient and it's exasperated with what it perceives as Kevin Lowe's get-out-of-jail-free card

Want to know what Tony Stark smells like? Head to Japan, where Iron Man 3 wills screen in 4D

Japanese cinemagoers will be treated to the 4DX experience when Iron Man 3 opens there in May, which includes moving seats, wind machines and, yes, scent devices

10 stupidest things ever said by tech's smartest people

(b) 15 April 2013

At least three dead, 144 injured as two bombs hit Boston Marathon finish line

Associated Press, National Post Staff and Tristin Hopper | 13/04/15 | Last Updated: 13/04/16 3:23 AM ET

An unidentified Boston Marathon runner leaves the course crying near Copley Square following an explosion in Boston Monday, April 15, 2013. Winslow Townson / the Associated Press

National Post Newsletters

Get National Post delivered to your inbox

Post Picks

Anger at Oilers president shows how Edmonton fans are growing impatient as team continues to lose

Cam Cole: The public is becoming generally impatient and it's exasperated with what it perceives as Kevin Lowe's get-out-of-jail-free card

Want to know what Tony Stark smells like? Head to Japan, where Iron Man 3 wills screen in 4D

Japanese cinemagoers will be treated to the 4DX experience when Iron Man 3 opens there in May, which includes moving seats, wind machines and, yes, scent devices

10 stupidest things ever said by tech's smartest people

(c) 15 April 2013

You are viewing an archived web page, collected at the request of Internet Archive Global Events using Archive-It. This page was captured on 17:13:59 Apr 19, 2013, and is part of the 2013 Boston Marathon Bombing collection. The information on this web page may be out of date. See All versions of this archived page. Videos Metadata

News › World news › Boston Marathon bombing

Boston Marathon blasts: three dead and more than 100 injured – as it happened

- At least three dead, many injured in two finish-line blasts
- Runners were still finishing the race when blasts hit
- Hotline for families of victims: 617-635-4500
- For witnesses who may have information: 800-494 8477
- Boston marathon: hunt begins for perpetrators - live

Tom McCarthy in New York
guardian.co.uk, Monday 15 April 2013 18.25 EDT

(d) 15 April 2013

LOGISTICS & TRANSPORTATION | 4/15/2013 @ 6:02PM | 56,720 views

Two Dead, More Than 100 Injured In Boston Marathon Bombings

This post was last updated at 7 PM ET on Monday, April 15.

Two bombs ripped though the crowds lining the route of the Boston Marathon this afternoon, killing two spectators, injuring more than 100 people, and leaving the nation on edge as word of the explosions –and confused reports of additional devices–flared across television and social media.

Most Read on Forbes

David M. Ewalt
Forbes Staff
+ Follow (493)

I've been writing about technology for over a decade, covering everything from video

(e) 15 April 2013

Boston Marathon Explosion Video Footage (GRAPHIC VIDEO, LIVE UPDATES)

Posted: 04/15/2013 3:31 pm EDT | Updated: 04/16/2013 12:11 am EDT

BOSTON MARATHON EXPLOSION

ALcomMobile Deceased Boston Marathon explosion suspect: 'I don't have a single American friend' | http://t.co/nerYDcdXfa
20 minutes ago from HootSuite

globaltvnews Police will execute 'controlled explosion' at secured house as part of #BostonMarathon bombing investigation http://t.co/0DDV07ANif
26 minutes ago from HootSuite

SkyNews Boston Marathon Explosion: Uncle Speaks Out http://t.co/LJNuS3rWJI
41 minutes ago from SkyNews Alerts – Latest

ALcomMobile Boston Marathon explosion: Uncle of suspects says they put shame on family and entire Chechnya ethnicity http://t.co/LPM29H2lVk
56 minutes ago from HootSuite

NBCLA MT @NBCNews "I feel blessed": Boston Marathon runner who escaped bombing unscathed witnessed Texas plant explosion http://t.co/liFw2Gxr4B
2 hours ago from SocialFlow

ALcomMobile Trying to catch up on the Boston Marathon explosion investigation... here's what we know. http://t.co/BgIpSvClEb
2 hours ago from HootSuite

(f) 15 April 2013

FIG. 20: Coverage of the Boston Marathon Bombing from different sites on specific point of time (April 15, 2013). Note that April 15, 2013 is the creation date of the Web pages that were used to create this story.

CHAPTER 2

BACKGROUND

There has been much interest in crafting digital narratives out of online resources and social media content to create stories using curation tools (e.g., Storify, Scoop.it) [250]. Despite the flexibility of these tools, they do not preserve the content of the resources (for example, Figure 4 in Chapter 1). While Web archives are solutions for preserving the Web, they lack tools that allow users to understand the archived collections. In our work, we will address integrating the storytelling techniques that users already are familiar with and Web archives which provide persistent data.

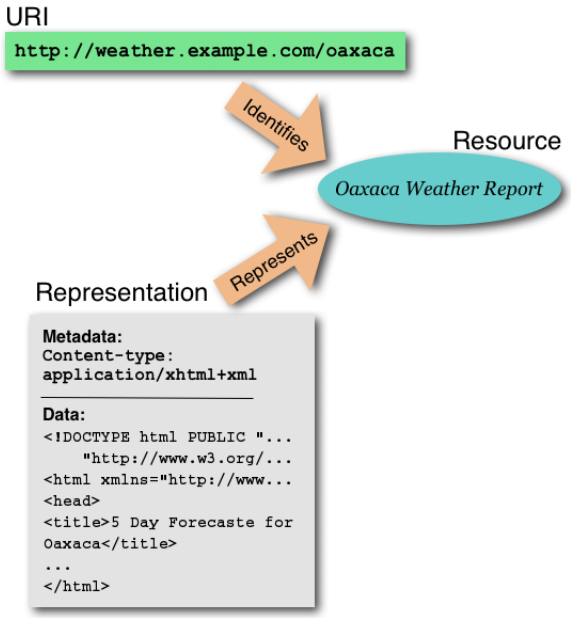
In this chapter, we briefly introduce the topics and concepts necessary to adequately understand the problem we are investigating in this dissertation. Moreover, the chapter includes the necessary terminology and definitions that will be discussed and utilized extensively throughout the next chapters. We introduce the anatomy of Web archives and content curation tools along with examples and illustrations on how they are used. The Memento Framework terminology adopted in the rest of the paper will be introduced as well.

2.1 THE WEB AND WEB ARCHIVES

The Web has become a major holder of our cultural record. Consequently, Web preservation is a fundamental precondition for research in history, sociology, political science, media, literature, and other related disciplines [240]. Before explaining the current trends in Web archives, we will briefly introduce the conventions of the Web as a primer to the background information needed to discuss the framework.

2.1.1 ARCHITECTURE OF THE WEB

The World Wide Web (WWW, or simply Web) was invented approximately 25 years ago by Tim Berners-Lee [44] as an information space for sharing documents and resources globally and providing distributed access for these resources, which are identified by Uniform Resource Identifiers (URI) [45]. The W3C's Architecture of the WWW [145] is illustrated in Figure 21. The figure demonstrates the relation between the URI, resource, and representation. As shown in the figure, resources are identified by URIs and when a URI is dereferenced, a representation (e.g., HTML, PDF) of the current state of the resource is returned to the user-agent (e.g., a browser). The common client-server relationships exist



```
Metadata:
Content-type:
application/xhtml+xml

Data:
<!DOCTYPE html PUBLIC "...
    "http://www.w3.org/...
<html xmlns="http://www...
<head>
<title>5 Day Forecaste for
Oaxaca</title>
...
</html>
```

FIG. 21: The relationship between identifier, resource, and representation [145].

in the context of the Web over the HyperText Transfer Protocol (HTTP), which determines the form of the representation through content-negotiation [103]. The time dimension has been absent from HTTP until the Memento protocol was introduced. We provide details about the Memento protocol in Section 2.1.3.

2.1.2 LONGEVITY OF URIS

Because the Web is a highly dynamic information space, the resources change, disappear, and move from one location to another frequently [177]. Many studies have shown that the expected lifetime of a Web page is short (between 44 to 100 days) [124, 182, 241, 201, 335, 34, 162] and Web resources disappear quickly [277, 302, 180]. In 1997, Brewster Kahle claimed that the average lifetime of a Web page was only 44 days [162]. In a subsequent study in 2001, Lawrence et al. [201] claimed that pages disappear after an average time of only 75 days. In a *Washington Post* article that was published in 2003, the expected lifetime of a Web page was estimated to be 100 days [335].

SalahEldeen et al. [276, 278] measured loss based on analyzing six different event-centeric datasets of resources shared in social media. They found that resources linked to in social media disappear (i.e., HTTP 404) at the rate of 11% per year for the first year, and 7% each year afterwards.

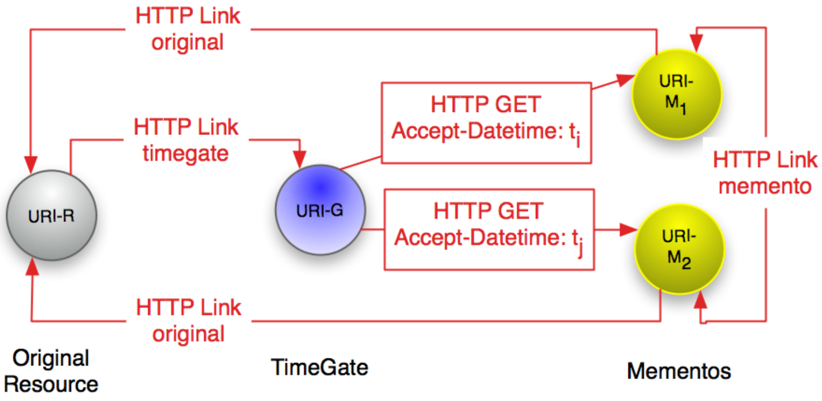


FIG. 22: Memento Framework. Source: `http://www.mementoweb.org/guide/quick-intro/`

Marshall et al. [222] showed that there are many reasons why URIs, or even entire websites, break such as: hacking, loss of account, owner deletion, server/service discontinued, etc. Lawrence et al. [201] found that many URI citations in computer science related papers have become invalid by a year or two after their publications. McCown et al. [227] conducted a study on articles published in *D-Lib Magazine* and found that the half-life of links in the articles is 10 years. Wallace Koehler [182] estimated the half-life of a random Web page is approximately two years.

The sites of Content Management Systems (CMS) such as MediaWiki, the platform used by Wikipedia[1] (the most popular information resource in the world with more than 500 million unique visitors monthly), typically links to external references in each article [10]. There are about 128,604 articles with dead links in Wikipedia references [2]. A recent Harvard study found that 49% of the URIs referenced in U.S. Supreme Court decisions are now dead [209].

The ephemeral nature of the Web highlights the importance of Web archives for historical purposes and records management compliance, capturing information. Providing tools that help normal users to understand the holdings of Web archives is important for raising awareness about Web archiving.

2.1.3 MEMENTO FRAMEWORK

As we have discussed in the previous section, the Web contains a huge amount of resources and these resources change over time. Web archives hold a substantial amount of

[1] `http://www.wikipedia.org/`

the Web. Integrating these archived resources into the live Web is important to give users access to archival content and allow people to browse the past [328].

The Memento protocol [328, 327, 329] was introduced in 2009 to allow temporal navigation of the Web. Memento is an HTTP protocol extension that enables time travel on the Web by linking current resources with their prior state. Memento introduces content negotiation in the datetime dimension using a special HTTP header, Accept-Datetime, that is sent by the user-agent to the TimeGate, a special resource that is aware of prior versions of Web resources, to indicate the preferred datetime (Figure 22). The Memento framework introduces new Relation Types for the HTTP "Link" header to convey typed links among Original Resources, TimeGates, Mementos, and TimeMaps.

Memento defines the following terms, which we will adopt in the rest of the dissertation:

- 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. URI-T(URI-R) = {URI-M$_1$, URI-M$_2$, ..., URI-M$_n$}.

Although Memento is supported with an effective set of client tools (e.g., MementoFox [282], Memento for Chrome [286], and for mobile iOS and Android devices [325], Mink [170], mCURL [23]), 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).

2.1.4 WEB ARCHIVING

Ben Saad and Gançarski [42] defined Web archiving as "the process of continuously collecting and preserving portions of the World Wide Web for future generations". An archived page is a snapshot of how this archived page looked at a particular point in time.

Ainsworth et al. estimated the coverage of Web resources in Web archives in "How Much of the Web Is Archived?" [13]. They sampled 4000 URIs from DMOZ, Delicious, Bitly, and search engines and measured their coverage in the public Web archives and the number and frequency of archived versions. They found that, according to the URI source, the archived percentage varies from 16% to 79%. These numbers increased in 2013 to be from 33% to 95% [24].

Helen Hockx-Yu provided a high-level overview of the Web archiving processes [133]. Based on her framework, the main processes of Web archiving are:

- *Selection* is determining the websites to be included in the Web archive collection.
- *Harvesting* (or crawling) automatically downloads copies of the specified websites. This starts with a list of seed URIs, then continues to crawl the hyperlinks.
- *Storage* is retaining the archived copies on a storage medium using archival formats, such as ARC [165] and WARC (Web ARChive) [314].
- *Access* refers to replaying and allowing the users to access the archived materials.
- *Digital Preservation* is the set of standards and technologies that are needed to ensure access to Web archives over time.

Selection is a key issue for Web archiving. Selection for Web archives can be manual (e.g., specifying seed URIs by people creating collections) or automatic (e.g., reading the URIs from a public directory). It is important to have a selection policy to ensure continuity and consistency in selection and revision [225]. The Archive-It service of the Internet Archive enables easy collection setup and management for institutions. As we detail in Section 2.1.5, the selection of the seed URIs that compose collections in Archive-It is personal, depending mainly on the domain knowledge of the curator, which suggests a need for tools that automatically discover new seed URIs.

Julien Masanès [225, 224] expressed a vision of the main issues involved by Web archiving such as the selection, storage, and preservation of Web content and the challenges that face them. Adrian Brown [53] also provided a practical guide for archiving the Web and the process of the archiving from selection, collection, storage, and delivery to the user.

It is a challenge for Web archiving institutes to balance between the completeness and quality of archived materials meanwhile avoid wasting time and space for storing and indexing [42]. The limitation of the resources of Web archives such as the storage, site politeness rules, etc., brought much attention from many researchers to optimize the processes of Web archiving lifecycle such as the selection, storage, and preservation [72, 42, 74, 248, 25, 26, 43].

AlSum et al. worked on enriching APIs for Web archives to support the access process [25, 26]. However, better APIs do not directly support increased archive exploration by humans. Rather than developing custom exploration interfaces for Web archives, we plan to utilize existing interface tools, such as Storify, which users are familiar to the general public.

Lin et al. [208] proposed Warcbase, which is a scalable Web archiving platform for storing, managing, and analyzing web archives to support the exploration and discovery in web archives. Warcbase applies modern "big data" infrastructure: HBase [69], an open-source platform for managing large semi-structured datasets, and Hadoop [68, 89], the open-source implementation of the MapReduce programming model. In a later work, Lin [207] describes how to scale down the infrastructure of Warcbase for providing new opportunities for personal web archiving.

Ben Saad et al. [42] proposed a framework that optimized page indexing and storage by discovering patterns from the temporal evolution of page changes using the data from the archive of French public TV channels. They claimed that these patterns can be a useful tool to predict changes and thus efficiently archive Web pages.

Focused crawling has become an active area of research to make such collection-building crawls more effective [67, 43, 100]. As Bergmark et al. mentioned [43], the goal of the focused crawl is to make a "best-first" crawl of the Web.

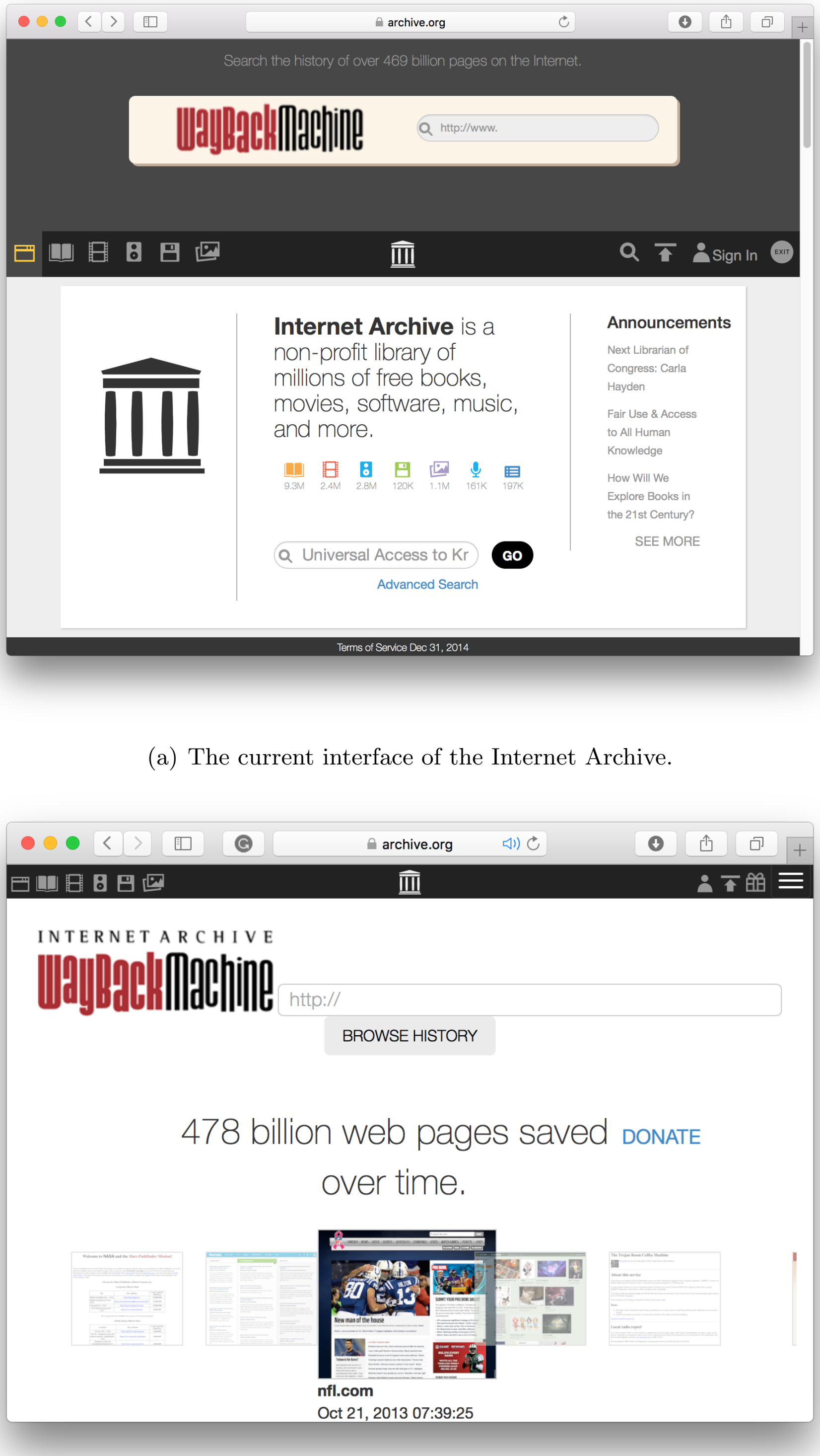
The previous techniques have focused on optimizing Web archives materials during the life cycle of Web archiving. Although these techniques are a good trend to avoid wasting time and space for storing and indexing Web pages, there is also a need to check the quality of archived materials that already exist in Web archives.

Excluding the off-topic pages from TimeMaps will significantly affect large-scale studies on archived materials. For example, the thumbnail summarization work [27] that was done by AlSum and Nelson would show off-topic pages in the generated summaries if these pages have not been detected before generating the summaries.

2.1.5 TYPES OF WEB ARCHIVES

With the significant growth in the amount of data, multiple Web archiving initiatives were started to archive different aspects of the Web [37, 113]. Web archives are those institutions that preserve much of the cultural discourse by archiving the Web [213]. We can categorize Web archiving initiatives into based on the scope and the purposes of their creation:

- Non-proprietary initiatives for archiving and preserving the entire Web, such as Internet Archive [3].
- Subscription services that allow institutions to create theme-based collections, such as Archive-It and the Web Archiving Service [7].
- On-demand free archiving services, such as Archive.is [1] and WebCite [99].



archive.org

Search the history of over 469 billion pages on the Internet.

Internet Archive is a non-profit library of millions of free books, movies, software, music, and more.

Announcements

Next Librarian of Congress: Carla Hayden

Fair Use & Access to All Human Knowledge

How Will We Explore Books in the 21st Century?

SEE MORE

Terms of Service Dec 31, 2014

(a) The current interface of the Internet Archive.

archive.org

478 billion web pages saved over time. DONATE

nfl.com

Oct 21, 2013 07:39:25

(b) The current interface of the Wayback Machine.

FIG. 23: The current interfaces of the IA and its Wayback Machine. Retrieved in March 2016.

- National libraries, such as the UK Web Archive [6], the Pandora project [65] at the National Library of Australia, the Greek Web project [198], etc.

In our work, we generate summary stories from the archived collections, such as the ones in Archive-It. Next, we will provide details on how these collections are created and how the people use Archive-It.

The Internet Archive

The Internet Archive is the largest and oldest Web archive [238], holding over 450 billion Web pages as far back as 1996 [164]. The Internet Archive was founded by Brewster Kahle to maintain an archive of the entire Web by taking periodic snapshots of pages then providing an access to these snapshots via the Wayback Machine [319].

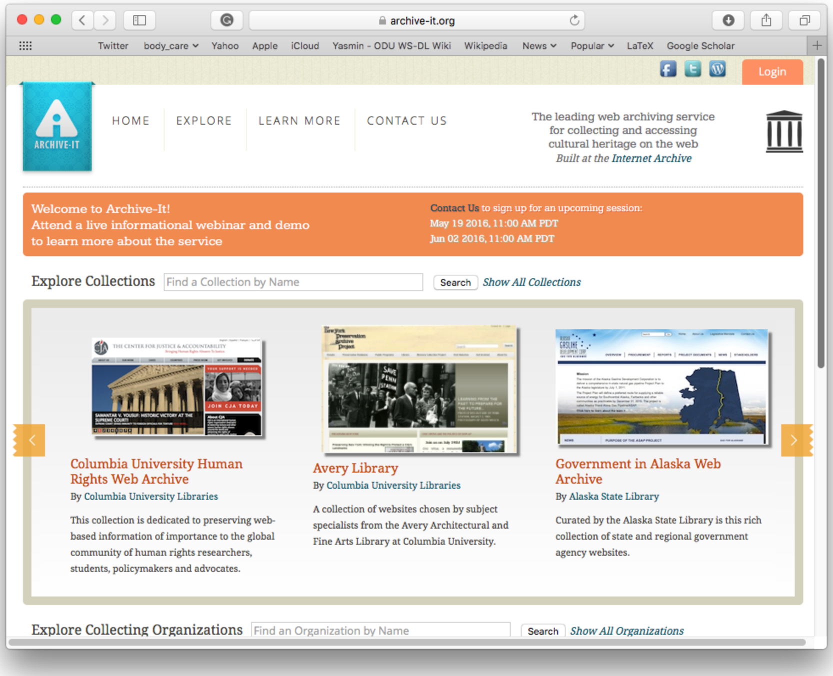
The Internet Archive enables users to see archived versions of Web pages across time, which the archive calls a “three dimensional index” [98]. In addition to Web pages, it also includes texts, audio, moving images, recordings, video games, TV broadcasts, and software in addition to a number of other projects such as the NASA Images Archive, the wiki-editable library catalog, and the Open Library. Moreover, the IA provides the Archive-It subscription service for institutions to build their own collections. Figure 23(a) shows the current homepage of the Internet Archive.

Access to the vast archive of Web pages is available through the Wayback Machine (Figure 23(b)), which allows archives of the World Wide Web to be accessed [319]. The Wayback Machine receives more than 82 million requests per day [18]. In Chapter 4, we study how humans and robots access and use Web archives using a dataset from the Wayback Machine’s access logs.

The Internet Archive announced an initiative to fix the broken links across the Internet to make URIs persistent on the live Web, such as using the archived pages for citing the references on Wikipedia [270, 8]. The content from Web archives can be used to fill the gaping holes left by dead pages on the live Web. This was a strong motive for us to create stories that are composed of persistent resources and integrate these stories with a storytelling service that people already know how to use.

Archive-It

Archive-It is a collection development service that has been operated by the Internet Archive since 2006. Archive-It provides Web archiving practices to a large number of organizations in the United States. As of May 2016, Archive-It was used by over 400 institutions in 48 states and featured over 9B archived Web pages in nearly 3,500 separate collections.



The leading web archiving service for collecting and accessing cultural heritage on the web
Built at the Internet Archive

Welcome to Archive-It!
Attend a live informational webinar and demo to learn more about the service

Contact Us to sign up for an upcoming session:
May 19 2016, 11:00 AM PDT
Jun 02 2016, 11:00 AM PDT

Explore Collections Find a Collection by Name *Show All Collections*

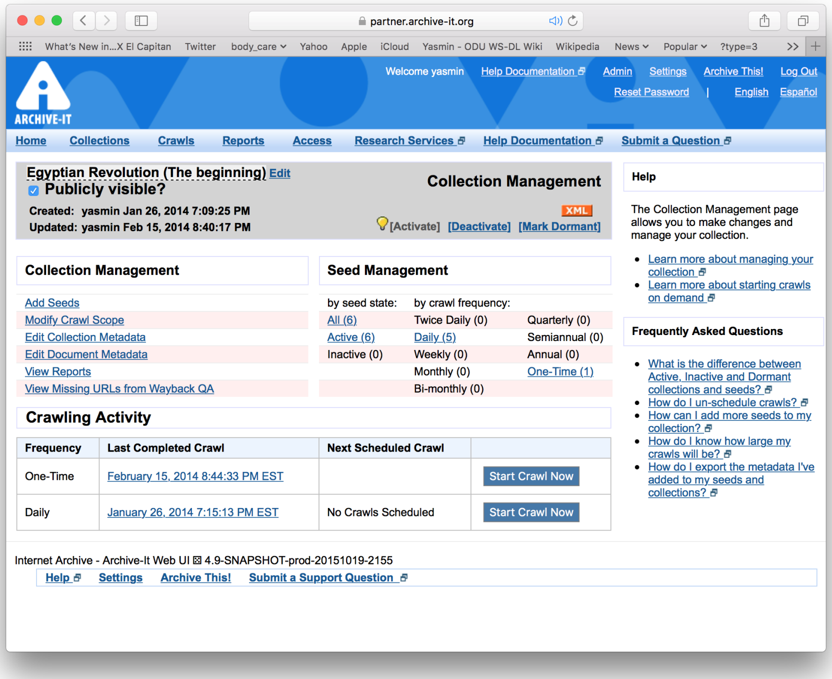
Columbia University Human Rights Web Archive
By Columbia University Libraries
This collection is dedicated to preserving web-based information of importance to the global community of human rights researchers, students, policymakers and advocates.

Avery Library
By Columbia University Libraries
A collection of websites chosen by subject specialists from the Avery Architectural and Fine Arts Library at Columbia University.

Government in Alaska Web Archive
By Alaska State Library
Curated by the Alaska State Library is this rich collection of state and regional government agency websites.

Explore Collecting Organizations Find an Organization by Name *Show All Organizations*

(a) The Archive-It interface from the user's view.



Egyptian Revolution (The beginning) Edit
☑ **Publicly visible?**
Created: yasmin Jan 26, 2014 7:09:25 PM
Updated: yasmin Feb 15, 2014 8:40:17 PM

Collection Management
[Activate] [Deactivate] [Mark Dormant]

Collection Management

- Add Seeds
- Modify Crawl Scope
- Edit Collection Metadata
- Edit Document Metadata
- View Reports
- View Missing URLs from Wayback QA

Seed Management

by seed state:	by crawl frequency:	
All (6)	Twice Daily (0)	Quarterly (0)
Active (6)	Daily (5)	Semiannual (0)
Inactive (0)	Weekly (0)	Annual (0)
	Monthly (0)	One-Time (1)
	Bi-monthly (0)	

Crawling Activity

Frequency	Last Completed Crawl	Next Scheduled Crawl	
One-Time	February 15, 2014 8:44:33 PM EST		Start Crawl Now
Daily	January 26, 2014 7:15:13 PM EST	No Crawls Scheduled	Start Crawl Now

Help

The Collection Management page allows you to make changes and manage your collection.

- Learn more about managing your collection
- Learn more about starting crawls on demand

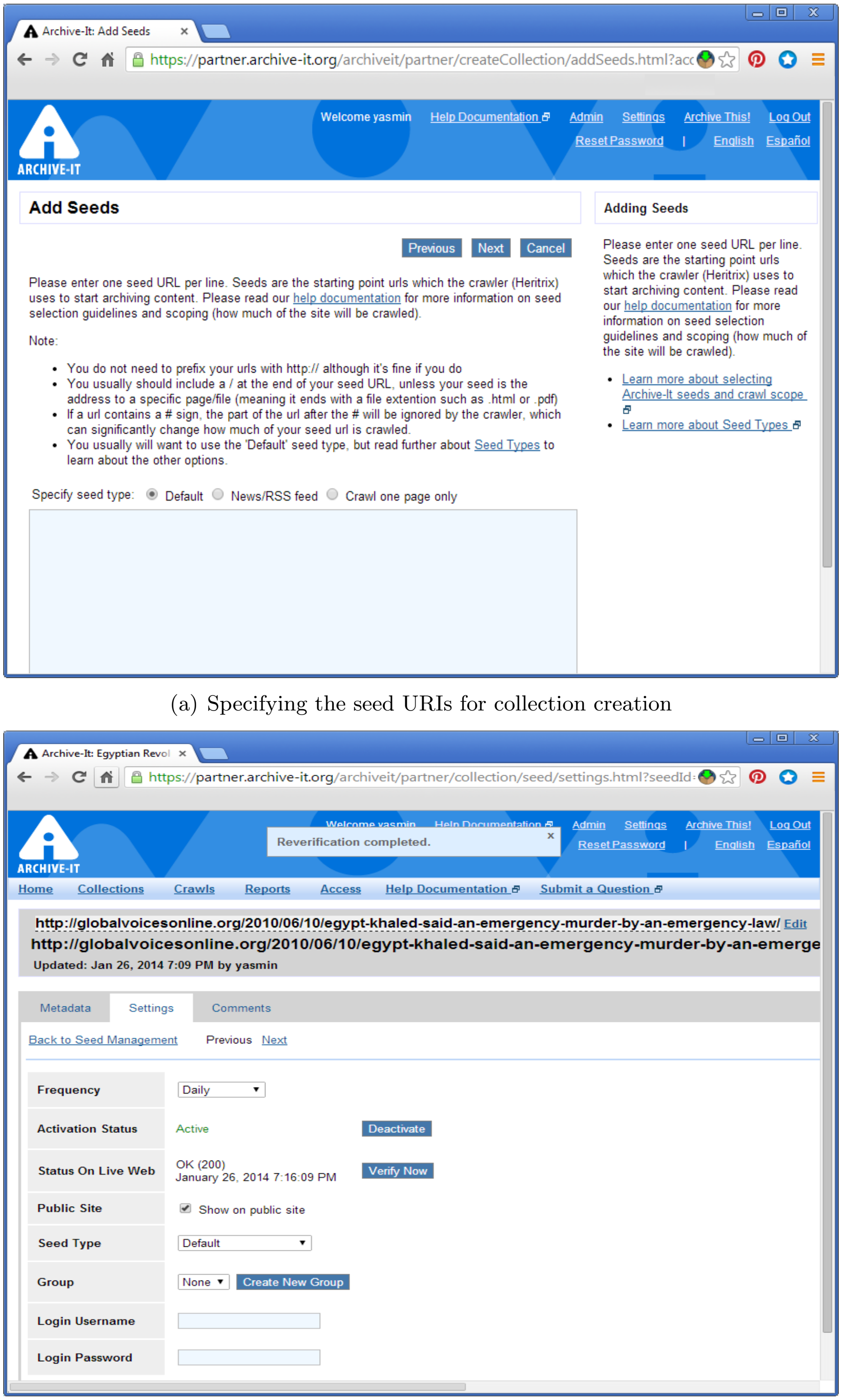
Frequently Asked Questions

- What is the difference between Active, Inactive and Dormant collections and seeds?
- How do I un-schedule crawls?
- How can I add more seeds to my collection?
- How do I know how large my crawls will be?
- How do I export the metadata I've added to my seeds and collections?

Internet Archive - Archive-It Web UI ⊠ 4.9-SNAPSHOT-prod-20151019-2155

(b) The Archive-It interface from the curator's view.

FIG. 24: Current Archive-It interfaces. Retrieved in January 2016.



Archive-It: Add Seeds

https://partner.archive-it.org/archiveit/partner/createCollection/addSeeds.html?acc

Add Seeds

Please enter one seed URL per line. Seeds are the starting point urls which the crawler (Heritrix) uses to start archiving content. Please read our help documentation for more information on seed selection guidelines and scoping (how much of the site will be crawled).

Note:

- You do not need to prefix your urls with http:// although it's fine if you do
- You usually should include a / at the end of your seed URL, unless your seed is the address to a specific page/file (meaning it ends with a file extention such as .html or .pdf)
- If a url contains a # sign, the part of the url after the # will be ignored by the crawler, which can significantly change how much of your seed url is crawled.
- You usually will want to use the 'Default' seed type, but read further about Seed Types to learn about the other options.

Specify seed type: ● Default ○ News/RSS feed ○ Crawl one page only

Adding Seeds

Please enter one seed URL per line. Seeds are the starting point urls which the crawler (Heritrix) uses to start archiving content. Please read our help documentation for more information on seed selection guidelines and scoping (how much of the site will be crawled).

- Learn more about selecting Archive-It seeds and crawl scope
- Learn more about Seed Types

(a) Specifying the seed URIs for collection creation

Archive-It: Egyptian Revol

https://partner.archive-it.org/archiveit/partner/collection/seed/settings.html?seedId:

Reverification completed.

http://globalvoicesonline.org/2010/06/10/egypt-khaled-said-an-emergency-murder-by-an-emergency-law/ Edit

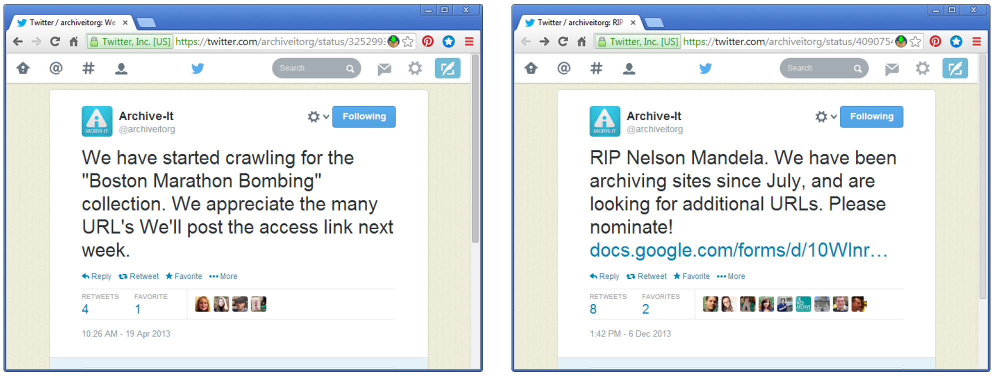
http://globalvoicesonline.org/2010/06/10/egypt-khaled-said-an-emergency-murder-by-an-emerge

Updated: Jan 26, 2014 7:09 PM by yasmin

Frequency	Daily ▾	
Activation Status	Active	Deactivate
Status On Live Web	OK (200) January 26, 2014 7:16:09 PM	Verify Now
Public Site	☑ Show on public site	
Seed Type	Default ▾	
Group	None ▾	Create New Group
Login Username		
Login Password		

(b) Specifying the parameters of crawl, such as the depth, the frequency, etc.

FIG. 25: For creating the collection, the curator specifies the seeds and the parameters of crawls. Retrieved in January 2016.


Twitter / archiveitorg: We
Twitter, Inc. [US] https://twitter.com/archiveitorg/status/325299
Search
Following
Reply
Retweet
Favorite
More
Twitter / archiveitorg: RIP
Twitter, Inc. [US] https://twitter.com/archiveitorg/status/409075
Search
Following
Reply
Retweet
Favorite
More

Archive-It
@archiveitorg

We have started crawling for the "Boston Marathon Bombing" collection. We appreciate the many URL's We'll post the access link next week.

RETWEETS 4 FAVORITE 1

10:26 AM - 19 Apr 2013

Archive-It
@archiveitorg

RIP Nelson Mandela. We have been archiving sites since July, and are looking for additional URLs. Please nominate! docs.google.com/forms/d/10WInr…

RETWEETS 8 FAVORITES 2

1:42 PM - 6 Dec 2013

(a) acquiring seed URIs about Boston Marathon Bombing

(b) acquiring seed URIs about Nelson Mandela

FIG. 26: Starting Archive-It collections that are built by the Archive-It team.

Archive-It provides their partners with tools that allow them to create themed collections of archived Web pages hosted on Archive-It machines (Figure 24). This is done by the curator specifying a set of seeds URIs that the curator believes best exemplifies the topic of the collection (Figure 25(a)). These URIs should be crawled periodically (the frequency is tunable by the curator), and to what depth (e.g., follow the pages linked to from the seeds two levels out), as shown in Figure 25(b). The Heritrix [233, 290] crawler, an open source Web crawler developed by Internet Archive specially designed for Web archiving, at Archive-It crawls/recrawls these seeds based on the predefined frequency and depth to build a collection of archived Web pages.

Archive-It provides tools that allow collection curators to perform collection management as well as quality control for the crawls. However, the tools are currently focused on crawl issues such as the mechanics of HTTP (e.g., how many HTML files and other file types, how many 404 missing URIs), as shown in Figure 8.

Archive-It collections are stored in the WARC file format [314], a revision of the Internet Archive's ARC file format that has traditionally been used to store Web crawls. The resources of the collection are combined and aggregated in a large WARC file. Each resource has a header containing metadata about how the resource was requested followed by the HTTP header and the response.

Starting Collections in Archive-It

Choosing seed URIs for a collection, especially collections centered around a specific event, is currently more art than science. Archive-It staff members create collections of global events (e.g., Arab Spring collection[2], SOPA Blackout collection[3]) under the name of Internet Archive Global Events. They collect the seed URIs by asking people to nominate URIs that are related to these events (Figure 26). Figure 26(a) shows an Archive-It request for URI nominations about the Boston Marathon Bombing, and Figure 26(b) contains a request to the community to nominate URIs about Nelson Mandela. The seed URIs are manually collected by people based on domain knowledge, which means there is no policy for automatically collecting the seed URIs. The result is a list of ad hoc URIs that are manually collected by users. Discovering seed URIs for building a collection is not easy.

Browsing Archive-It Collections

Archive-It provides a listing of all URIs in the collection along with the number of times and dates over which each site was archived, as well as a full-text search of archived sites, as we showed before in Figure 5 in Chapter 1. The main interface of the curated collection contains metadata about the collection that is added by the collection curator (Figure 24(b)). The Archive-It interface consists mainly of a list of seed URIs in alphabetical order in which the crawl information of each URI is available (Figure 6(a) and 6(b)). Clicking on any URI in the list presents a table listing dates when the mementos were captured (Figure 6(c)). Clicking on any date displays the archived version of the Web page at that date (Figure 6(d)). There is no tool to help users to understand the collection and gain insight about it other than the descriptive metadata on the collection page. To understand the collection, the user must go through all the URIs and browse all their relevant archived copies.

2.2 CONTENT CURATION PLATFORMS

Because of the sheer volume of information on the Web, there is a trend for creating tools that allow users to select and organize Web resources to create a narrative or story for a certain topic of interest [223]. These tools, which are called content curation tools, allow users to choose, collect, and manage their own narratives or stories (e.g., Storify, Scoop.it).

Ann Handley [122] defined content curation as:

> Content curation is the act of continually identifying, selecting and sharing the best and most relevant online content and other online resources (e.g., articles, blog posts, videos, photos, tools, tweets, etc.) on a specific subject to match the needs of a specific audience

[2] `https://archive-it.org/collections/2349/`

[3] `https://archive-it.org/collections/3010/`

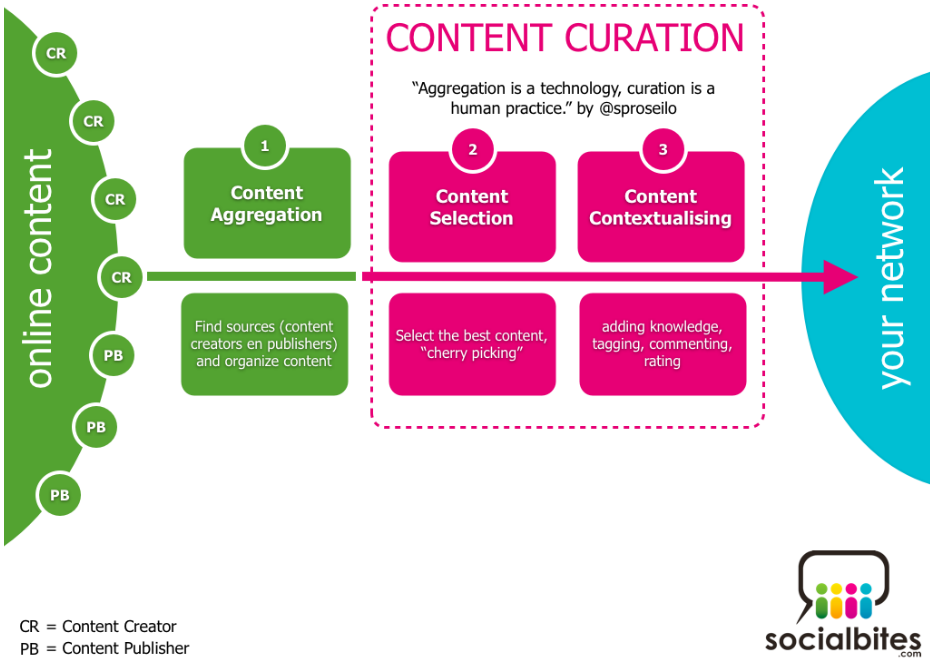

online content
CR
CR
CR
CR
PB
PB
PB
1
Content Aggregation
Find sources (content creators en publishers) and organize content
CONTENT CURATION
"Aggregation is a technology, curation is a human practice." by @sproseilo
2
Content Selection
Select the best content, "cherry picking"
3
Content Contextualising
adding knowledge, tagging, commenting, rating
your network
CR = Content Creator
PB = Content Publisher
socialbites.com

FIG. 27: The Process of Content Curation. Source: `http://socialmediatoday.com/pamdyer/1629516/60-content-curation-tools`

Content curation platforms allow users to create narratives or stories from the Web resources. The process of content curation (Figure 27) starts with selecting resources related to a topic of interest to the user (i.e., content aggregation) and then adding context to the collected content. The definition of "Digital Curation", as defined by the UK Digital Curation Centre, is "Digital curation, broadly interpreted, is about maintaining and adding value to, a trusted body of digital information for current and future use" [41, 111]. Holton et al. [134] described content curation tools as a filtering method for huge streams of social media.

Curation is important for people to handle information overload in digital resources [211]. Based on the analysis of 100 Web artifacts (e.g., blog posts, online news articles, videos including the comments in these posts), Liu [211] identifies seven curatorial activities that are interconnected: collecting, organizing, preserving, filtering, crafting a story, displaying, and facilitating discussions. Content curation existed before the Information Age, and for librarians it is much the same as what librarians used to do in the past [266, 41]; the curation of reference materials was being used a long time ago by patrons in the form of encyclopedias and specialized reference books.

Most of the content curation tools are general-purpose collection tools (i.e., they are not limited to news only, there are many forms of curation, such as video curation and

TABLE 1: Examples of different types of content curation tools.

Social bookmarking services	Diigo Reddit	diigo.com reddit.com
Visual bookmarking services	Pinterest Symbaloo	pinterest.com symbaloo.com
Hybrid curation tools	Storify Scoop.it	storify.com scoop.it

product curation) [114]. The art of finding, aggregating, filtering, selecting, curating, and republishing high-quality news stories on a specific topic or interest is important for librarians as well as journalists [250].

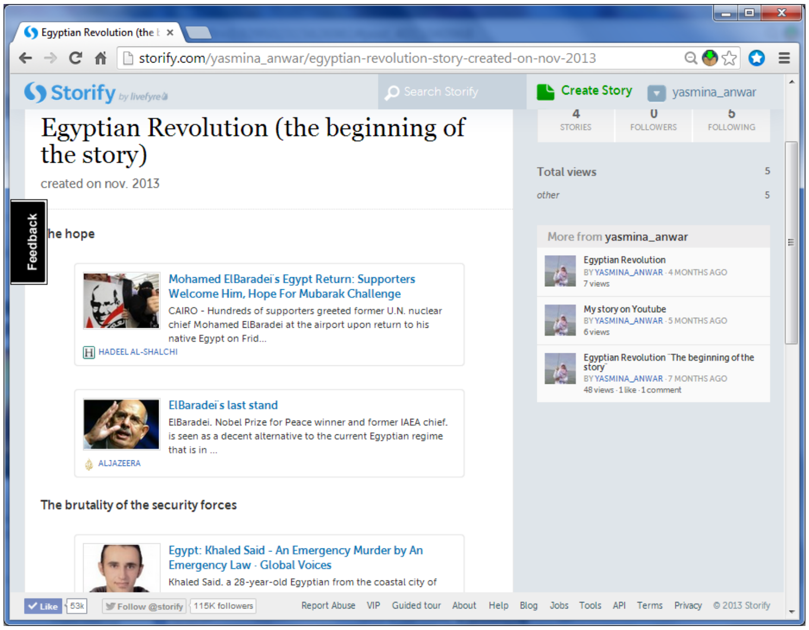
Recently, there have been many tools developed for digital curation. Table 1 shows examples of different kinds of curation tools [266] and examples for each kind: social bookmarking services, visual bookmarking services, and hybrid curation tools that are used for bookmarking and creating stories. Although recently there is increasing interest in content curation by users on social media platforms, relatively little attention has been given to the analysis of content curation platforms. In the next section, we will highlight some of the most popular curation tools and also the studies that have been done for understanding the nature of these tools.

2.2.1 STORIFY

Storify is a social networking curation service launched in 2010 that allows users to create a "story" of their own choosing, consisting of manually chosen Web resources, arranged with a visually attractive interface, clustered together with a single URI and suitable for sharing. Storify is one of the most prominent platforms for creating stories from many social media channels. Storify has a global rank of 5,410 as measured by Alexa[4] and has 850,000 users [261]. It provides a graphical interface for selecting URIs of Web resources and arranging the resulting snippets and previews, with a special emphasis on social media (e.g., Twitter, Facebook, YouTube, Instagram). The gathered resources of the story can be reordered and annotated. An example that shows a generated story about the Egyptian Revolution is illustrated in Figure 28.

The problem of Storify and the other curation tools is the persistency of resources that have been used to create a story or narrative. Like the problem of citation using non-persistent articles, if the URI that is chosen to be included in the story disappears, it will be difficult to know what it is about and the context will not be kept.

[4] http://www.alexa.com/siteinfo/storify.com/, accessed on May. 27, 2016


Egyptian Revolution (the b
storify.com/yasmina_anwar/egyptian-revolution-story-created-on-nov-2013
Storify by livefyre
Search Storify
Create Story
yasmina_anwar
4 STORIES
0 FOLLOWERS
5 FOLLOWING
Feedback
Like 53k
Follow @storify 115K followers
Report Abuse VIP Guided tour About Help Blog Jobs Tools API Terms Privacy © 2013 Storify

Egyptian Revolution (the beginning of the story)

created on nov. 2013

he hope

Mohamed ElBaradei's Egypt Return: Supporters Welcome Him, Hope For Mubarak Challenge

CAIRO - Hundreds of supporters greeted former U.N. nuclear chief Mohamed ElBaradei at the airport upon return to his native Egypt on Frid...

HADEEL AL-SHALCHI

ElBaradei's last stand

ElBaradei. Nobel Prize for Peace winner and former IAEA chief. is seen as a decent alternative to the current Egyptian regime that is in ...

ALJAZEERA

The brutality of the security forces

Egypt: Khaled Said - An Emergency Murder by An Emergency Law · Global Voices

Khaled Said. a 28-year-old Egyptian from the coastal city of

Total views 5

other 5

More from yasmina_anwar

Egyptian Revolution
BY YASMINA_ANWAR · 4 MONTHS AGO
7 views

My story on Youtube
BY YASMINA_ANWAR · 5 MONTHS AGO
6 views

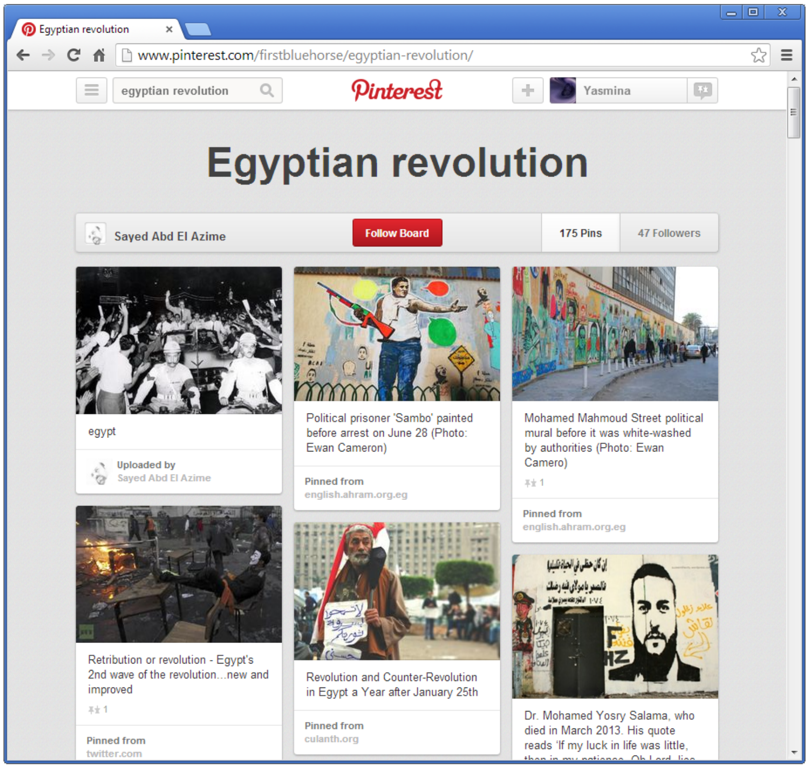
Egyptian Revolution "The beginning of the story"
BY YASMINA_ANWAR · 7 MONTHS AGO
48 views · 1 like · 1 comment

FIG. 28: A story about the Egyptian Revolution on Storify. Source: `https://storify.com/yasmina_anwar/egyptian-revolution-story-created-on-nov-2013`

Storify has been used in many studies by journalists [304] and also to explore how curation works in the classroom [231, 197]. Cohen et al. [78] believed that Storify can be used to encourage students to become empowered storytellers and researchers. Laire et al. [197] used Storify to study the effect of social media on teaching practices and writing activities.

Kieu et al. [174] proposed a method for predicting the popularity of social curation content based on a dataset from Storify. They specified the popularity of social curation using the number of views of the content. They used a machine learning approach based on curator and curation features (for example, the number of followers, the number of stories for the users, and the time that the user started using Storify) from stories. They found that combining the curator features (social features) and the curation features (content features) improves the performance of predicting the popularity.

We use Storify to present automatically created summaries of collections of archived Web pages in a social media interface that is more familiar to users (as opposed to custom interfaces for summaries, e.g. [252]). Since the stories in Storify are created by humans, we model the structural characteristics of these stories, with particular emphasis on "popular" stories (i.e., the top 25% of views, normalized by time available on the Web) (see Chapter 6).



Egyptian revolution

egypt

Uploaded by Sayed Abd El Azime

Political prisoner 'Sambo' painted before arrest on June 28 (Photo: Ewan Cameron)

Pinned from english.ahram.org.eg

Mohamed Mahmoud Street political mural before it was white-washed by authorities (Photo: Ewan Camero)

1

Pinned from english.ahram.org.eg

Retribution or revolution - Egypt's 2nd wave of the revolution...new and improved

1

Pinned from twitter.com

Revolution and Counter-Revolution in Egypt a Year after January 25th

Pinned from culanth.org

Dr. Mohamed Yosry Salama, who died in March 2013. His quote reads 'If my luck in life was little,

FIG. 29: A story about the Egyptian Revolution on Pinterest. Source: `https://www.pinterest.com/makarems/egyptian-revolution/`

2.2.2 PINTEREST

Pinterest is the most popular curation service with nearly 100 million users [142], as of September 2015. Pinterest has a global rank of 32 as measured by Alexa[5]. Pinterest is currently estimated as the third most popular social media website in the United States behind Facebook and Twitter [345, 118]. Pinterest's users pin images and videos onto boards to tell stories with pictures and videos found all over the Web, with the option of adding metadata to the resource [347]. Pinterest revolves around the metaphor of a pinboard, in which the user pins photos or videos of interest to create theme-based image/video collections such as hobbies, fashion, events, etc. In Pinterest, each pin includes the number of times it has been liked or re-pinned, along with a feed of any comments it has received. Users also can browse other pinboards for images. An example of a board about the Egyptian Revolution is shown in Figure 29.

Many studies have been conducted to study data curation using datasets from Pinterest [121, 283, 347, 112]. Zhong et al. [347] studied why and how people curate using datasets from Pinterest in January 2013 and Last.fm in December 2012. They found that curation tends to focus on items that may not be highly ranked in popularity and search rankings, which slightly contradicts our finding [19] based on a dataset from Storify.

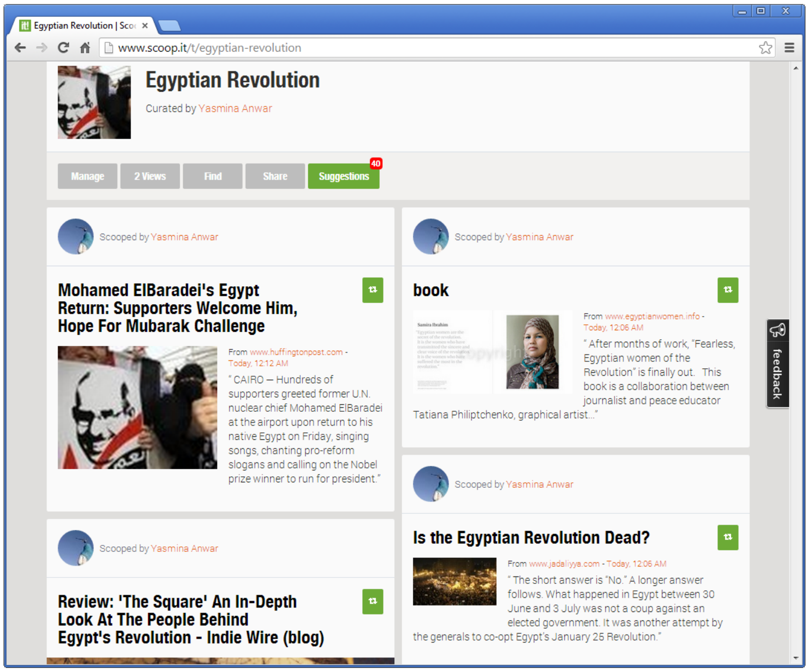
The most used subject areas by Pinterest users are food and drinks, décor and design, and apparel and accessories [118]. Most of the pins on Pinterest come from blogs, and a large number of pins were uploaded by the users from their own systems. Based on analyzing Pinterest data, Hall and Zarro [118] found that of the source type in their sample, there were 0.5% from archives, libraries, and museums.

2.2.3 SCOOP.IT

Scoop.it allows users to organize online content into magazine format by pulling information from various sources. The user specifies keywords to represent a specific topic so the content from multiple social media channels (e.g., Twitter, Facebook, Google, Scoop.it topics, and RSS Feeds) will be suggested. The user can edit the list of keywords at any time. There is also a bookmarklet to allow a user to add any page of interest. The primary feature that highlights Scoop.it is automatic suggestion, which allows users to get the latest resources that are related to a particular topic, then the user can publish the update and share it.

Antonio et al. [30] and Tuffley et al. [322] studied the potential of Scoop.it for facilitating learning and engaging the digital information literacy skills among high school students.

[5] `http://www.alexa.com/siteinfo/pinterest.com/`, accessed on May. 27, 2016


Egyptian Revolution | Scoo ×
www.scoop.it/t/egyptian-revolution
Manage
2 Views
Find
Share
Suggestions
40
feedback

Egyptian Revolution

Curated by Yasmina Anwar

Scooped by Yasmina Anwar

Mohamed ElBaradei's Egypt Return: Supporters Welcome Him, Hope For Mubarak Challenge

From www.huffingtonpost.com - Today, 12:12 AM

" CAIRO — Hundreds of supporters greeted former U.N. nuclear chief Mohamed ElBaradei at the airport upon return to his native Egypt on Friday, singing songs, chanting pro-reform slogans and calling on the Nobel prize winner to run for president."

Scooped by Yasmina Anwar

book

From www.egyptianwomen.info - Today, 12:06 AM

" After months of work, "Fearless, Egyptian women of the Revolution" is finally out. This book is a collaboration between journalist and peace educator Tatiana Philiptchenko, graphical artist..."

Scooped by Yasmina Anwar

Review: 'The Square' An In-Depth Look At The People Behind Egypt's Revolution - Indie Wire (blog)

Scooped by Yasmina Anwar

Is the Egyptian Revolution Dead?

From www.jadaliyya.com - Today, 12:06 AM

" The short answer is "No." A longer answer follows. What happened in Egypt between 30 June and 3 July was not a coup against an elected government. It was another attempt by the generals to co-opt Egypt's January 25 Revolution."

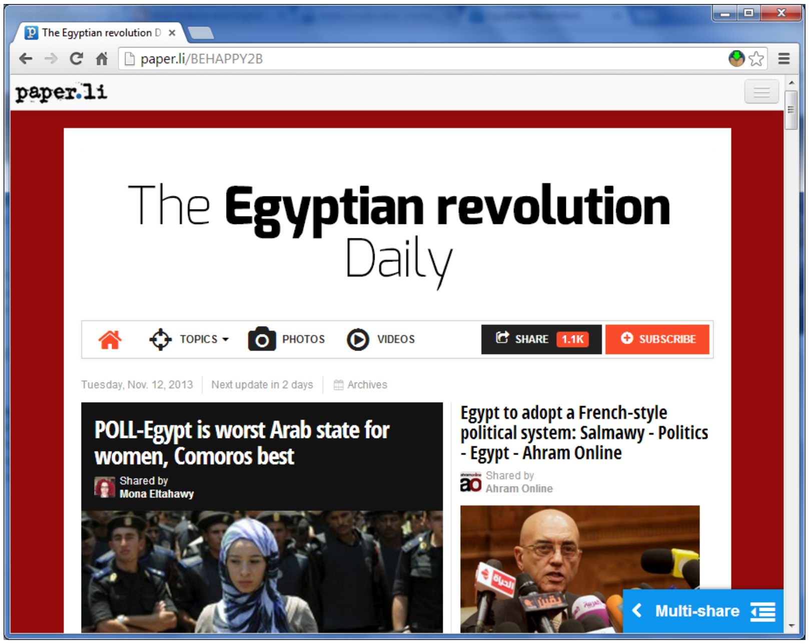
FIG. 30: A story about the Egyptian Revolution on Scoop.it. Source: `http://www.scoop.it/t/egyptian-revolution-the-beginning-of-the-story`

They found that it is important for the students to know how to prioritize the selected Web pages they collect to create stories. They also found that Scoop.it facilitates engagement, but it has less effect on improving the digital information literacy skills of the students.

Saaya et al. [272, 273] introduced a content-based recommendation framework for automatically assigning new resources of a collection to users based on the content of the collection. Their method depends on capturing the essence of the collection using features extracted from the pages, such as titles and descriptions, then classifying a given URI as belonging to a particular collection. They used three information retrieval approaches (TF-IDF, which is calculated using Lucene [126]) and two other classification approaches: Naive Bayes Multinomial (NBM) [173] and Support Vector Machines (SVM) [80] in Weka [119].

2.2.4 PAPER.LI

Paper.li enables users to create and publish their own topic based newspapers, called a *paper*. It allows users to choose the sources, such as Facebook, Google+, Twitter, RSS feeds, and other sources based on keywords. After this, it creates a paper automatically that contains the most recent related materials, such as text (e.g., blogs, news articles),



The Egyptian revolution D

paper.li/BEHAPPY2B

The Egyptian revolution Daily

Tuesday, Nov. 12, 2013 | Next update in 2 days | Archives

POLL-Egypt is worst Arab state for women, Comoros best

Shared by Mona Eltahawy

Egypt to adopt a French-style political system: Salmawy - Politics - Egypt - Ahram Online

Shared by Ahram Online

FIG. 31: A story about the Egyptian Revolution on Paper.li. Source: `http://paper.li/BEHAPPY2B`

photos (e.g., Flickr, Twitpic), and videos (e.g., YouTube, Vimeo). Paper.li automatically detects the relevant content daily, and then updates the paper [141].

Figure 31 contains an example of a paper about the Egyptian Revolution on Paper.li.

2.2.5 OTHER TOOLS

There are many other content curation tools that allow users to bookmark, collect, and organize their favorite resources manually, for example, Facebook[6] timeline, Twitter[7] timeline, Roojoom[8] (presents the collected resources in a timeline), Pearltrees[9] (visually organizes the resources and place them in a tree), Bundlr[10] (presents the gathered resources in list view or grid view), Togetter[11] (a popular curation service in Japan, was being used for the social curation of microblogs such as tweets [96]) and TweetDeck[12] (a social media dashboard for organization tweets in a column-based interface).

[6] `https://www.facebook.com/`
[7] `https://www.twitter.com/`
[8] `https://www.roojoom.com/`
[9] `http://www.pearltrees.com/`
[10] `http://bundlr.com/`
[11] `http://togetter.com/`
[12] `https://tweetdeck.twitter.com/`

2.3 SUMMARY

In this chapter, we presented the definitions and terminology that are adopted in the rest of chapters. We presented a high level overview of the Web, the Web architecture, longevity of URIs, and the anatomy of Web archives and content curation tools. We explained processes of Web archiving life cycle and how its optimization processes has been handled in research. We also presented the types of Web archives, focusing on the Internet Archive and Archive-It. We illustrated how an archived collection is curated and browsed in Archive-It.

In the next chapter, we discuss the prior research that we leverage into our research. We also present and compare the related work of research that we conducted in the DSA framework.

CHAPTER 3

RELATED WORK

In this chapter, we provide an overview of the research that has been established in several fields related to the problems we investigate in this dissertation. We provide an overview of different methodology and techniques for handling collection understanding (Section 3.1). We present how the previous research handled telling stories with data through summarizing the work of narrative visualizations and time series visualizations (Section 3.2). The related techniques of text analysis and usage mining that we adopted in the DSA framework are presented in Section 3.3. Section 3.4 contains the related research of Web archives usage and mining the past web. At the end, we present the different notions of time (Section 3.5).

3.1 COLLECTION UNDERSTANDING

Collection understanding is the focus of gaining a comprehensive view of a collection [70]. Collection understanding is different from the Information Retrieval (IR) focus, which is about locating specific resources in the collection using a keyword or phrase [70]. In the following sections, we overview the previously suggested solutions for collection understanding. We overview document collections visualization, image collections summarization, and video abstraction. We also contrast these solutions against the solution we introduce in the DSA framework.

3.1.1 ENCODED ARCHIVAL DESCRIPTION

In order to preserve the evidentiary value of the collections and summarize their scope, archivists typically create a document containing detailed information about a specific collection of papers or records within an archive called a *finding aid* [106, 215]. A finding aid provides a comprehensive overview of a collection. it also provides a description of a collection's components parts in details.

An Encoded Archival Description (EAD)[1], a Document Type Definition (DTD) defined in the Extensible Mark-up Language (XML), has been developed as a machine readable encoding standard of finding aids created by archives, libraries, museums, and manuscript repositories to support the use of their holdings. EAD is the de facto standard for describing

[1] `http://www.loc.gov/ead/`

collections [258]. It has been developed mainly for supporting "collection understanding" [70], facilitating the goals of a standardized description [188], and allowing for the emergence of new archival descriptive practices [287]. The focus of EAD is on the structural content of archival description, not on its presentation.

Francisco-Revilla et al. [106] investigated the quality of finding aids and their impact on information visualization techniques by analyzing a set of 8729 finding aids aggregated by the Texas Archival Repository Online[2] (TARO) using VADA, a visual analytic tool for finding aids. They also discussed the aggregations of finding aids to specific aspects of EAD, EADs design and the actual encoding practices of EAD, and the problems associated with the EAD standard. They provided recommendations for improving the quality of finding aid data. They concluded that EAD allows great flexibility in the encoding of finding aids and this is a positive factor for encoding legacy data and accommodating the practices of multiple different archival repositories.

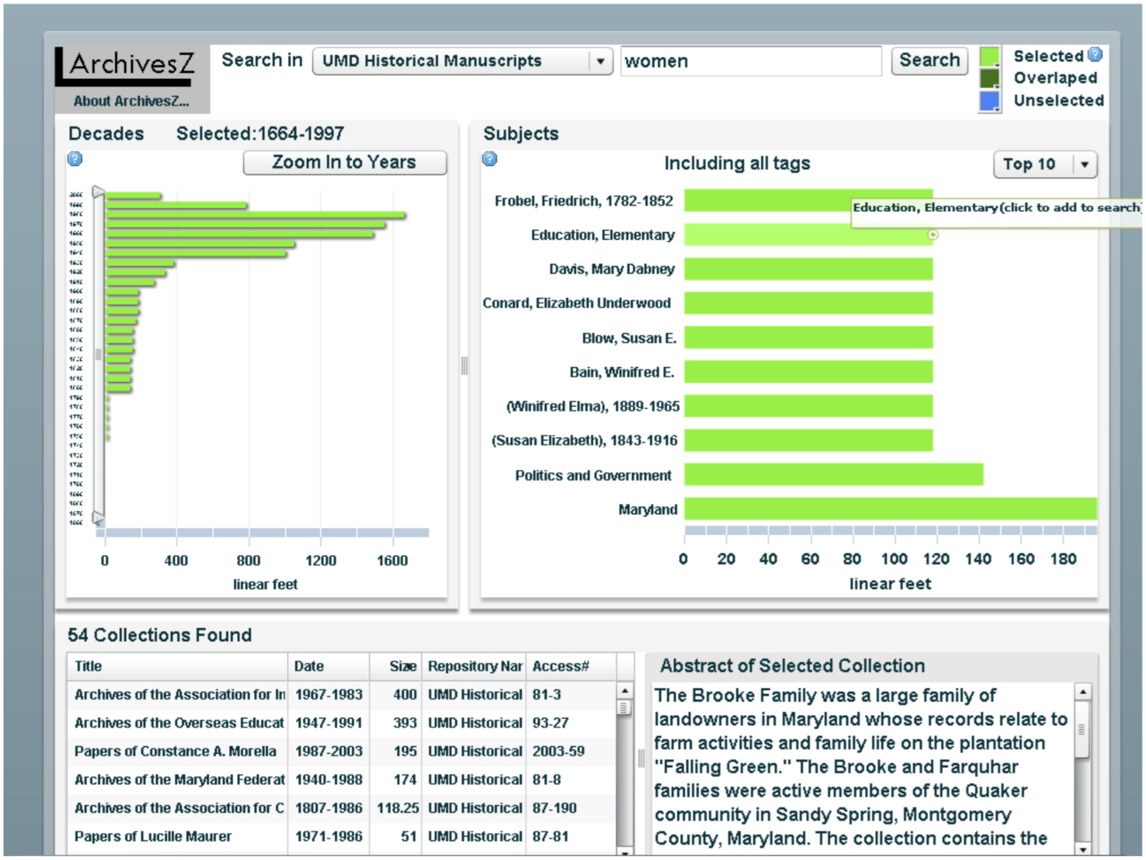
The potential of EAD is to enable finding aids to be encoded, searched, and displayed online. We believe that increased textual metadata (e.g., EAD) added to the interface as shown in Figure 1 is not a solution for getting the essence of the collection. Instead, we are informed by emerging trends in social media storytelling, which focus on a small number of exemplary pages (i.e., "high precision" in information retrieval terms) as chosen by a human.

3.1.2 VISUALIZING DOCUMENT COLLECTIONS

Since the digitization process has started, most institutions, e.g., libraries and archives, have focused on storing digital collections and making them accessible online [88]. Most of the current digital collection interfaces are text-based search with very limited browsing features. Much research has been dedicated to developing visualizations for viewing and querying documents, and towards graphical querying and browsing of the results [130, 12, 336, 138, 137].

On of the earliest efforts in the description of visualizing data via a "starfield display" was by Ahlberg and Shneiderman [12]. They presented multiple visualization techniques for presenting a large volume of data. They introduced the key visual information seeking concepts, which have been used in visualizing document collections: rapid filtering to reduce result sets, progressive refinement of search parameters, continuous reformulation of goals, and visual scanning to identify results. They added a number of new principles for supporting visual search such as dynamic query filters, use of a starfield display, and

[2] `https://www.lib.utexas.edu/taro/`



Search in UMD Historical Manuscripts women

54 Collections Found

Title	Date	Size	Repository Nar	Access#
Archives of the Association for In	1967-1983	400	UMD Historical	81-3
Archives of the Overseas Educat	1947-1991	393	UMD Historical	93-27
Papers of Constance A. Morella	1987-2003	195	UMD Historical	2003-59
Archives of the Maryland Federat	1940-1988	174	UMD Historical	81-8
Archives of the Association for C	1807-1986	118.25	UMD Historical	87-190
Papers of Lucille Maurer	1971-1986	51	UMD Historical	87-81

Abstract of Selected Collection

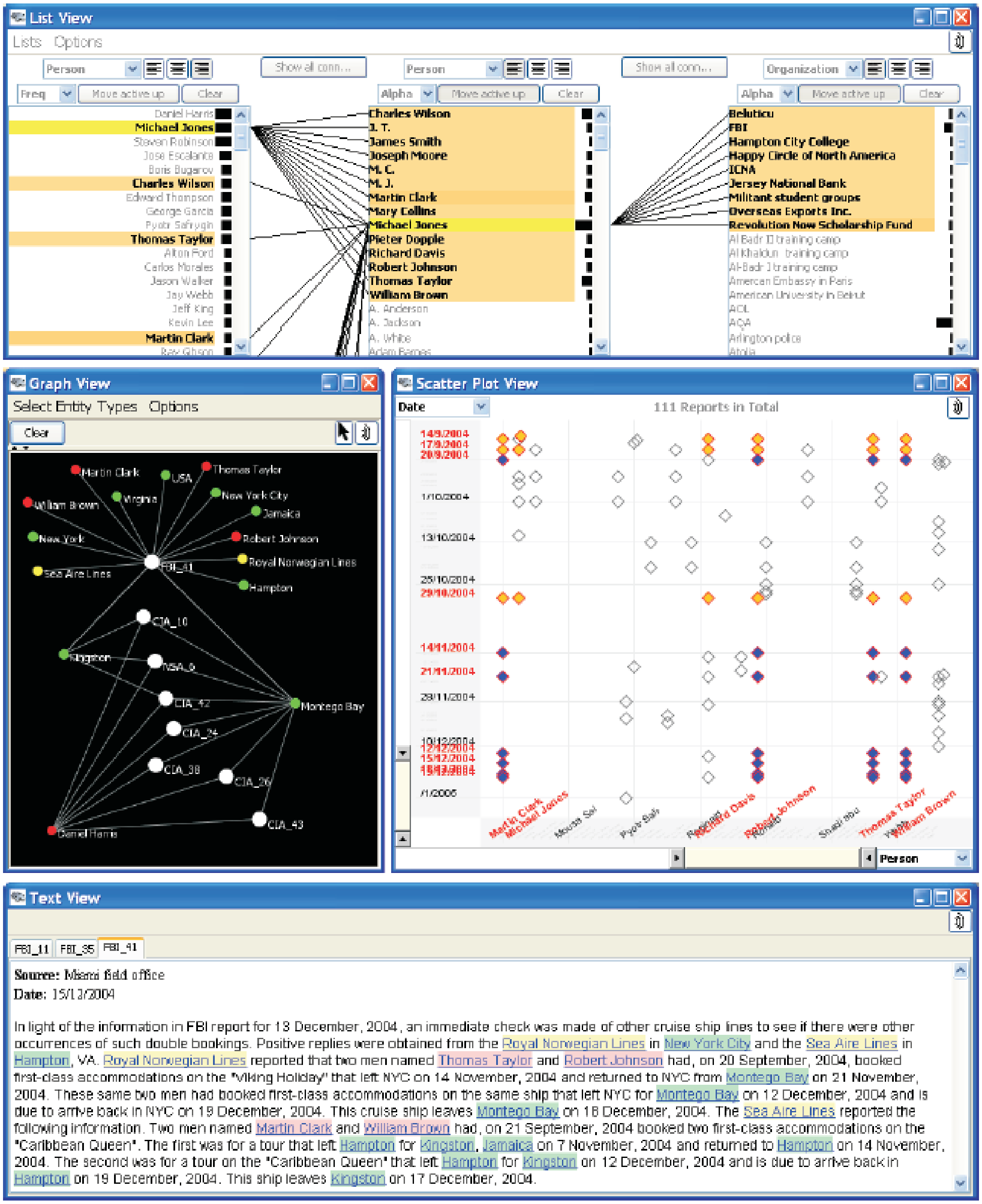
The Brooke Family was a large family of landowners in Maryland whose records relate to farm activities and family life on the plantation "Falling Green." The Brooke and Farquhar families were active members of the Quaker community in Sandy Spring, Montgomery County, Maryland. The collection contains the

FIG. 32: ArchivesZ: First Level Search Results. Gives a full overview of years and top 10 subjects by total linear feet [188].

tight coupling. Many of the principles Ahlberg and Shneiderman introduced are the basis for many of today's visualizations [88].

Karmer-Smyth [188] developed ArchivesZ, an information visualization for archived collections inspired by the availability of structured data in the EAD standard for encoding finding aids. The ArchivesZ prototype interface help users explore the metadata that describes archival collections through searching for content by year and subject in a tightly coupled dual histogram interface. ArchivesZ uses linear feet as a unit of measurement rather than the number of separate collections. A linear foot[3] is a measure of shelf space necessary to store documents. Therefore, ArchivesZ gives users a visual representation of the total amount of content available in an archive on a given topic. It also visualizes the overlapping assignment of subjects terms to archival collections. Therefore, all collections about the same subject can be grouped together, rather than a single tagged collection through leveraging the combination of key structured data elements of metadata about

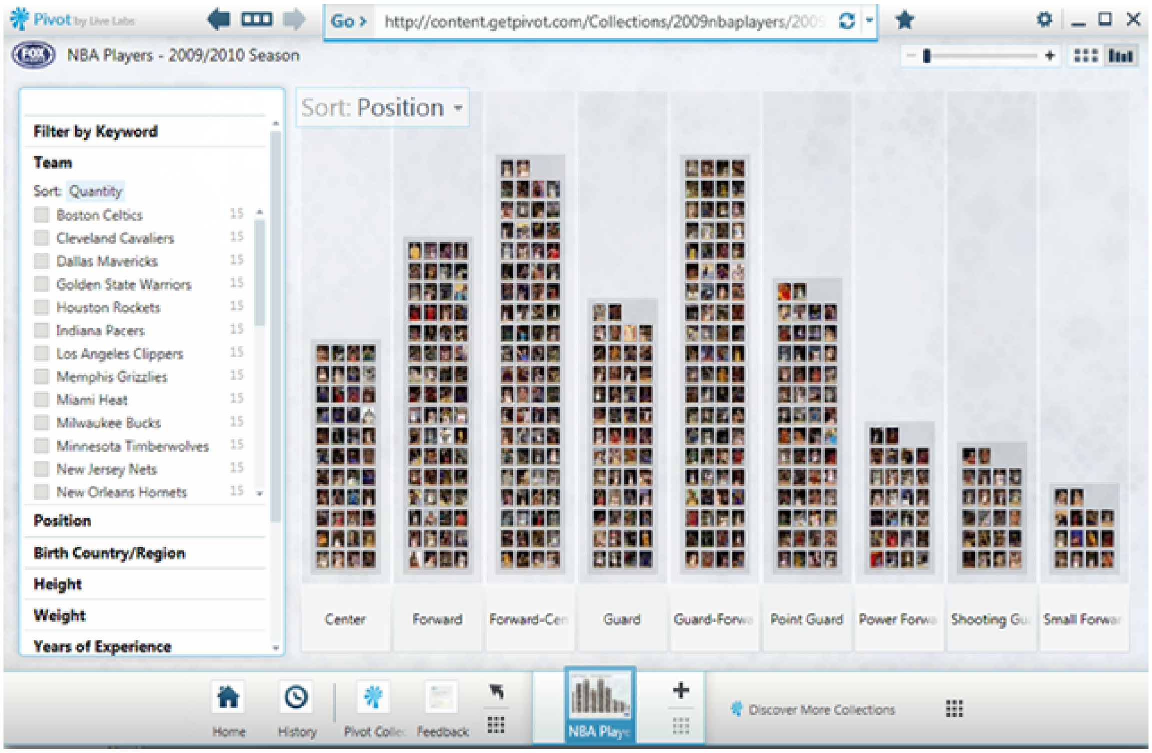
[3] http://www2.archivists.org/glossary/terms/l/linear-foot



Source: Miami field office
Date: 15/12/2004

In light of the information in FBI report for 13 December, 2004 , an immediate check was made of other cruise ship lines to see if there were other occurrences of such double bookings. Positive replies were obtained from the Royal Norwegian Lines in New York City and the Sea Aire Lines in Hampton, VA. Royal Norwegian Lines reported that two men named Thomas Taylor and Robert Johnson had, on 20 September, 2004, booked first-class accommodations on the "Viking Holiday" that left NYC on 14 November, 2004 and returned to NYC from Montego Bay on 21 November, 2004. These same two men had booked first-class accommodations on the same ship that left NYC for Montego Bay on 12 December, 2004 and is due to arrive back in NYC on 19 December, 2004. This cruise ship leaves Montego Bay on 18 December, 2004. The Sea Aire Lines reported the following information. Two men named Martin Clark and William Brown had, on 21 September, 2004 booked two first-class accommodations on the "Caribbean Queen". The first was for a tour that left Hampton for Kingston, Jamaica on 7 November, 2004 and returned to Hampton on 14 November, 2004. The second was for a tour on the "Caribbean Queen" that left Hampton for Kingston on 12 December, 2004 and is due to arrive back in Hampton on 19 December, 2004. This ship leaves Kingston on 17 December, 2004.

FIG. 33: Various views (List View, Graph View, Scatter Plot View, and Text View) for the visual analytic system, Jigsaw [305].



Sort: Position

Sort: Quantity

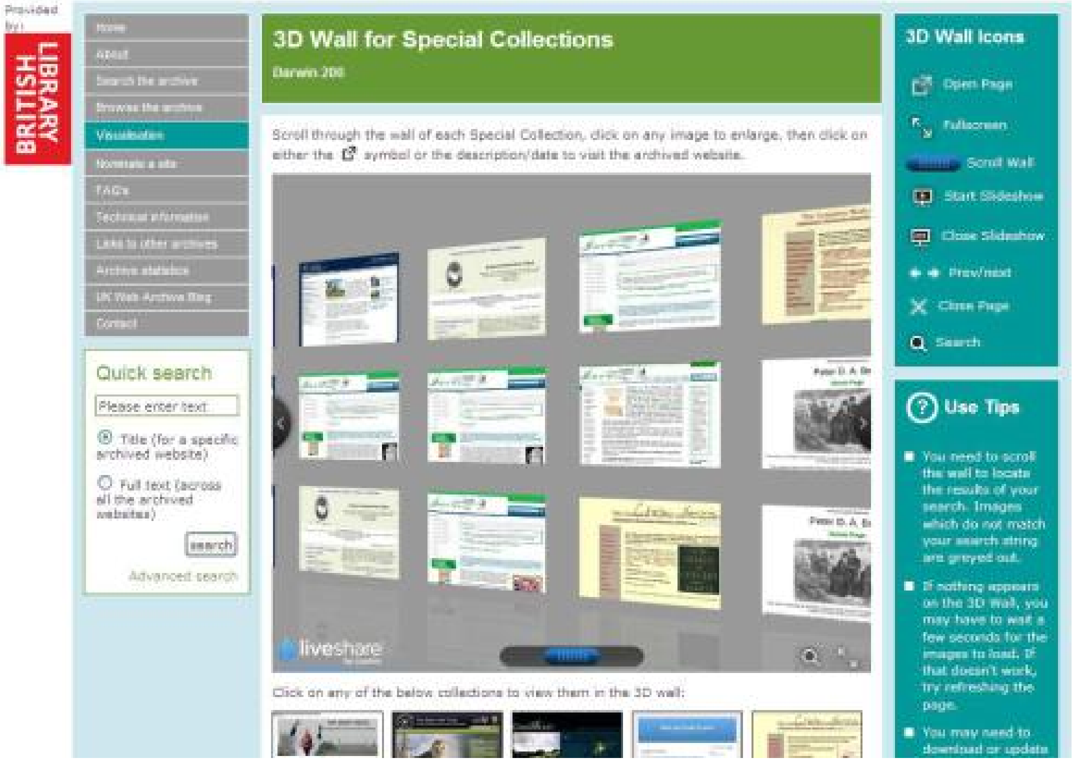
□ Boston Celtics	15
□ Cleveland Cavaliers	15
□ Dallas Mavericks	15
□ Golden State Warriors	15
□ Houston Rockets	15
□ Indiana Pacers	15
□ Los Angeles Clippers	15
□ Memphis Grizzlies	15
□ Miami Heat	15
□ Milwaukee Bucks	15
□ Minnesota Timberwolves	15
□ New Jersey Nets	15
□ New Orleans Hornets	15

FIG. 34: An example of Microsoft PivotViewer showing positions of NBA players from the 2009/2010 season. Source: `http://www.michaelmcclary.net/image.axd?picture=image_12.png`

archival collections. An example for a visualization of aggregate information about groups of archival collections is shown in Figure 32. The figure shows the range of decades covered by all collections, the top five subject terms based on the total linear feet worth of collections associated with that subject term, and a list of collections returned by the search.

Hearst et al. [130] experimented with the use of hierarchical metadata and hyperlinked images as results for the purpose of browsing and searching through information on the Web. A usability study conducted by the authors suggested that about 50% of users used images as a primary means of browsing and searching for information all the time. Their finding indicates that users are more inclined towards visual methods of querying and browsing rather than textual methods.

Many visual analytics tool were developed to visualize text documents [115, 305, 129]. Jigsaw [115] is a visual analytics system which provides a series of visual interfaces for investigative analysis across text documents' collections. Jigsaw is an important tool for analysts, especially when it comes to large text corpora, by highlighting inter-connections between entities across documents [305]. It provides multiple views (Figure 33):



3D Wall for Special Collections

Darwin 200

Scroll through the wall of each Special Collection, click on any image to enlarge, then click on either the symbol or the description/date to visit the archived website.

◉ Title (for a specific archived website)

○ Full text (across all the archived websites)

Click on any of the below collections to view them in the 3D wall:

Use Tips

- You need to scroll the wall to locate the results of your search. Images which do not match your search string are greyed out.
- If nothing appears on the 3D Wall, you may have to wait a few seconds for the images to load. If that doesn't work, try refreshing the page.
- You may need to download or update

FIG. 35: A 3D wall visualization for collections in U.K. Web Archive. Source: `http://takingaccountproject.wordpress.com/2012/03/14/uk-web-archive/`

1. The List View, which contains multiple reorderable lists of entities, uses colors to show connections between entities.

2. The Calendar View adds temporal context to the documents.

3. The Time Line View shows connections between entities and dates.

4. The Text View presents the actual reports with highlighting the entities.

5. The Scatter Plot View highlights pairwise relationships between any two entity types.

6. The Graph View shows the connection between entities and reports in a node-link diagram.

Although Jigsaw provides multiple views for large text corpora, it does not preserve hierarchies in a document collection. Furthermore, Jigsaw supports only text documents and cannot be used to visualize multimedia documents such as Web pages containing images and videos.

For temporal visualization of large document collections, ThemeRiver [129] provides contextual information through thematic changes within the documents over time. TIARA (Text Insight via Automated Responsive Analytics) [334] applies the ThemeRiver metaphor to visually summarize a text collection based on the topic content. It combines text analytics and interactive visualization to help users explore and analyze large collections of text.

PivotViewer [5, 86] is a Silverlight [4] application for exploring large datasets with a flexible visual interactive manner. It was released by Microsoft Live Labs in 2009. PivotViewer allows users to interact with massive amounts of data dynamically, uncovering trends and patterns in a visual format. Pivot can load any form of data and represent it as a deck of cards, with similar cards stacked together. By visualizing thousands of related items at once, users can see trends and patterns that would be hidden when looking at one item at a time [338]. Figure 34 has an example of Microsoft PivotViewer showing positions for NBA players the in 2009/2010 season.

The UK Web Archive provides a visualization for the collections through a 3D wall of sites allowing interaction through zooming (Figure 35).

Our initial attempt to browse Archive-It collections and highlight the collections' underlying characteristics was applying four alternate visualizations (Figure 36 and 37) for the Archive-It interface [251, 252]:

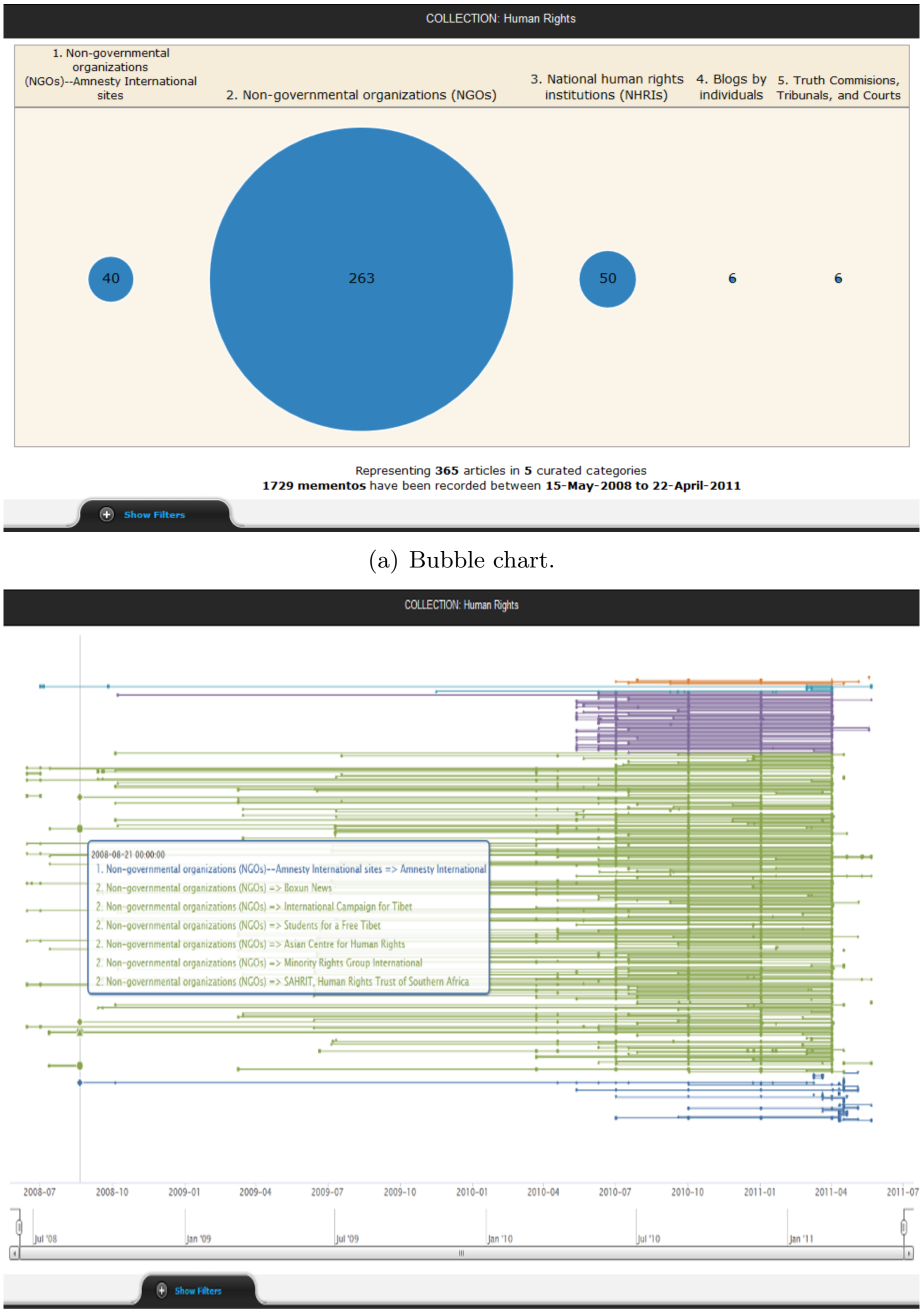
- *Bubble chart* provides a quick summary of the collection by displaying each group in the collection as a bubble, where the size of the bubble represents the number of sites in each group.

- *Image plot with histogram* allows the user to explore the collection by representing all sites in a collection in a graphical manner. Each screenshot is linked to a list of archived versions in Archive-It.

- *Wordle* [330] appears when hovering over any image in the image plot.

- *Timeline* provides an insight about the development of the collection over time. In this visualization, each site is represented as a single horizontal line, the length of which denotes the duration over which its archived copies have been captured. Each point on the line represents an archived copy of the site. Hovering over a point displays a list of archives of other sites captured on that same day.

For those collections that lack a curator-defined grouping, we also provided a heuristic-based categorization to make the new visualizations more meaningful. Figure 38(b) shows


COLLECTION: Human Rights
1. Non-governmental organizations (NGOs)--Amnesty International sites
2. Non-governmental organizations (NGOs)
3. National human rights institutions (NHRIs)
4. Blogs by individuals
5. Truth Commisions, Tribunals, and Courts
40
263
50
6
6
Show Filters

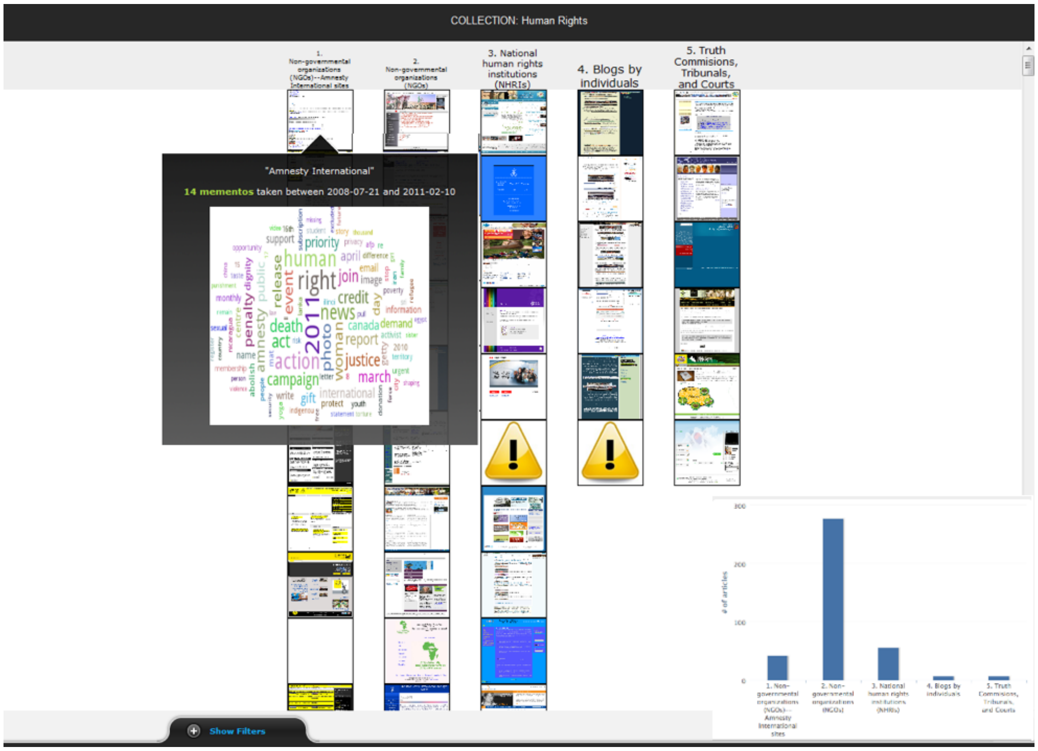
Representing **365** articles in **5** curated categories
1729 mementos have been recorded between **15-May-2008 to 22-April-2011**

(a) Bubble chart.

2008-08-21 00:00:00
1. Non-governmental organizations (NGOs)--Amnesty International sites => Amnesty International
2. Non-governmental organizations (NGOs) => Boxun News
2. Non-governmental organizations (NGOs) => International Campaign for Tibet
2. Non-governmental organizations (NGOs) => Students for a Free Tibet
2. Non-governmental organizations (NGOs) => Asian Centre for Human Rights
2. Non-governmental organizations (NGOs) => Minority Rights Group International
2. Non-governmental organizations (NGOs) => SAHRIT, Human Rights Trust of Southern Africa

(b) Timeline

FIG. 36: Different visualizations for exploring Human Rights collection at Archive-It.



COLLECTION: Human Rights

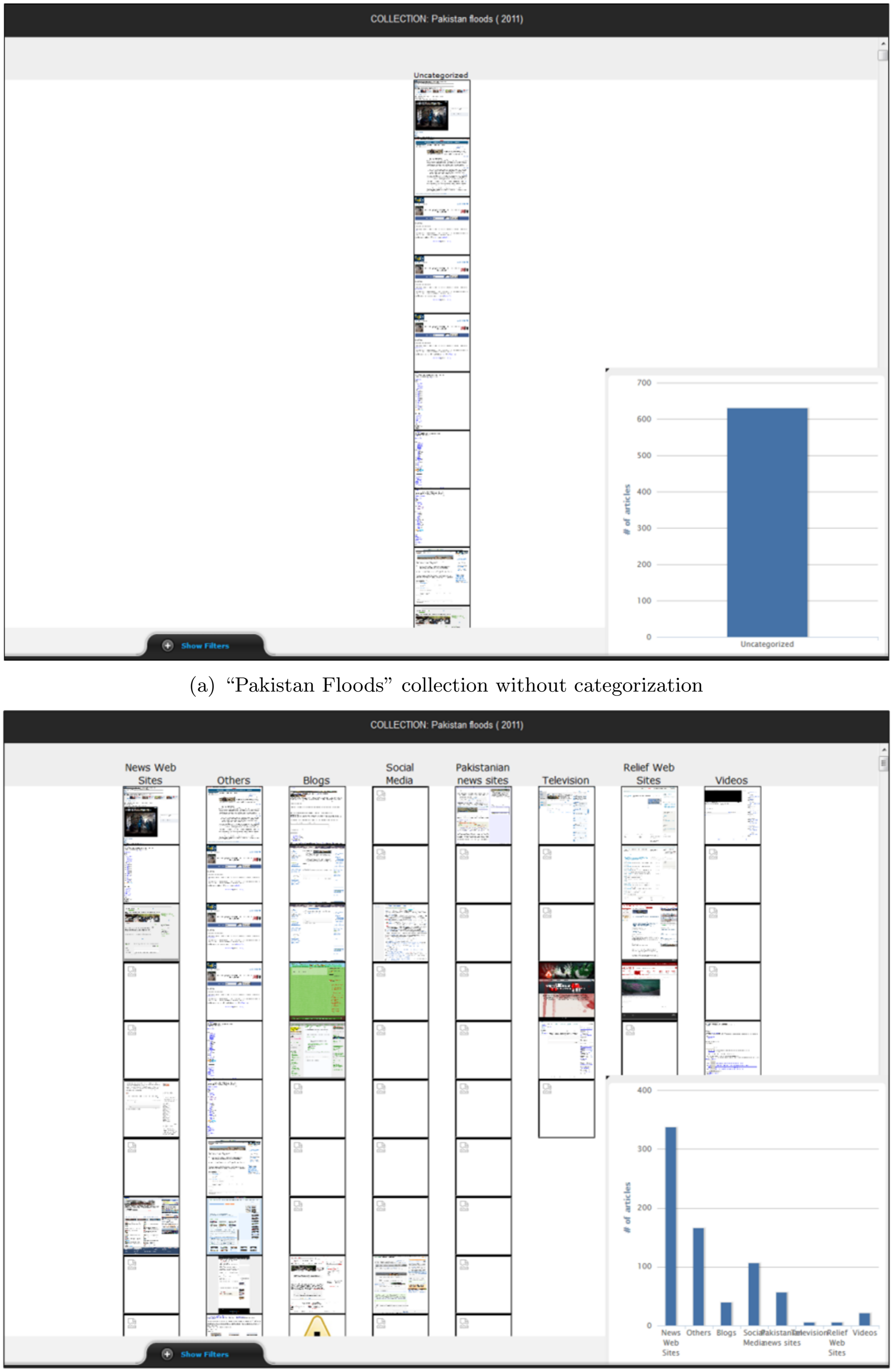
"Amnesty International"

14 mementos taken between 2008-07-21 and 2011-02-10

1. Non-governmental organizations (NGOs)--Amnesty International sites
2. Non-governmental organizations (NGOs)
3. National human rights institutions (NHRIs)
4. Blogs by individuals
5. Truth Commisions, Tribunals, and Courts
human
right
2011
death
penalty
amnesty
release
event
news
action
justice
campaign
woman
photo
report
march
credit
priority
support
international
# of articles
300
200
100
0
1. Non-governmental organizations (NGOs)--Amnesty International sites
2. Non-governmental organizations (NGOs)
3. National human rights institutions (NHRIs)
4. Blogs by individuals
5. Truth Commisions, Tribunals, and Courts
Show Filters

(a) Image plot with histogram, and wordle.

FIG. 37: Different visualizations for exploring Human Rights collection at Archive-It (continued).



(a) "Pakistan Floods" collection without categorization

(b) "Pakistan Floods" collection with categorization

FIG. 38: "Pakistan Floods" collection after and before applying categorization.

an example of “Pakistan Floods” collection before and after categorization in the image plot with histogram visualization.

One problem with the above approaches is that there is often an implicit assumption that everything in a collection is equally valuable and should be visualized. Some of the Web pages change frequently, some are near-duplicates, and some go off-topic and no longer contribute to the collection. Visualization techniques with an emphasis on recall (i.e., “here’s everything in the collection”) do not scale. Instead, we are informed by emerging trends in social media storytelling, which focus on a small number of exemplary pages (i.e., high precision) as chosen by a human, to sample from the collection by choosing representative pages that best exemplify the topic of the collection (Chapter 5).

3.1.3 IMAGE COLLECTION SUMMARIZATION

Because of the rapid growth of image collections, managing and understanding these collections have become necessary and have been handled by many researchers [116, 204, 244, 76, 84, 294, 27].

Nguyen et al. [244] identified three requirements to efficiently dealing with visual large collections: overview, visibility, and structure preservation. In their work, they provided solutions for each requirement and proposed a visualization scheme for interacting with large image collections. They used the structure of the collection as the main focus for creating overview about the collection through dividing the collection into groups using clustering techniques [147], then selecting a representative image from each group. They used a distance matrix for finding the clusters.

Graham et al. [116] introduced different techniques for browsing collections with thousands of time-stamped digital images: Calendar Browser and Hierarchical Browser. They provided clustering techniques based on the timing information that is attached to images’ timestamps to structure the collections. They also provided summarization techniques, so instead of displaying all the images the tool displays a set of representative images to be presented to the users. They specified four rules for choosing good representative images for the collection:

- One image from a sequence of images that have little time between one another.
- Images separated by largest difference between one another, so the photographs right before or after a long time interval is a good candidate.
- Images with high contrast and resolution information.
- The image that best represents the visual characteristics of the cluster in cluster-wide image analysis.

At the end, Graham et al. conducted a study to evaluate their developed browser against an unsummarized browser. The results showed that summarizing the collections led to a significant improvement and users preferred the summarized collection.

In the DSA framework, we use the quality of replaying mementos and the quality of the generated snippet from mementos to select the best representative mementos (Chapter 9).

In addition to using the timing information of digital photo collections, Jaffe et al. [146] considered a multitude of spatial, social, and temporal metadata dimensions for clustering and summarizing large collections. They used geo-referenced digital photographs, whereby photos are connected to metadata describing the geographic location in which they were taken [236], to create summarizations that can be used to assist in map-based browsing of images. They also developed the Tag Maps visualization for large-scale geo-referenced photo collections that exposed the textual topics which were tied to a specific location on a map.

Li et al. [204] proposed a framework for automatic organization and summarization of personal digital photos based on their creation timestamps and image contents. They first applied photo sequence partitioning by time then by content, and then they applied similarity on the color histogram of the images to observe the changes in the photos. They used selection criteria for choosing representative images (e.g., face criterion, time criterion).

Sinha [293, 294] proposed a framework for generating representative subset summaries from photo collections hosted on Web archives or social networks to create an overview summaries from large personal photo collections. They evaluated the framework using 40K personal photos of 16 different users collected from Flickr[4], Picasa[5] and other photo archives. They claimed that an effective subset summary should satisfy these properties: quality, diversity, and coverage. The results showed that summaries generated using their models outperformed baselines considerably.

Chu et al. [76] utilized the near-duplicate detection concept for automatic selection of representative photos. First, they applied time-based clustering technique, then they applied near-duplicate techniques (e.g., SVM-based determination model, orientation feature extraction) for choosing representative images.

AlSum et al. [27] proposed various summarization techniques to optimize the thumbnail creation for TimeMap based on information retrieval techniques. They found that SimHash similarity fingerprints have the best prediction for the visualization change. They

[4] `https://www.flickr.com/`

[5] `http://picasa.google.com/`

proposed three algorithms, threshold grouping, K clustering, and time normalization. They minimized the number of generated thumbnails by 9% - 27% on average.

Most of the image collections summarization techniques start with dividing the image collection by time, then cluster the images by content, and lastly select a representative image from each cluster. In our framework, we also slice the collection then cluster the mementos of each slice, and then based on quality metrics we select a representative page from each cluster (Chapter 9).

3.1.4 VIDEO ABSTRACTION

Multiple techniques of video abstraction have emerged to allow fast browsing of videos [341, 298, 123, 166, 206, 260]. Video abstracting can be either a video summary (still-image abstracts or keyframes) or video skimming (moving-image abstracts) [321, 205]. Numerous works have handled generating video summaries [320, 120, 92, 326, 31, 237, 87, 229, 243, 242] and video skimming [127, 313, 299, 75, 123].

There are different types of video summaries based on how the keyframes are extracted: shot-based, perceptual feature-based, sampling-based, cluster-based, and others. Some of these techniques are similar to the techniques we use in the DSA framework, for example the perceptual feature-based keyframe selection. The perceptual feature-based summaries depend on selecting the keyframes that differ from each others in terms of their features [247]. Examples of these features are the color, shape, motion, etc. [346].

This is similar to the grouping methodology we used for eliminating duplicates in individual TimeMaps (Chapter 9). We select the first memento of the TimeMap and compare it to other subsequent mementos. If the most recent memento exceeds a specific threshold, it is selected to be the current memento that we compare to the subsequent mementos. In the DSA framework, we use the text similarity between the mementos.

Jung et al. [160, 161] proposed a narrative-based abstraction framework for story-oriented videos (e.g., dramas) to understand the overall story of the video. To capture the human understanding of a story they analyzed the scenario writing rules and movies editorial techniques to establish the narrative structure. The model analyzes a story-oriented video, captures the narrative structure, and annotates narrative components according to their degree of progression to the overall story. They evaluated their framework through running multiple experiments to test the viewer's understanding and preference through comparing their method against ground truth dataset.

Video abstraction techniques are similar to what we do to generate summaries from archived collections. We slice the collection then we use the content as a feature to detect

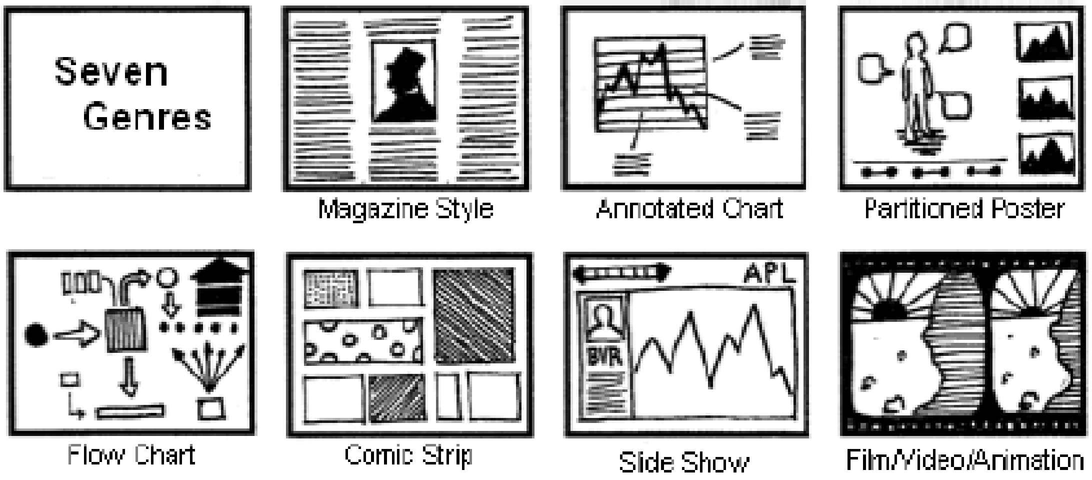


FIG. 39: Genres of narrative visualization [285].

the events based on similarity threshold between each cluster, then we select representative mementos based on selection metrics, and then arrange the selected mementos in a chronological order to compose a summary that gives users an overview of the collection (Chapter 9). We also were inspired by the work of Jung et al. [161] to capture how humans create a story, so we build a baseline for the human-generated stories based on analyzing a dataset of Storify stories.

3.2 TELLING STORIES WITH DATA

The definition of narrative in the Oxford English Dictionary is "an account of a series of events, facts, etc., given in order and with the establishing of connections between them." Stories of this form often have a beginning, middle, and end [323, 226]. Storytelling strategies vary among media and genre [285]. For example, the story in films is different in the structure from the written story which may have more narrative structure. Jonathan Harris defined "story" as follows: "I define 'story' quite loosely. To me, a story can be as small as a gesture or as large as a life."[285]. An event is defined as "an occurrence that happens at a specific time and draws continuous attention" [15].

We briefly review the literature on time-based storytelling techniques highlighting narrative visualizations and time series visualizations.

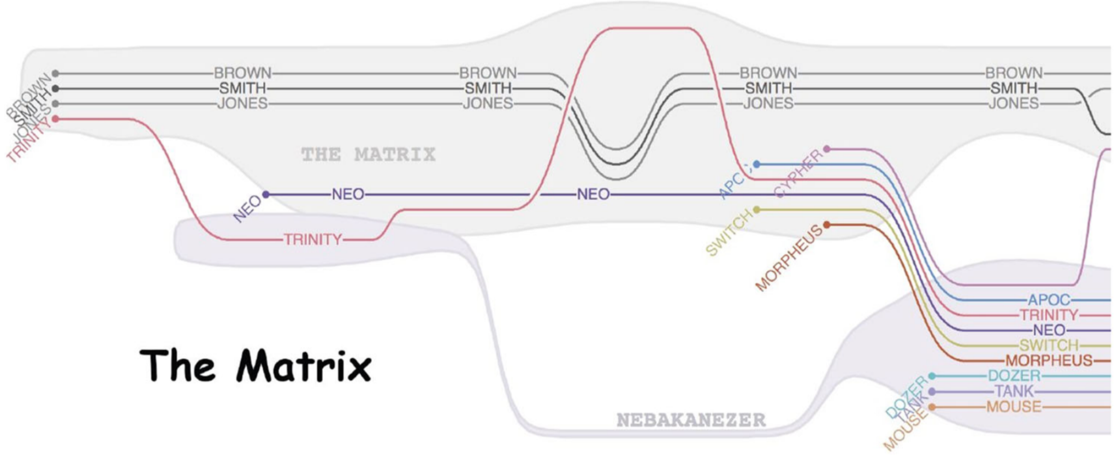


FIG. 40: Storyline visualization of the movie The Matrix [311].

3.2.1 NARRATIVE VISUALIZATIONS

Recently, there has been increased interest in leveraging narrative visualizations [139, 285] and telling stories with data techniques [285, 281, 109, 210].

Segel and Heer [285] investigated the design of narrative visualizations and identified techniques for telling stories with data graphics. They introduced seven genres of narrative visualization (Figure 39): magazine style, annotated chart, partitioned poster, flow chart, comic strip, slide show, and film/video/animation.

Hullman et al. [139] studied the effect of the sequences choices in the narrative visualization on end-users perception, based on a qualitative analysis of 42 narrative visualizations [140]. They studied the characteristics that made a visualization sequence successful. In particular, they focused on how the effects of sequencing style on user perception and message communication can be useful for linear and slideshow-style presentations. They also had previous studies about framing effects in narrative visualization.

Multiple storyline visualizations have been developed to illustrate the dynamic relationships between entities in a story [311, 210]. For example, Figure 40 shows a storyline visualization of the movie “The Matrix” that was developed by Tanahashi and Ma [311]. The main problem with storyline visualizations is the scalability and complexity [210].

Every story is made up of a sequence of events. In our framework, events are represented by Web pages from Archive-It collections, automatically discovered, arranged in a narrative structure ordered by time, and replayed through an appropriate visualization interface.

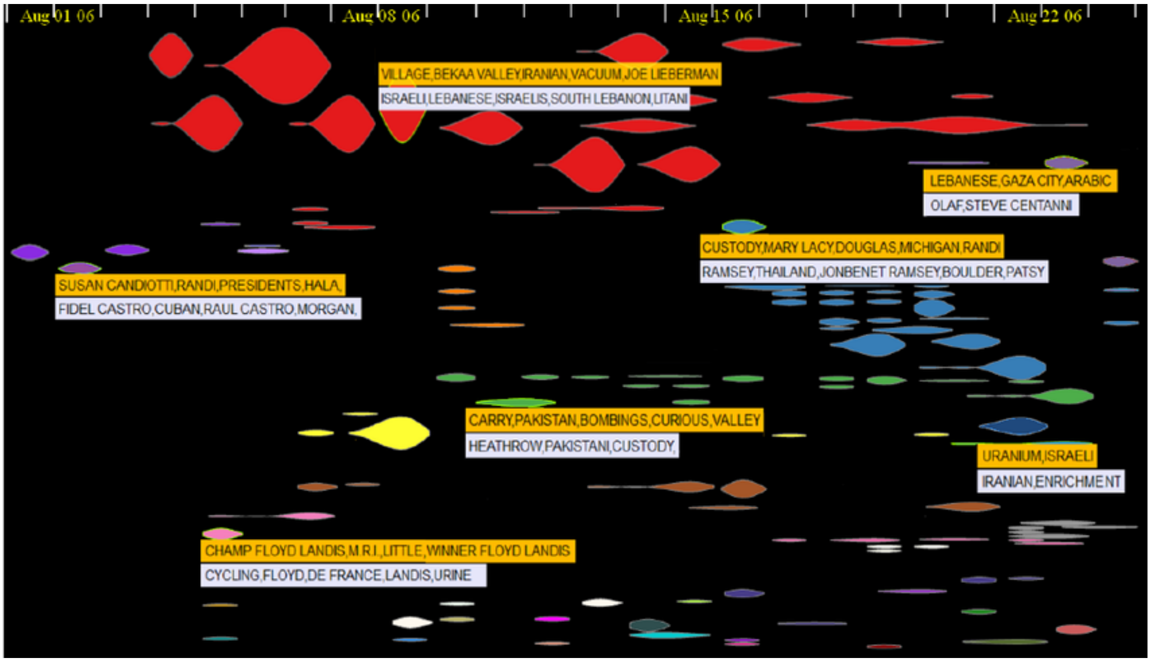


FIG. 41: CNN news from Aug. 1 to 24, 2006 in EventRiver [129].

3.2.2 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 text and its evolution over time [94, 212, 189, 268, 190].

Dou et al. developed LeadLine, an interactive visual analytics system to automatically identify meaningful events in the news and social media data and support exploration of the event [94]. 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 [212]. EventRiver allows users to browse, search, 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 long-term stories in a display that look like a river of events flowing over time (see Figure 41). In EventRiver, Luo et al. used text clustering to group the documents that are coherent in content and adjacent in time. EventRiver is different from ThemeRiver [129], which we explained in Section 3.1.2. ThemeRiver does

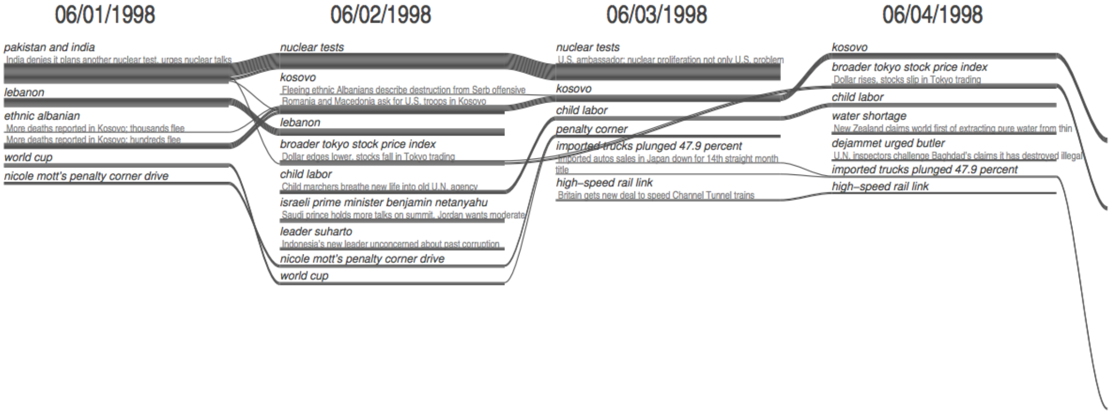


FIG. 42: An example of Story flow visualization [268].

not support event-related tasks which means the occurrence of an event at a specific time, while EventRiver supports event browsing.

There is also a wide range of tools developed for visualizing time series event [189, 190, 268] for example, CloudLines, which is a visual analytics technique to visualize context as a continuous flow [189]. CloudLines provides a compact visualization for time series event data with a lens-based interaction for direct access to overlapping events. Another example is CAST, 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 [268]. In CAST, Rose et al. [268] used a clustering algorithm on extracted keywords from the documents in the corpus and then captured temporal information for tracking and adapting to evolving stories. CAST system uses node-link-based visualization and depicts the topical change over time (Figure 42).

Most of the previous research divided the collections by time and content to reflect the evolution of the corpus through time. In our framework, we slice the collection to predefind number of slices then cluster the pages of each slice by content so each cluster represents a specific event (Chapter 9). We provide an insight into the evolution of archived collections through time through generating broad stories from these temporal collections. Furthermore, we support event-related tasks through allowing generating stories from different URIs at the same time. Such a story is an important information source in a wide variety of applications, such as social and cultural studies.

3.3 INFORMATION RETRIEVAL MEASURES

As the volume of content swirling around the Web continues to grow, it is not easy to specify the materials relevant to a specific topic. Information retrieval is finding the relevant set of documents to a specific request based on the information need of users [59]. In the DSA framework, we adopt information retrieval techniques to specify the set of resources that are central to what the story of the collection is about. We will compute the "aboutness" of the individual pages within the collection to eliminate the non relevant pages. Furthermore, we will adopt multiple techniques for clustering the pages based on their content.

3.3.1 THE NOTION OF ABOUTNESS

The "aboutness" is the description of the intellectual content of documents for retrieval purposes [61]. The aboutness of a document has long been central to Web science and information retrieval (IR) systems, including Web search engines. The IR system's goal is determining how related a document is, in terms of its aboutness, to a user-specified query [29].

Many studies [59, 108, 29, 167] have been performed on the key aspects of aboutness such as the page's titles [179], tags [177], key terms [342, 144, 254], lexical signatures [181], or summaries [220]. A lexical signature (LS) is a small set of terms derived from a document that captures the aboutness of that document [178]. LS has been used widely for finding the missing pages (i.e., HTTP 404) on the Web [255, 181, 178]. Typically a lexical signature of a Web page is the top n terms from the page, sorted by its TF-IDF values [181].

For calculating the aboutness of a collection, there are machine learning statistical models, such as topic modeling and detection tools to discover the abstract topics that occur in a collection of documents [47, 253].

Jatwot and Ishizuka [150] used statistical analysis among the text features for summarizing the content of webpages through time. In a later study, Jatowt et al. [153] proposed methods for detecting the degree of freshness of linked pages based on comparing the pages with the previously viewed pages by users. That resulted in reducing the cost and time of browsing by informing the users with what they have not yet viewed. They incorporated the mechanism of the personalized freshness detection into the browser.

Paranjpe [254] concentrated on document's aboutness using words and phrases presented in the document that best reflect the central topics of that document. She used a machine learning approach to identify the rank of words and phrases in the document according to their relevancy to that document. Paranjpe also used the click data of a search engine.

Another perspective on textual aboutness has been introduced by Kehoe and Gee [167, 29]. They examined social tagging from a corpus linguistic perspective to represent the aboutness of pages, based on data from using from Delicious[6] (a social bookmarking site that allows users to add tags to their bookmarks).

3.3.2 TOPIC DETECTION AND TRACKING (TDT)

Topic Detection and Tracking (TDT) refers to automatic techniques for finding topically related material in streams of data and organizing stories by the events that they discuss [333, 200, 199]. One of the fundamental concepts that distinguishes TDT is the notion of event or topic. In TDT, "a topic is a seminal event or activity along with all directly related events and activities" and an event means something that happens at some specific time and place. Fiscus et al. considered a story is on-topic when it discuss events that are related to the topic's events [104].

TDT evaluation tasks cover many of the topics we use in the DSA framework, such as "link detection", which means detecting clusters of stories that discuss the same topic. There have been many models for link detection, such as statistical language model techniques [202, 199] and vector space approaches [279, 340].

There are also intensive studies that have been conducted for investigating TDT for news Web pages. Lavrenko et al. extended the relevance model from working with short queries to work with stories for comparing two stories in link detection [200, 199]. For measuring the similarity between the two models, they used the Kullback-Leibler divergence [192, 279], which is a standard way to compare two probability distributions.

Mori et al. [235] proposed a new approach for topic tracking from the Web pages that are returned by a search engine. They used the temporal information of the Web pages (the Creation-Datetime and the Last-Modified) to cluster the Web pages and create temporal clusters for the relative events.

We use a clustering algorithm to determine if two pages are about the same event. We also use different similarity measures to detect the off-topic pages (Chapter 8).

3.3.3 SIMILARITY MEASURES

In archived collections, it is possible to find many Web pages about the same event. That results in having many possible candidate pages that contain similar content [175, 340, 339]. According to estimates, as many as 40% of the Web pages are duplicates of other pages

[6] `https://delicious.com/`

[217]. Similarity comparison should be applied on these pages to detect if they are (near-)duplicates or they are talking about the same topic.

There are multiple techniques for measuring the similarity between pages. Cosine Similarity is one of the most well-known and effective similarity measures in IR and text mining. It is based on cosine correlation between two vectors where each vector contains one component for each term in the document [292]. Cosine similarity is based on the vector space model [280].

Other measures of similarity have been used, such as the Dice Coefficient [90], the Levenshtein Edit Distance [203], and the Jaccard similarity coefficient [310].

There have been many studies for detecting (near-)duplicates between documents [132, 230, 318, 46, 216, 259, 193, 262]. SimHash [71] is a useful and efficient hash-based method for detecting the near-duplicates between Web pages based on the difference of the pages' fingerprints. SimHash maps high dimensional vector to an f-bit fingerprint where f is very small, for example, 64. These fingerprints are then used for comparing documents. SimHash is effective in comparing two documents because it is fast [132].

Singh Manku et al. [216] investigated detecting the near-duplicates in Web crawls by comparing the crawled pages with their previous copies. In their work, they illustrated that Simhash is appropriate for detecting near-duplicates from large repositories.

Other near-duplicate detection techniques have proposed, such as Locality Sensitive Hashing (LSH) by Indyk and Motwani [143] and Shingling algorithm by Broder et al. [51, 52].

In Chapter 8, we use multiple similarity measures for detecting off-topic pages in Web archives. We evaluate the methods using multiple evaluation metrics that will be explained in Section 3.3.5. We also use the SimHash method to detect the near-duplicates in individual TimeMaps because of its time efficiency (Chapter 9).

3.3.4 TERM FREQUENCY - INVERSE DOCUMENT FREQUENCY

Term frequency (TF) refers to how often a term appears in a document. The probability that a term that occurs very frequently in a document is likely to be more relevant for that document than a term that occurs less frequently. Inverse Document Frequency (IDF), which is first introduced by Sparck Jones [301], is the number of documents in the collection that contain a specific term. Combined TF and IDF, TF-IDF, reflects how important a word is to a document in a collection by providing an accurate measure of the terms local (within the document) and global (within the entire corpus) importance [267].

The TF-IDF weighting developed for vector space retrieval has shown remarkable effectiveness [267]. We used the TF and TF-IDF weightings for creating Wordles [102, 330] for the collections and the Web pages within the collections in our preliminary work [252] to understand Archive-It collections. We calculate the TF-IDF for mementos then apply the cosine similarity to compare the $aboutness(URI\text{-}R@t_0)$ with $aboutness(URI\text{-}R@t)$ by calculating the similarity between the mementos (Chapter 8).

3.3.5 PERFORMANCE MEASURES

There are multiple methods to evaluate a retrieval system's performance: precision and recall [83, 217], F-measure [217], Discounted Cumulative Gain (DCG) [148, 149], etc. Precision is a measure of specificity, meaning the fraction of retrieved documents that are relevant to the information need. For example, when searching for "Egyptian Revolution", the number of the correctly related documents to the "Egyptian Revolution" divided by the number of all returned results is the precision. Recall is the fraction of the relevant documents that are retrieved divided by all the relevant documents in the corpus (collection). In our previous example, recall will be the number of relevant documents to the Egyptian Revolution divided by all the number of all the documents in the collection that should be returned. F-measure is the harmonic mean of the precision and recall. The DCG is a popular method that is often used in IR for measuring the usefulness or gain of a document based on its rank in the result. The DCG is based on the assumptions that highly relevant documents are more useful when they have higher ranks than less relevant documents. The Normalized DCG (nDCG) is now popular for comparing lists that vary in length and taking the average over multiple queries.

3.3.6 WEB USAGE MINING

The breadth and depth of research in the area of Web usage mining is massive and increasing [35, 194, 295, 66, 221]. Web usage mining involves discovering usage patterns from Web data using data mining [303]. The results obtained from Web usage mining can be used in different applications, such as Web traffic analysis, site modification, system improvement, personalization, business intelligence, and usage characterization. We use Web usage mining techniques to provide traffic analysis and usage characterization for Web archives and extract the abstract models for accessing Web archives (Chapter 4). In this section, we briefly review the related work of Web usage mining. We will highlight the work of usage mining on Web archives data in Section 3.4.1.

Adams et al. [11] explored the usage patterns of scientific and historical data repositories. 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 issue of archiving the Web.

A challenge 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 [309, 91, 196, 117]. Doran et al. [93] characterized robot detection techniques into four categories: syntactical log analysis, traffic pattern analysis, analytical learning techniques, and Turing test systems. In our study of Web archiving usage (Chapter 8), we used syntactical log analysis (simple processing by finding the self-identified robots) and traffic pattern analysis (specifying features for contrasting robots with humans).

3.4 TRENDS IN WEB ARCHIVING

In this section, we highlight the research that has been conducted on mining the past Web.

3.4.1 THE USAGE OF WEB ARCHIVES

Understanding the current demand for access to Web archives can provide insights into how to make the best use of limited archiving and access resources. In our prior work that formed the foundation of the DSA framework, we provided the first analysis of user access to a large Web archive [18]. We analyzed the Web server logs from the Internet Archive's Wayback Machine to extract the user access patterns in Web archives and study why and how people come to Web archives (Chapter 4).

Costa et al. studied the search behavior characterization of Web archives based on a quantitative analysis of the Portuguese Web Archive (PWA) search logs [82, 81]. Costa et al. compared between the search patterns of Web archives and Web search engines. Despite the different information needs of Web archives and Web search engine users, the search patterns for Web archives had shown adoption of Web search engine technologies. They found that most Web archive users conducted short sessions. In our study, the most frequent sessions are composed of one request. One important finding from analyzing the search interactions of the PWA logs is that the users prefer older documents. This is in contrast to what we found, that Web archive users have significant repetitions for requests in 2011 (the year prior to our sample) [18].

What is missing from digital libraries and Web archives and the effect of this on the satisfaction of users' needs and expectations has been widely investigated [317, 63, 348, 291]. Thelwall and Vaughan [317] studied the coverage of the Internet Archive for the Web. The results showed an unintentional international bias in the archive coverage through an uneven representation of different countries in the archive. The reason for unbalanced representation of countries is the visibility of the websites (i.e., the number of inlinks of websites). The results also showed that the language of the websites does not have an effect on how the Internet Archive indexes the websites.

Carmel et al. [63] suggest a tool to dynamically analyze the query logs of the digital library system, identify the missing content queries, and then direct the system to obtain the missing data.

AlSum et al. studied the coverage of twelve Web archives using three datasets from the live Web, Web server access logs of the archives, and full-text search of the archives to create profiles for the twelve archives [28]. They discovered that IA has the largest and widest coverage of all the archives, which matches our results of checking the coverage of other archives in a previous study [16].

3.4.2 MINING THE PAST WEB

Web archives are becoming commonly used in social science and humanities research. Archiving the political process has become popular, both in terms of Web pages [284, 264, 105], and YouTube and blogs [62, 219]. Mining the past Web is different from Web content mining because of the temporal dimension of the archived content [157, 186]. The benefit of utilizing the Web archives for knowledge discovery has been discussed many times [32, 157, 154]. Below, we outline some of the approaches that have been used for mining the past Web using Web archives data.

Jatowt and Tanaka [157] discussed the benefits of utilizing the content of the past Web for knowledge discovery. They discussed two mining tasks on Web archive data: temporal summarization and object history detection. They also presented different measures for analyzing the historical content of pages over a long time frame for choosing the important versions to be mined. They used a vector representation for the textual content of page versions using a weighting method, e.g., term frequency. They presented a change-detection algorithm for detecting the change in the past versions of a page through time.

In a later study, Jatowt et al. [156] proposed an interactive visualization system called Page History Explorer (PHE), an application for providing overviews of the historical content of pages and also exploring their histories. They used change detection algorithms based on the content of archived pages for summarizing the historical content of the page to present only the active content to users. They also extended the usage of term clouds for representing the content of the archived pages.

Figure 43 displays the history view of the BBC Homepage[7] in PHE. This visualization displays clouds of top 20 terms over the specified time period in the top frame. Additionally, tag clouds consisting of up to 20 terms, for smaller time periods, are shown below the top

[7] BBC homepage: `http://www.bbc.co.uk/`

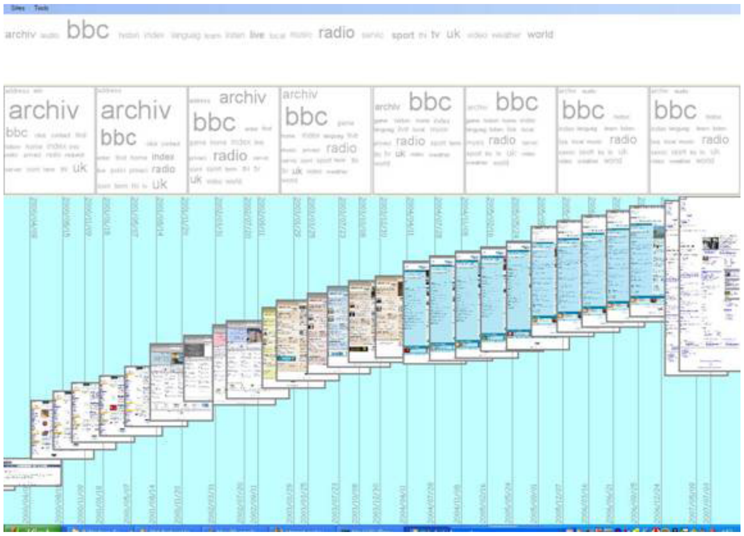


FIG. 43: The history view of BBC homepage (www.bbc.co.uk) in Page History Explorer [155].

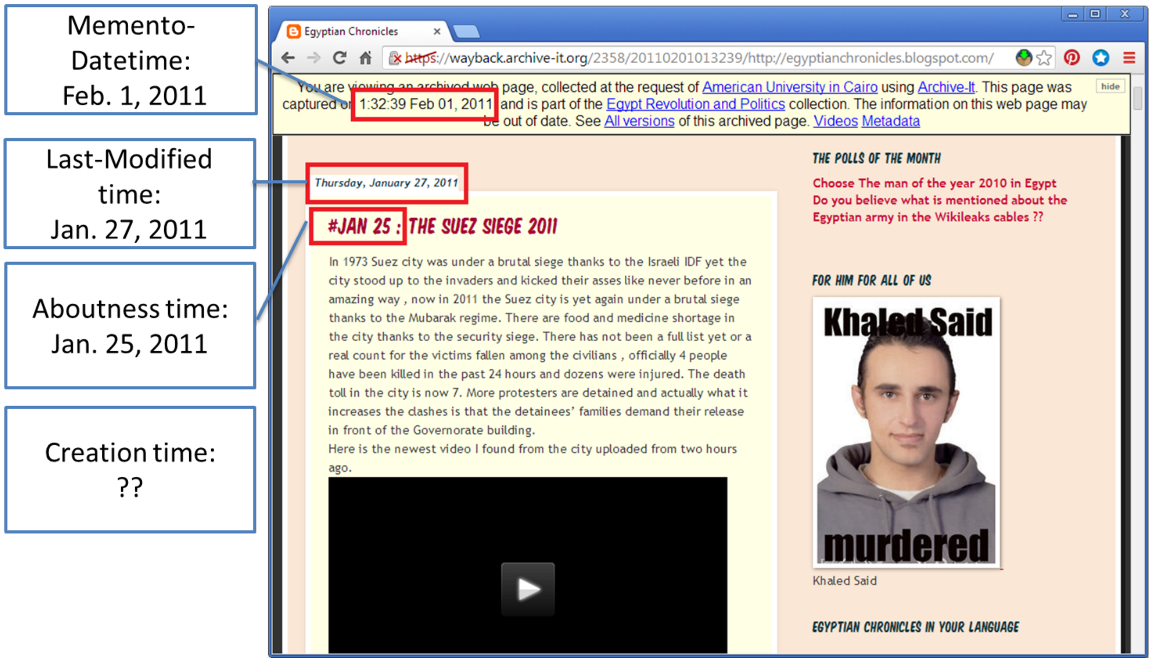
frame. Each snapshot on the timeline represents the view of a Web page as it existed during that period. This visualization system can be used to visualize how a single Web page in a collection changes over time by representing its various mementos over the timeline.

To help people in understanding Web content change, Teevan et al. [316] introduced DiffIE, a browser plug-in that caches the page a user visits, and then detects and highlights any changes to that page since user's last visit. They compared the Document Object Model representation of page's text to highlight the differences.

Tools like PHE and DiffIE are a good way to understand the changes of Web pages through time. In our work of detecting off-topic pages in Web archives (Chapter 8), we are not looking for a deep reading between versions, but rather flagging off-topic pages for non-consideration for other processes (e.g., thumbnail generation [27]).

Spaniol and Weikum [300] used Web archives data to track the evolution of entities (e.g., people, places, things) through time and visualize them. This work is a part of the LAWA project (Longitudinal Analytics of Web Archive data), a focused research project for managing Web archive data and performing large-scale data analytics on Web archive collections. Jatowt et al. [154] also utilized the public archival repositories for automatically detecting the age of Web content through the past snapshots of pages.

Web archiving research has focused on the selection, storage, and preservation of Web content and solving the challenges that face them [225]. Despite the existence of crawl quality tools that focus on directly measurable things like MIME types, response codes, etc., there are no tools to assess if a page has stayed on-topic through time. One of the


Memento-Datetime: Feb. 1, 2011
Last-Modified time: Jan. 27, 2011
Aboutness time: Jan. 25, 2011
Creation time: ??
Egyptian Chronicles
https://wayback.archive-it.org/2358/20110201013239/http://egyptianchronicles.blogspot.com/
hide

You are viewing an archived web page, collected at the request of American University in Cairo using Archive-It. This page was captured on 1:32:39 Feb 01, 2011 and is part of the Egypt Revolution and Politics collection. The information on this web page may be out of date. See All versions of this archived page. Videos Metadata

Thursday, January 27, 2011

#JAN 25 : THE SUEZ SIEGE 2011

In 1973 Suez city was under a brutal siege thanks to the Israeli IDF yet the city stood up to the invaders and kicked their asses like never before in an amazing way , now in 2011 the Suez city is yet again under a brutal siege thanks to the Mubarak regime. There are food and medicine shortage in the city thanks to the security siege. There has not been a full list yet or a real count for the victims fallen among the civilians , officially 4 people have been killed in the past 24 hours and dozens were injured. The death toll in the city is now 7. More protesters are detained and actually what it increases the clashes is that the detainees' families demand their release in front of the Governorate building.
Here is the newest video I found from the city uploaded from two hours ago.

THE POLLS OF THE MONTH

Choose The man of the year 2010 in Egypt
Do you believe what is mentioned about the Egyptian army in the Wikileaks cables ??

FOR HIM FOR ALL OF US

Khaled Said
murdered

Khaled Said

EGYPTIAN CHRONICLES IN YOUR LANGUAGE

FIG. 44: A memento from the "Egypt Revolution and Politics" collection in Archive-It has different notions of times.

parts of the framework in this dissertation is assisting curators in identifying the pages that are off-topic in a TimeMap.

In Chapter 4, we provide the first analysis of user access to a large Web archive. We investigated multiple questions that help us in shaping the problem of the dissertation such as how Web archives are being used and by who, why people come to Web archives, where do people come from, etc.

3.5 DETERMINING DATETIME OF WEB PAGES

Each Web page can have four notions of time [239]:

- Creation-Datetime (CD) is the datetime the resource was created
- Last-Modified (LM) is the datetime the resource last changed
- Memento-Datetime (MD) is the datetime the resource was crawled
- Aboutness Time (AT) is the datetime of an event that the page contains.

An example of the AT is a page published today about events in the past or future. The different notions of time are best exemplified in the example in Figure 44 that shows different times for a page in the "Egypt Revolution and Politics" collection in Archive-It.

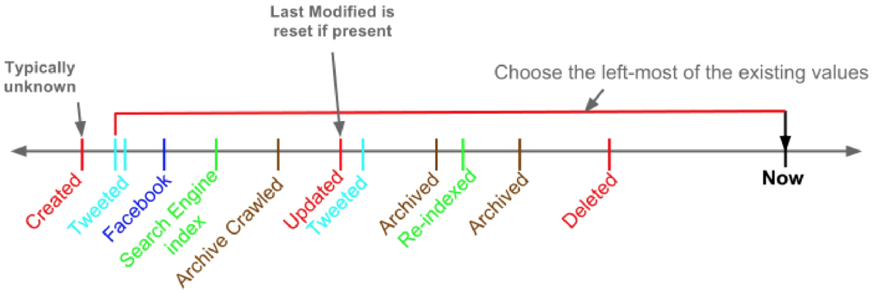


FIG. 45: The timeline of a shared resource and the proposed process of carbon dating [277].

The figure shows a homepage of a blog that was last modified on Jan. 27, 2011 with a post about an event on Jan. 25, 2011. The page was crawled on Feb. 1, 2011, but we do not know exactly when it was created. The times of this page are different from each other (the page is created at t_1, about an event that happened at t_2, modified at t_3, crawled at t_4), which illustrates the need for identifying the different datetimes of the pages in the collection.

There has been research in the area of automatically estimating the creation dates of content elements of pages [152, 154, 157]. 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. [151] and Nunes et al. [246].

SalahEldeen et al. [277] 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. They illustrated a timeline of resources along with how they estimated the age of the resource in Figure 45.

Defining the different notions of a Web page is important. In the DSA framework, we define the different notions of time for the best representative mementos, then we sort them chronologically to be visualized. We will provide more details on how we extract the notions of time of the mementos in Chapter 9.

3.6 SUMMARY

In this chapter, we presented an overview of the research that has been established to summarize different kinds of collections: image collections, document collections, and videos (Section 3.1). We presented how the archival community has attempted to solve the problem of collection understanding through the development of new standards. Then, we provided an overview of the techniques for exploring and understanding the document and image collections. We also presented the related research of telling stories with data focusing on the work of narrative visualizations and time series visualizations (Section 3.2). The principles upon which we build the following chapters, such as the techniques of Information retrieval, document similarity, and Web usage mining are presented in Section 3.3 and Section 3.4. We also presented the related research of Web archives usage and mining the past web. At the end, we present the different notions of time (Section 3.5).

With the knowledge we have gained about the problem of collection understanding and how the previous solutions focused on visualizing everything in the collection without scaling, we proceed in the next chapters with our proposed solution that depend on selecting the best representative pages to summarize an archived collections.

CHAPTER 4

HOW PEOPLE USE WEB ARCHIVES

In this chapter, we present our preliminary work in gaining an understanding about how people use Web archives. We answer these questions: How do people use Web archives? Why do users come to Web archives? Where do Web archive users come from? Why do sites link to the past? Does the destination affect the number of pages the users browse, or does it affect the duration that the users spend on the archive?

Section 4.1 handles how people access the Wayback Machine to understand the user access models of Web archives through analyzing user access logs of the IA's Wayback Machine [18]. We also studied the linking to Web archives and why people come to Web archives (Section 4.2). In the next sections, we examine each of these aspects and explain how they are shaping our understanding of the problem that we are studying [17, 16].

4.1 USER ACCESS PATTERNS IN WEB ARCHIVES

In our desire to provide better archive interfaces, we first begin by examining how archives are used in the absence of interface tools. We also planned to look for any correlation between archival usage and current events but we did not find promising results to complete this study.

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 browse Web archives?
- Do they have extended browsing sessions, going 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 of the previous two patterns?
- Are robot accesses similar to human accesses?

```
0.247.222.86 - - [02/Feb/2012:07:03:46 +0000] "GET
http://wayback.archive.org/web/*/http://www.cnn.com HTTP/1.1"
200 96433 "http://www.archive.org/web/web.php" "Mozilla/5.0
(Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko)
Chrome/16.0.912.77 Safari/535.7"

0.247.222.86 - - [02/Feb/2012:07:03:55 +0000] "GET
http://web.archive.org/web/20130318135600/http://www.cnn.com/ HTTP/1.1"
200 18875 "http://wayback.archive.org/web/*/http://www.cnn.com"
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8)  AppleWebKit/535.7
(KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"}

0.179.81.310_0 - - [02/Feb/2012:13:46:16 +0000] "GET
http://wayback.archive.org/web/20071015000000*/http://9gag.com HTTP/1.1"
200 118819 "http://fr.wikipedia.org/wiki/9gag" "Mozilla/5.0
(Windows NT 5.1; rv:9.0.1) Gecko/20100101 Firefox/9.0.1"

0.251.197.1210_0 - - [02/Feb/2012:18:40:57 +0000] "GET
http://web.archive.org/web/20071008113630/http://www.filg.uj.edu.pl/
ifa/przeklad/przeklad2/poezja2.html HTTP/1.1" 200 25335
"http://info-poland.buffalo.edu/web/arts_culture/literature/poetry/
szymborska/poems/link.shtml" "Mozilla/4.0 (compatible; MSIE 8.0;
Windows NT 5.1; Trident/4.0; .NET CLR 1.1.4322; .NET CLR 2.0.50727)"

0.83.5.950_0 - - [02/Feb/2012:03:18:56 +0000] "GET
http://web.archive.org/ HTTP/1.1" 302 0
"http://www.google.co.uk/search?gcx=c&sourceid=chrome&ie=UTF-8
&q=website+archiver" "Mozilla/5.0 (X11; Linux i686)
AppleWebKit/535.1 (KHTML, like Gecko) Chrome/14.0.835.186 Safari/535.1"
```

FIG. 46: Sample of the Wayback Machine access log (line breaks and new lines added for readability).

4.1.1 WAYBACK MACHINE ACCESS LOGS

A web server log file is a plain text file that records the activity of the submitted requests from users of the web server. The Wayback Machine access logs contain the following fields[1]: client IP address, access time, HTTP request method (GET or HEAD), URI, the protocol (HTTP), HTTP status code (200, 404, etc.), bytes sent, referring URI, and User-Agent. A segment of five requests from the Wayback Machine server log is shown in Figure 46. The first example is a request for a TimeMap, while the second one is a request for a memento. The last three examples are different cases for how the users linked to the Wayback Machine. In the third example, the referrer is Wikipedia, which links to a partial TimeMap (TimeMap for a year only). The fourth example shows an example of an external referrer. The fifth request shows an example of a Google referrer.

4.1.2 METHODOLOGY

The Wayback access logs were sampled using two probability techniques [315, 171]: cluster sampling, which is choosing a cluster of data randomly, and random sampling, where each sampling unit has an equal chance of being included. We performed cluster sampling by choosing a week (Feb. 2-8, 2012) and random sampling by taking a random slice from each day of that week. Each sample comprised a slice of 2M requests to the Wayback Machine Web server.

Then, we applied Web usage mining techniques [35, 194, 295, 66, 303, 232, 159] on the logs to extract user access patterns for Web archives from the Wayback access logs. We first applied data preprocessing techniques (data cleaning, user identification, session identification) to determine the server sessions from the log file [79]. Then, we performed feature extraction, robot detection [309, 91, 196, 117, 93], and user access pattern detection.

4.1.3 ABSTRACT MODELS FOR ACCESSING WEB ARCHIVES

Based on analyzing samples from the Web server logs of the Internet Archive's Wayback Machine (Figure 46), we provided answers for the previous questions [18]. Through our analysis, we discovered four major patterns for Web archive access (Figure 47).

[1] Apache Combined Log File Format: `https://httpd.apache.org/docs/1.3/logs.html#combined`

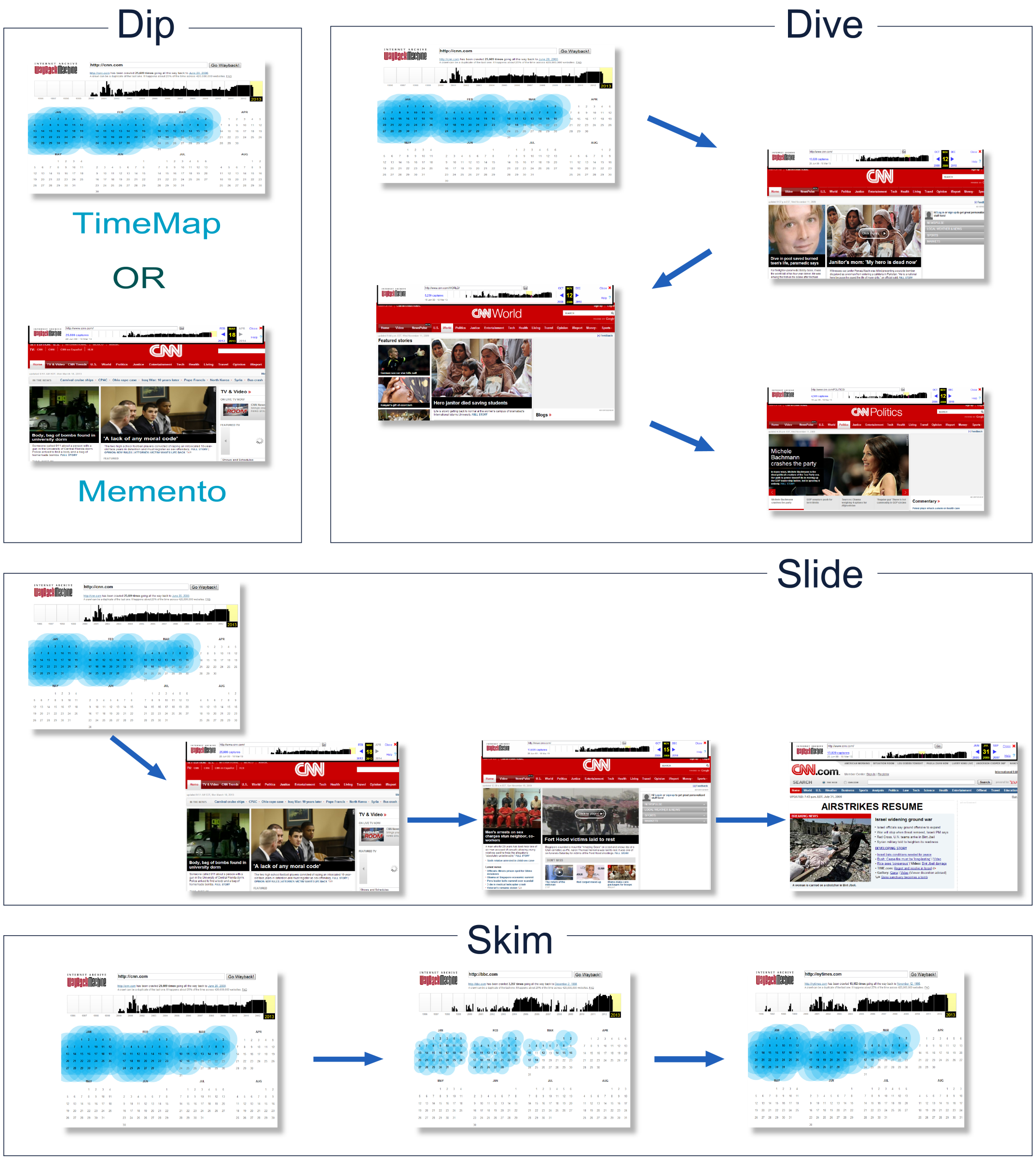


FIG. 47: User access patterns in Web archives (Dip, Dive, Slide, and Skim).

Pattern 1: Dip

Dip is the pattern where a user accesses only one URI. The request can be for a URI-T or a URI-M.

$$Dip = \{\text{URI-X}_i |\ i = 1 \text{ and URI-X} \in \{\text{URI-T, URI-M}\}\} \tag{1}$$

Pattern 2: Slide

Slide is the pattern in which a user accesses the same URI-R at different Memento-Datetimes. In this pattern, the user requests a URI-R and walks through time browsing its different copies.

$$\begin{aligned} Slide &= \{\text{URI-X}_i |\ i > 1, \text{URI-X} \in \{\text{URI-T, URI-M}\} \text{ and} \\ &\quad \text{URI-R}(\text{URI-X}_i) = \text{URI-R}(\text{URI-X}_{i-1})\} \end{aligned} \tag{2}$$

Navigation between different URI-Ms can be done in many ways, e.g., directly from URI-M_1 to URI-M_2 (URI-R@t_1 ⇒ URI-R@t_2) or from URI-M_1 to URI-M_2, but in the middle the user returns to the TM URI-R to choose between the available datetimes (URI-R@t_1 ⇒ URI-T ⇒ URI-R@t_2).

Pattern 3: Dive

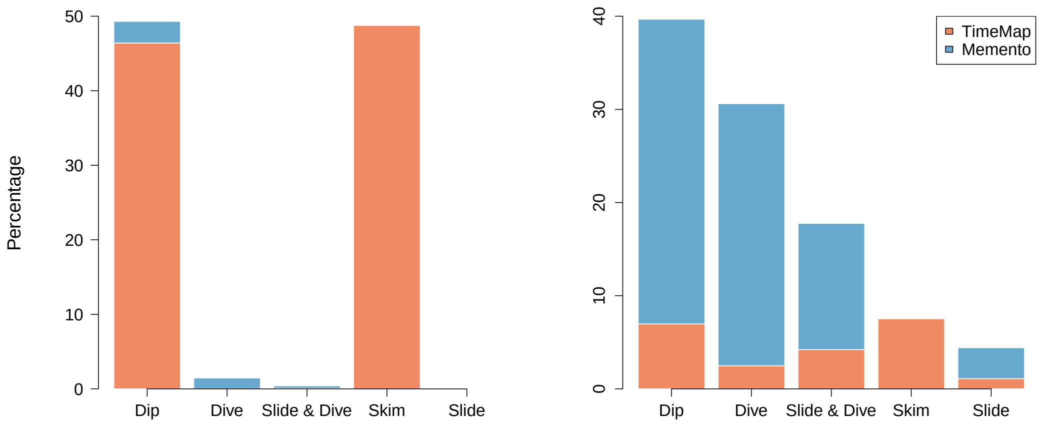
Dive is when a user accesses different URI-Rs at nearly the same datetime. In this pattern, the user accesses one URI-R at a specific time, URI-R_1@t_0, then navigates to different hyperlink(s) of URI-R_1's page (e.g., URI-R_2@t_0) and so on.

$$\begin{aligned} Dive &= \{\text{URI-X}_i |\ i > 1, \text{URI-X} \in \{\text{URI-T, URI-M}\} \text{ and} \\ &\quad \text{URI-R}(\text{URI-M}_i) \neq \text{URI-R}(\text{URI-M}_{i-1})\} \end{aligned} \tag{3}$$

Pattern 4: Skim

Skim is when a user accesses a number of different TimeMaps for different URI-Rs. Skim does not include any access for mementos.

$$Skim = \{\text{URI-X}_i |\ i > 1 \text{ and URI-X} \in \{\text{URI-T}\}\} \tag{4}$$



(a) Robots (34203 sessions)

(b) Humans (3431 sessions)

FIG. 48: Robots and humans exhibit different access patterns.

4.1.4 ROBOT VS. HUMAN ACCESS PATTERNS

Because of the increasing numbers of Web crawlers that are engaged in Web harvesting, many studies have been conducted for investigating the robot detection problem [309, 196]. We used different types of robot detection techniques [245, 312, 263, 64, 306, 93]. First, we applied syntactical log analysis by checking the User-Agent field to identify the self-identified robots. Second, we applied traffic pattern analysis techniques to distinguish humans from robots based on their navigational behavior.

We found 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). Robot accesses can be improved via APIs [26, 25, 36, 38], and the low number of human accesses suggests that better discovery tools are needed.

4.1.5 QUANTIFICATION OF THE WEB ARCHIVE USER ACCESS PATTERNS

We used the Web access logs we described in Section 4.1.2 to quantify the user access patterns for Web archives. We extracted the requested URIs for each session then we identified them based on their type, URI-M or URI-T. We also extracted the URI-R of each requested URI to compare it with the other URI-Rs from the same session. Because of the existence of different forms of URIs which refer to the same Web site [228], we applied URI canonicalization for the URI-Rs to normalize them under one host [85]. The percentages of each of the four patterns exhibited in robot and human sessions are shown in Figures 48(a) and 48(b) along with the percentages of requests to TimeMaps and mementos for each pattern.

Dip

Dip represents the most repeated pattern for humans (33% of all sessions) and robots (49% of all sessions). URI-Ms contribute to 83% of human sessions that exhibit the Dip pattern, although 94% of the robot Dips are requests for URI-Ts.

Slide

There are only a few humans who access the Web archives broadly then navigate away (4.2% of all sessions). Robot sessions do not have this pattern with a noticeable percentage (0.1% of all sessions).

Dive

Dive represents the second highest percentage of human sessions, 29.7%. The robot sessions which were composed of this pattern crawl the Web sites deeply, but they are not a significant number of sessions. Ainsworth et al. looked at the temporal coherence of mementos and the temporal drift (i.e., the difference between the target datetime originally required and the Memento-Datetime returned by an archive) in the browsing Web the archives [14]. They found that embedded resources have Memento-Datetimes that are different from the datetimes of the embedding HTML mementos. We suggest that using actual user walks within the Web archives (Dives) will guide the study of temporal drift in Web archives through analyzing actual user experience.

Skim

Robot sessions exhibit this pattern 48.7% of the time.

Slide and Dive

A large number of human sessions consist of at least two occurrences of the Dive and Slide patterns. In these sessions, the users request URI-R$_1$ and browse its different copies at different times (URI-R$_1$@t_1 $\Rightarrow$ URI-R$_1$@t_2 $\Rightarrow$ URI-R$_1$@t_3), then dive through a hyperlink (URI-R$_2$@t_3) from URI-R$_1$@t_3, then repeat Dive or Slide. In contrast, users may start

TABLE 2: The length of all Slides, Dives, and Skims

User	Pattern	Median	Mean	SD
Robots	Slide	3	3	1.4
	Dive	3	15	53.2
	Skim	3	21	267.0
Humans	Slide	3	4	3.4
	Dive	4	8	14.3
	Skim	3	6	7.2

by going deeply through different mementos for different URI-Rs (Dive pattern), then go broadly through one of these mementos to browse other captures at different times (Slide pattern) (e.g., URI-R$_1$@t_1 ⇒ URI-R$_2$@t_1 ⇒ URI-R$_3$@t_1 ⇒ URI-R$_3$@t_2, etc.). The percentage of human sessions that were composed of a combination of these two patterns is 17.2%. We calculated the number of Slides and Dives for these sessions and found 1167 Slides and 1942 Dives. For robot sessions that were composed of Slide and Dive, we found 328 Slides and 571 Dives.

Pattern Length

Each pattern is made up of a number of requests, which we call the pattern length. We calculated the pattern length for all sessions. The median, mean, and standard deviation of the lengths of each pattern for robots and humans are summarized in Table 2. Humans do longer Dives than Slides and Skims, while robots do longer Skims than Dives and Slides.

4.1.6 TEMPORAL ANALYSIS

Figure 49 shows both the unique and total number of mementos referenced grouped by the year of their Memento-Datetime. Although there is no clear temporal preference for any one year of the unique mementos, there were a significant number of repeated requests for mementos from 2011. This locality of reference suggests that there is an important benefit to be gained by caching the mementos from the recent past. Figure 50 shows that the total number of mementos available for 2011 was similar to previous years. In both Figures 49 and 50, pre-2001 data is included although in those years the archives are too sparse for meaningful comparison with later years.

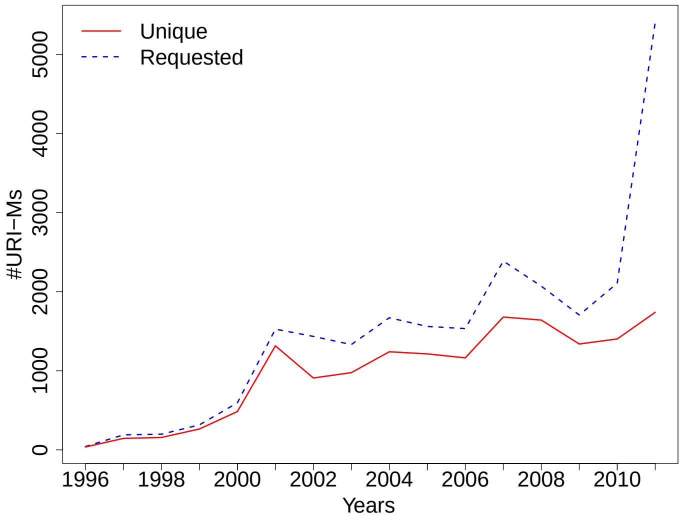


FIG. 49: Distributions of the years for the unique and requested mementos by humans.

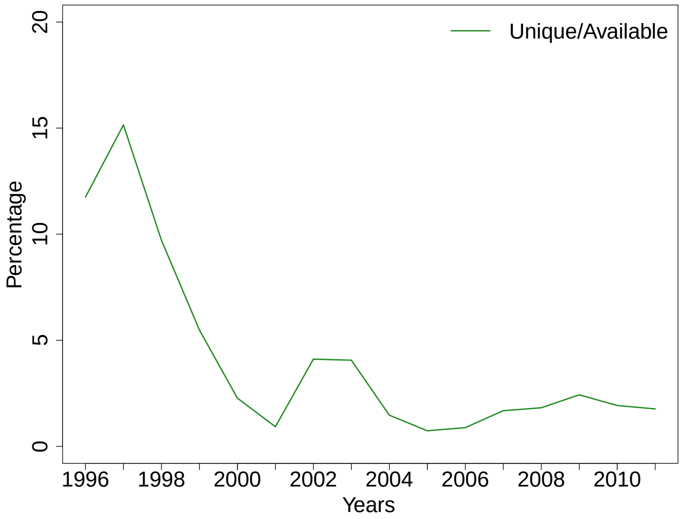


FIG. 50: The proportion of unique URI-Ms requested out of the potential requested for each year.

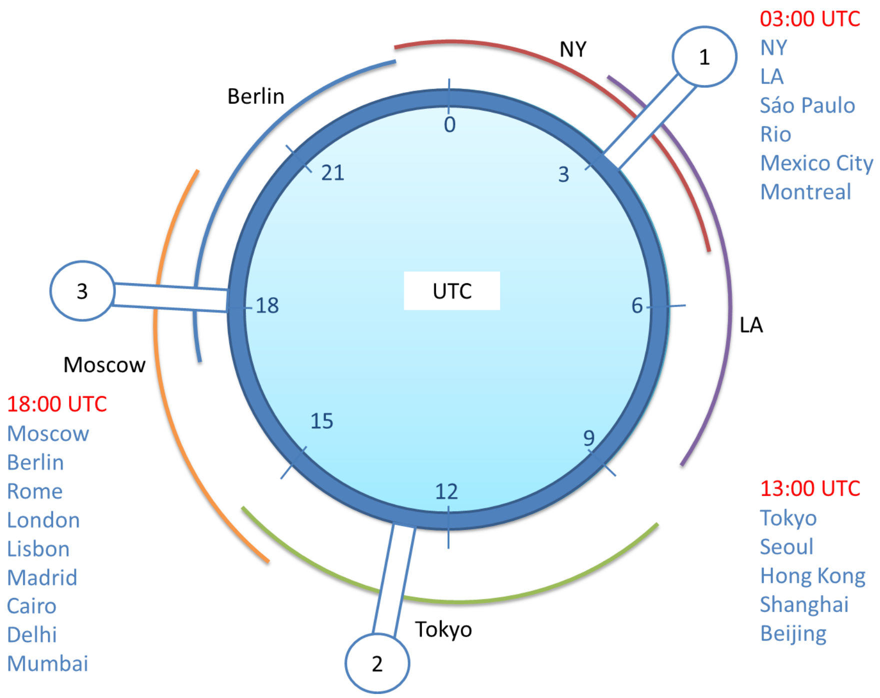


FIG. 51: The dataset of 6M HTTP requests is constructed from slices of 2M each from 03:00, 13:00, and 18:00 UTC on February 2, 2012. The peak hours in NY, LA, Tokyo, Moscow, and Berlin are indicated by arcs.

4.2 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 [16, 17]:

- 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?
- Why do sites link to the past?
- Does the referrer affect the length of the sessions?

4.2.1 METHODOLOGY

Because we checked the language of the content of requested pages, we picked samples from the Wayback Machine access logs that are representative for the peak times of major cities around the world, as shown in Figure 51. These samples covered the peak times of Internet traffic for many countries with speakers of different languages to avoid biasing the results. According to previous studies, the hours between 6 p.m. to 12 a.m. (i.e., midnight) are considered to be peak times for Internet traffic [265, 107, 332]. Home internet use has been well-studied, at least in the United States, and reveals that people engage in a wide range of activities, including commerce, entertainment, job and career enrichment, classes, and news [135, 297]. Note that even though we focused on choosing samples that cover the peak times in multiple cities, each sample also covers work hours for other cities of the world. For example, the 13:00 UTC sample that covers the peak time of Moscow, Berlin, etc., will also cover the work hours for New York City (8am Eastern Time). Note that the IA anonymized the client IP addresses, so it is not possible to geolocate the incoming requests. Furthermore, in the interest of further protecting the anonymity of their users, the Internet Archive recently announced they are encrypting all traffic to their site [50, 163, 307].

4.2.2 LANGUAGES USED IN THE WAYBACK MACHINE

Upon analyzing the user access logs, we identified 52 different languages from the successful requests [17]. We found that English is the most used language on the Wayback Machine, followed by many European languages. We noticed that despite the existence of Web archives in Europe, the requests to the IA from speakers of European languages represent 13% of the top 10 list for human requested pages and 18.5% of the top 10 list for the robot requests. We assume that this is because of the popularity of the Internet Archive, so most of the people who know about Web archiving may only know about the IA.

4.2.3 TEMPORAL DISTRIBUTION OF THE REFERRED URI-MS

Figure 52(a) shows the total number of mementos which were pointed to by the referrers, grouped by the year of their Memento-Datetime. There is a significant bias toward 2008, then 2007, and then a bias against the more distant past. We found 14 URI-Ms all from a single Web site that link to a datetime in 2099. We assume that the referrer wants to redirect the site's visitors to the most recent copy of the linked Web page.

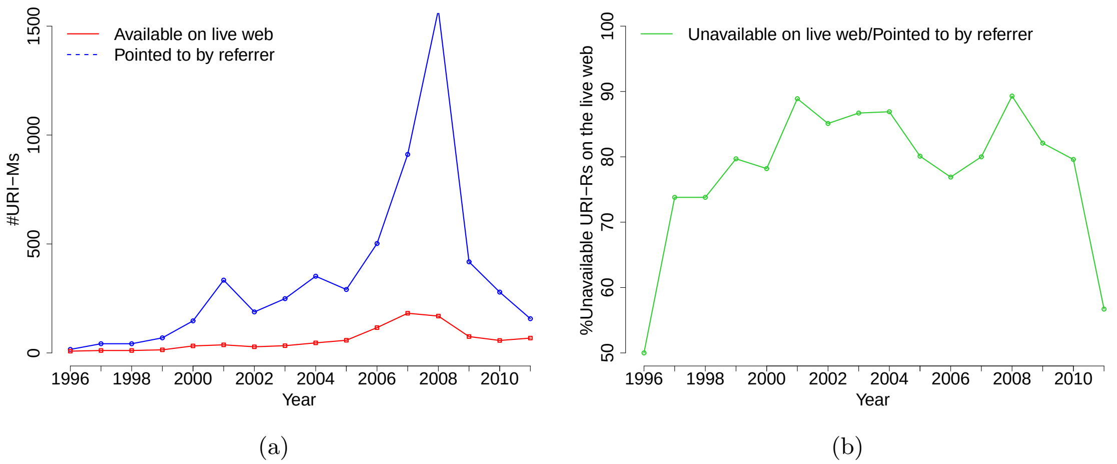


FIG. 52: (a) The temporal distribution of URI-Ms pointed to by the referrers and the number of relative URI-Rs of these URI-Ms that are currently available on the live Web. (b) The percentage of unavailable URI-Rs of these URI-Ms on the live Web.

4.2.4 WHY DO WEB SITES LINK TO THE WAYBACK MACHINE?

Because the Web is ephemeral and the expected lifetime of a Web page is short, Web archives are important to webmasters and third parties for preserving and saving many Web sites. Figure 52(b) clarifies that most people link to the Wayback Machine because they did not find the pages on the live Web. The figure shows that for most of the years, more than 70% of the referred pages in the archive no longer exist on the live Web. About 83% of all referred-to URI-Rs do not currently exist on the live Web.

4.2.5 DOES THE REFERRER AFFECT THE SESSION LENGTH AND DURATION?

When we started to analyze the Wayback machine access logs we expected to see long browsing sessions that may have a story. However, we found that of all the sessions, 50% were composed of one request only (Dips). We investigated if the type of the referrer could be a reason for these Dips.

In this section, we give an analysis of the sessions after dividing them based on their destination (i.e., the referrer field) into four categories: sessions from external Web sites, sessions from search engines, sessions from the archive homepage, sessions with no referrer (e.g., sessions that came from direct address such as a link in an email, etc.). For sessions with no referrer, which we call “direct address”, there is not much information about how they link to the archive (e.g., link in an email or bookmarking) from the logs.

TABLE 3: The median and the mean of session length and session duration of the sessions that were divided based on the referrer.

	Session Length		**Session Duration in seconds**		**Empirically Enhanced Session Duration in seconds**	
Referrer	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**
External Sites	1	2.9	74	171.2	71	99.3
Search Engines	6	11.4	92	190.3	79	176.8
Archive Home-Page	6	11.3	95	199.9	73	177.6
Direct Address	2	7.2	136	326.2	61	215.8

Session Length

We found that 77% of the Dips sessions came from external Web sites. Table 3 shows a summary of median and mean values for the session lengths and durations of the four categories of sessions. Note that session duration in the middle do not include the one request sessions, and the last two columns represent the session duration with estimating the one request session using the mean value of the inter-request time of the two-requests sessions. The mean of each group of sessions are: 71 seconds for the sessions from external Web site, 16 seconds for the sessions from search engines, 10 seconds for the sessions from the archive homepage, and 77 seconds for the sessions from direct address.

The left two columns of Table 3 show that the median and mean values of session length for the sessions that came from search engines and the archive homepage are much larger than the median and mean values that came from the external Web sites. That means the people who know about the archive browse more pages than the users who come from external Web sites. The sessions that come from direct address also have longer session lengths than the sessions from the external Web sites.

It is rare to have a long session length when referred by an external Web site. However, the three-request sessions represent the highest percentage, with 12% of the sessions that came by the search engines. The Dips represent only 7% of all the sessions that came through the search engines. Of the sessions that started on the archive homepage, 11% are Dips.

Session Duration

Table 3 shows a summary of the effect of the destination on the session duration. The two columns in the middle contain the medians and the means for each group of sessions, excluding the sessions that were composed of one request only. We notice that the sessions that came from direct address had longer durations than the rest of the groups, furthermore, the smallest median and mean are for the sessions that came from the external referrers.

Since the one-request sessions represented a large portion of the sessions that came from external Web sites (64%), we did not want to totally disregard them. Thus, we estimated a value for the duration of the session that were composed of one-request based on the inter-request time between the two requests of the two-request sessions. We calculated the mean of the two-request sessions of each group and assigned this value to the one request sessions, then recalculated the median and the mean of the sessions, which we named "Empirically Enhanced Session Duration". The mean of inter-request time of the two-request sessions of the four groups of sessions are: 71 seconds for the sessions from external Web sites, 16 seconds for the sessions from the search engine, 10 seconds for the sessions from the archive homepage, and 77 seconds for the sessions from direct address. The results are shown in the rightmost columns of Table 3. There is a significant difference between the values of the mean before and after adding the estimated duration of the one-request sessions, especially for the sessions from the external Web sites.

4.3 SUMMARY

One of the concerns in the Web archiving world is how to generate more interest in and use of Web archives. We studied how and why people browse Web archives to gain insight about the user access patterns in Web archives based on samples from the Internet Archive's public Wayback Machine. In our studies, we noticed that Web archives are not well-known by the general Web population, and those who do know about Web archives consider them difficult to use. We found that although the Internet Archive's Wayback Machine receives more than 82 million requests per day, based on our dataset, robots outnumber humans 10:1 [18]. Furthermore, the humans that visit the Internet Archive's Wayback Machine typically visit a single page and then leave; depending on the source this can be as often as 64% of the time (in Web analytics terminology, this is known as an undesirably high "bounce rate") [16, 17]. These results indicate the need for tools that support increased archive exploration by humans.

We identified four major Web archive access patterns: Dip (a single access), Slide (the same page at different archive times), Dive (different pages at approximately the same archive time), and Skim (lists of what pages are archived, i.e., TimeMaps) [18]. Robots are limited almost exclusively to Dips and Skims, but human accesses are more varied between all four types. We also uncovered the temporal preference of unique archived Web pages and found that no overall preference for a particular time, but the recent past (within the last year) shows significant repeat accesses.

We also found that most human users come to Web archives because they do not find the requested pages on the live Web [16]. About 65% of the requested archived pages no longer exist on the live Web. We find that more than 82% of human sessions connect to the Wayback Machine via referrals from other Web sites, while only 15% of robots have referrers. Most of the links (86%) from Web sites are to individual archived pages at specific points in time, and of those, 83% no longer exist on the live Web. Finally, we find that users who come from search engines browse more pages than users who come from external Web sites.

To help users in understanding the holdings of the archived collections, we provide a new framework that generate summaries of those collections (semi-)automatically in the next chapters. We also provide tools to assist the curators for detecting the off-topic pages in Web archives and increase the quality of these collections for utilizing their content for knowledge-discovery.

CHAPTER 5

THE DSA FRAMEWORK: GENERATING STORIES FROM ARCHIVED COLLECTIONS

This chapter describes the abstract model for generating stories from archived collections, along with the terminology and definitions that represent the basics of the Dark and Stormy Archives[1] (DSA) framework. Storytelling has become a popular technique in social media for selecting resources (e.g., tweets, videos, Web pages) and arranging them to create a narrative or a story of a particular topic of interest. Every story is made up of a sequence of events. In the DSA framework, events are exemplified by corresponding Web pages from Archive-It collections, automatically discovered, arranged in a narrative structure ordered by time, and replayed through an appropriate visualization interface. The DSA framework also provides tools for computing the "aboutness" of the pages in the collection, and then detecting the off-topic archived pages. Our plan of work is motivated by a likely usage scenario of trying to discover and browse the mementos that represent a story, such as those mementos in Figures 12 - 20 of Chapter 1, and then pushing them to existing tools, such as Storify (Figures 53 and 54). A usage scenario for generating a story from an archived collection is presented in Section 5.1. Section 5.2 contains the definitions and terminology of the DSA framework that will be adopted in the rest of dissertation. Section 5.3 describes different possible types of stories that can be extracted from archived collections. The methodology to achieve the usage scenario will be summarized in Section 5.4, then detailed in the following chapters.

5.1 USAGE SCENARIO

Our research goal can be summarized with the following scenario. Lori wants to have the story of the Egyptian Revolution to show to her children when they grow, as narrated by Figures 12, 14, and 16 of Chapter 1. She knows that many of the Web pages that make up the story will not survive long enough to show her children, so she uses Web archives. She knows about Archive-It collections, but it is not easy to browse the collection to create an overview from the seed URIs in the collection. Through a Web-based interface that is integrated with Archive-It, as an output of the DSA framework, she will easily create stories automatically. Those stories will provide different perspectives about the collection

[1] Inspired by "It was a dark and stormy night", a well-known storytelling trope: `https://en.wikipedia.org/wiki/It_was_a_dark_and_stormy_night/`

with the flexibility to specify different parameters, such as the source type, the periods of the story, etc.

The first step will be building a baseline for the characteristics of human-generated stories and for what we can sample from the archived collections. The next step is computing the “aboutness” of the pages to exclude the non-relevant pages and eliminate the duplicates. The last step will be dynamically finding the set of relevant Web pages that best represent the collection based on the kind of story Lori wants to create. In this step, the datetimes of the Web pages should be determined. Suppose the software finds more than one related page for each event of the story, then the software would choose the best candidates for each event of the story and then visualize these candidates using Storify, with which Lori is already familiar (Figures 53 and 54). The software will also provide other visualizations for the story. Lori has the ability to generate different stories, and she is also able to specify the boundary times of the story. After creating the story, Lori can save and share the story with her friends.

5.2 CONVENTIONS AND DEFINITIONS OF THE DSA FRAMEWORK

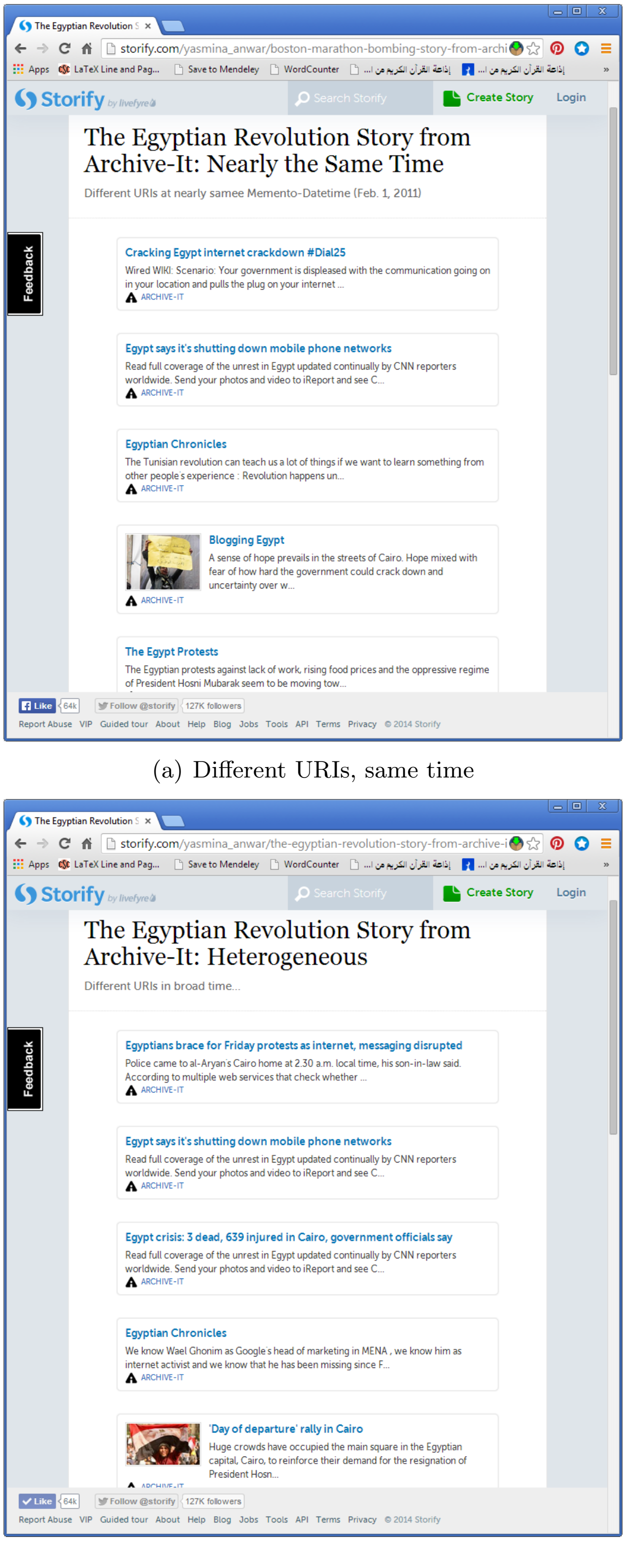
In this section, we give the notions and the conventions we will use for defining the DSA framework. It is possible for a collection to be summarized with more than one kind of story (depending on the nature of the collection as well as curator or user preferences). Before specifying the possible types of stories (Section 5.3), we first define the archived collections.

An Archive-It collection (C) is a set of seed URIs collected by the users from the Web (W), where $C \subset W$. Each seed URI has mementos.

A collection C can be formally defined as following:

$$\begin{aligned} C &= \{URI\text{-}T_1, URI\text{-}T_2, ..., URI\text{-}T_n\} \text{ where} \\ & \quad URI\text{-}T = \{URI\text{-}M_1, URI\text{-}M_2, ..., URI\text{-}M_x\} \\ & \quad \text{ and } URI\text{-}M_i \text{ is } URI\text{-}R@t_i \end{aligned} \tag{5}$$

In the DSA framework, we apply IR and machine learning techniques to identify and select different sets of k mementos that compose stories, in which each story (S) provides an overview about the collection (Figure 55). So, we extract stories from a collection, $C \rightarrow S$, where $C \subset S$.


The Egyptian Revolution S
storify.com/yasmina_anwar/boston-marathon-bombing-story-from-archi
Apps
LaTeX Line and Pag...
Save to Mendeley
WordCounter
إذاعة القرآن الكريم من ...
إذاعة القرآن الكريم من ...
Storify by livefyre
Search Storify
Create Story
Login
Feedback

The Egyptian Revolution Story from Archive-It: Nearly the Same Time

Different URIs at nearly samee Memento-Datetime (Feb. 1, 2011)

Cracking Egypt internet crackdown #Dial25
Wired WIKI: Scenario: Your government is displeased with the communication going on in your location and pulls the plug on your internet ...
ARCHIVE-IT

Egypt says it's shutting down mobile phone networks
Read full coverage of the unrest in Egypt updated continually by CNN reporters worldwide. Send your photos and video to iReport and see C...
ARCHIVE-IT

Egyptian Chronicles
The Tunisian revolution can teach us a lot of things if we want to learn something from other people's experience : Revolution happens un...
ARCHIVE-IT

Blogging Egypt
A sense of hope prevails in the streets of Cairo. Hope mixed with fear of how hard the government could crack down and uncertainty over w...
ARCHIVE-IT

The Egypt Protests
The Egyptian protests against lack of work, rising food prices and the oppressive regime of President Hosni Mubarak seem to be moving tow...

Like 64k
Follow @storify 127K followers
Report Abuse VIP Guided tour About Help Blog Jobs Tools API Terms Privacy © 2014 Storify

(a) Different URIs, same time

The Egyptian Revolution S
storify.com/yasmina_anwar/the-egyptian-revolution-story-from-archive-i
Apps
LaTeX Line and Pag...
Save to Mendeley
WordCounter
إذاعة القرآن الكريم من ...
إذاعة القرآن الكريم من ...
Storify by livefyre
Search Storify
Create Story
Login
Feedback

The Egyptian Revolution Story from Archive-It: Heterogeneous

Different URIs in broad time...

Egyptians brace for Friday protests as internet, messaging disrupted
Police came to al-Aryan's Cairo home at 2.30 a.m. local time, his son-in-law said. According to multiple web services that check whether ...
ARCHIVE-IT

Egypt says it's shutting down mobile phone networks
Read full coverage of the unrest in Egypt updated continually by CNN reporters worldwide. Send your photos and video to iReport and see C...
ARCHIVE-IT

Egypt crisis: 3 dead, 639 injured in Cairo, government officials say
Read full coverage of the unrest in Egypt updated continually by CNN reporters worldwide. Send your photos and video to iReport and see C...
ARCHIVE-IT

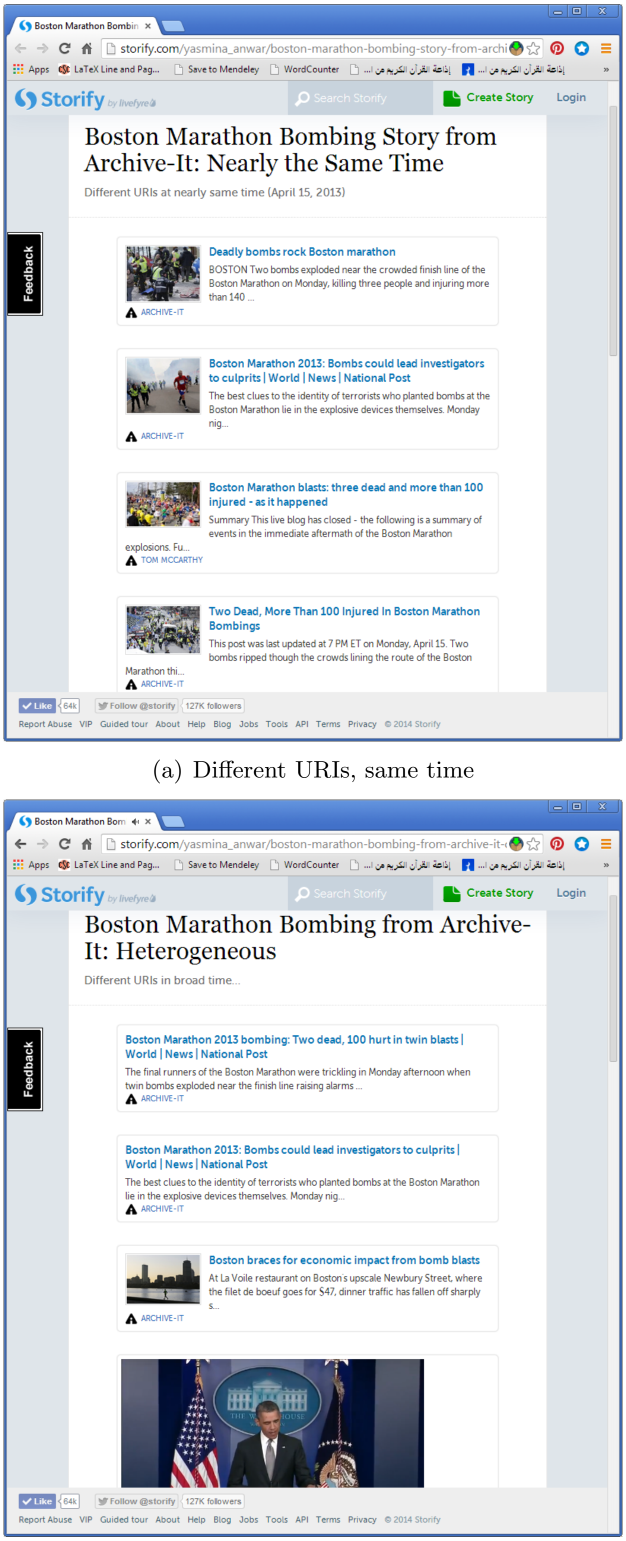
Egyptian Chronicles
We know Wael Ghonim as Google's head of marketing in MENA , we know him as internet activist and we know that he has been missing since F...
ARCHIVE-IT

'Day of departure' rally in Cairo
Huge crowds have occupied the main square in the Egyptian capital, Cairo, to reinforce their demand for the resignation of President Hosn...
ARCHIVE-IT

Like 64k
Follow @storify 127K followers
Report Abuse VIP Guided tour About Help Blog Jobs Tools API Terms Privacy © 2014 Storify

(b) Different URIs, different times

FIG. 53: Different kinds of stories created manually by selecting URIs from "Egypt Revolution and Politics" collection



Boston Marathon Bombing Story from Archive-It: Nearly the Same Time

Different URIs at nearly same time (April 15, 2013)

Deadly bombs rock Boston marathon

BOSTON Two bombs exploded near the crowded finish line of the Boston Marathon on Monday, killing three people and injuring more than 140 ...

ARCHIVE-IT

Boston Marathon 2013: Bombs could lead investigators to culprits | World | News | National Post

The best clues to the identity of terrorists who planted bombs at the Boston Marathon lie in the explosive devices themselves. Monday nig...

ARCHIVE-IT

Boston Marathon blasts: three dead and more than 100 injured - as it happened

Summary This live blog has closed - the following is a summary of events in the immediate aftermath of the Boston Marathon explosions. Fu...

TOM MCCARTHY

Two Dead, More Than 100 Injured In Boston Marathon Bombings

This post was last updated at 7 PM ET on Monday, April 15. Two bombs ripped though the crowds lining the route of the Boston Marathon thi...

ARCHIVE-IT

Like 64k
Follow @storify 127K followers
Report Abuse VIP Guided tour About Help Blog Jobs Tools API Terms Privacy © 2014 Storify

(a) Different URIs, same time

Boston Marathon Bombing from Archive-It: Heterogeneous

Different URIs in broad time...

Boston Marathon 2013 bombing: Two dead, 100 hurt in twin blasts | World | News | National Post

The final runners of the Boston Marathon were trickling in Monday afternoon when twin bombs exploded near the finish line raising alarms ...

ARCHIVE-IT

Boston Marathon 2013: Bombs could lead investigators to culprits | World | News | National Post

The best clues to the identity of terrorists who planted bombs at the Boston Marathon lie in the explosive devices themselves. Monday nig...

ARCHIVE-IT

Boston braces for economic impact from bomb blasts

At La Voile restaurant on Boston's upscale Newbury Street, where the filet de boeuf goes for $47, dinner traffic has fallen off sharply s...

ARCHIVE-IT

Like 64k
Follow @storify 127K followers
Report Abuse VIP Guided tour About Help Blog Jobs Tools API Terms Privacy © 2014 Storify

(b) Different URIs, different times

FIG. 54: Different kinds of stories created manually by selecting URIs from “2013 Boston Marathon Bombing” collection

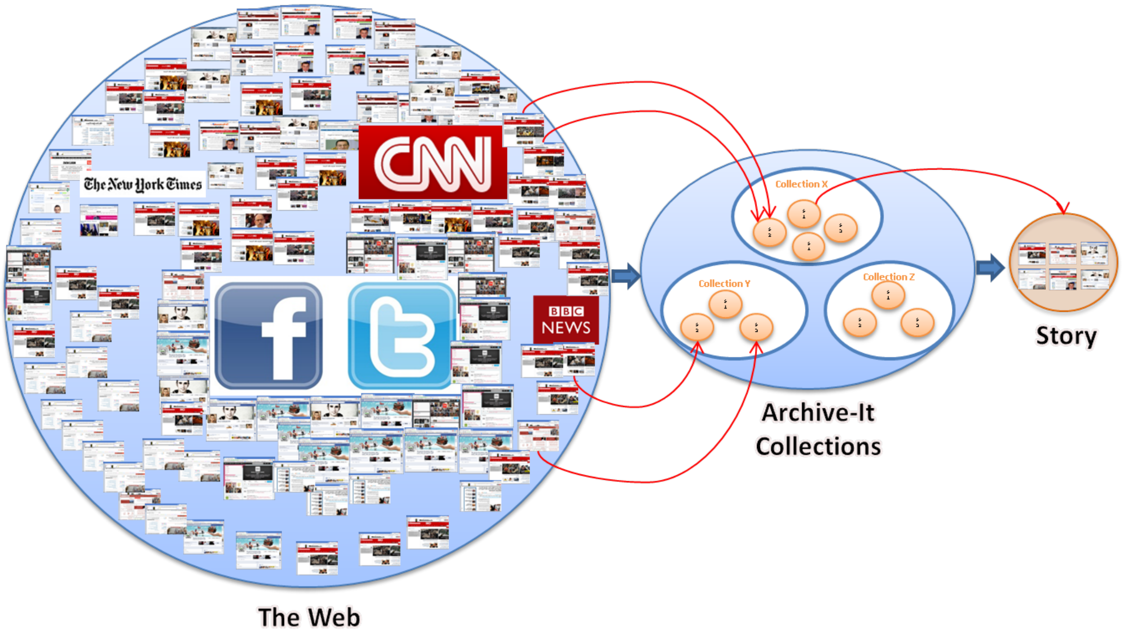
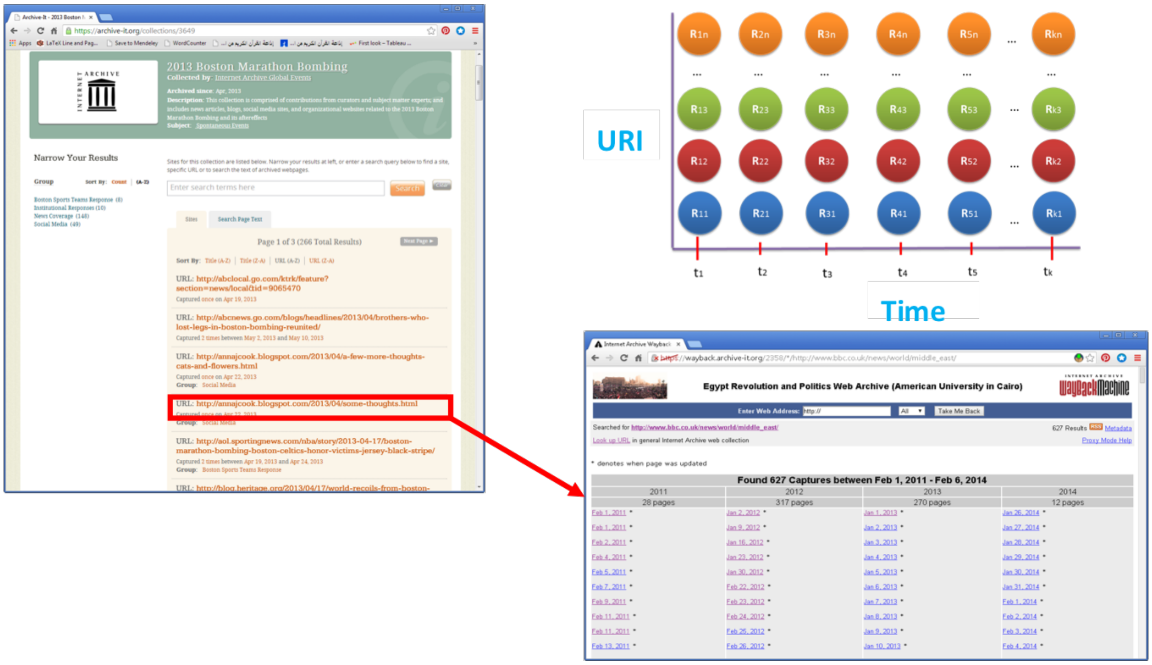


FIG. 55: Collections in Archive-It can be thought of as thematic samples from the live Web. In the DSA framework, we sample k mementos from the pages of the collection to create a summary story.



2013 Boston Marathon Bombing

Collected by: Internet Archive Global Events

Archived since: Apr, 2013

Description: This collection is comprised of contributions from curators and subject matter experts; and includes news articles, blogs, social media sites, and organizational websites related to the 2013 Boston Marathon Bombing and its aftereffects

Subject: Spontaneous Events

Narrow Your Results

Group Sort By: Count | (A-Z)

Boston Sports Teams Response (8)
Institutional Responses (10)
News Coverage (148)
Social Media (49)

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here

Page 1 of 3 (266 Total Results)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

URL: http://abclocal.go.com/ktrk/feature?section=news/local&id=9065470
Captured once on Apr 19, 2013

URL: http://abcnews.go.com/blogs/headlines/2013/04/brothers-who-lost-legs-in-boston-bombing-reunited/
Captured 2 times between May 2, 2013 and May 10, 2013

URL: http://annajcook.blogspot.com/2013/04/a-few-more-thoughts-cats-and-flowers.html
Captured once on Apr 22, 2013
Group: Social Media

URL: http://annajcook.blogspot.com/2013/04/some-thoughts.html
Captured once on Apr 22, 2013
Group: Social Media

URL: http://aol.sportingnews.com/nba/story/2013-04-17/boston-marathon-bombing-boston-celtics-honor-victims-jersey-black-stripe/
Captured 2 times between Apr 19, 2013 and Apr 24, 2013
Group: Boston Sports Teams Response

URL: http://blog.heritage.org/2013/04/17/world-recoils-from-boston-

Egypt Revolution and Politics Web Archive (American University in Cairo)

Enter Web Address: http://

Searched for http://www.bbc.co.uk/news/world/middle_east/ 627 Results

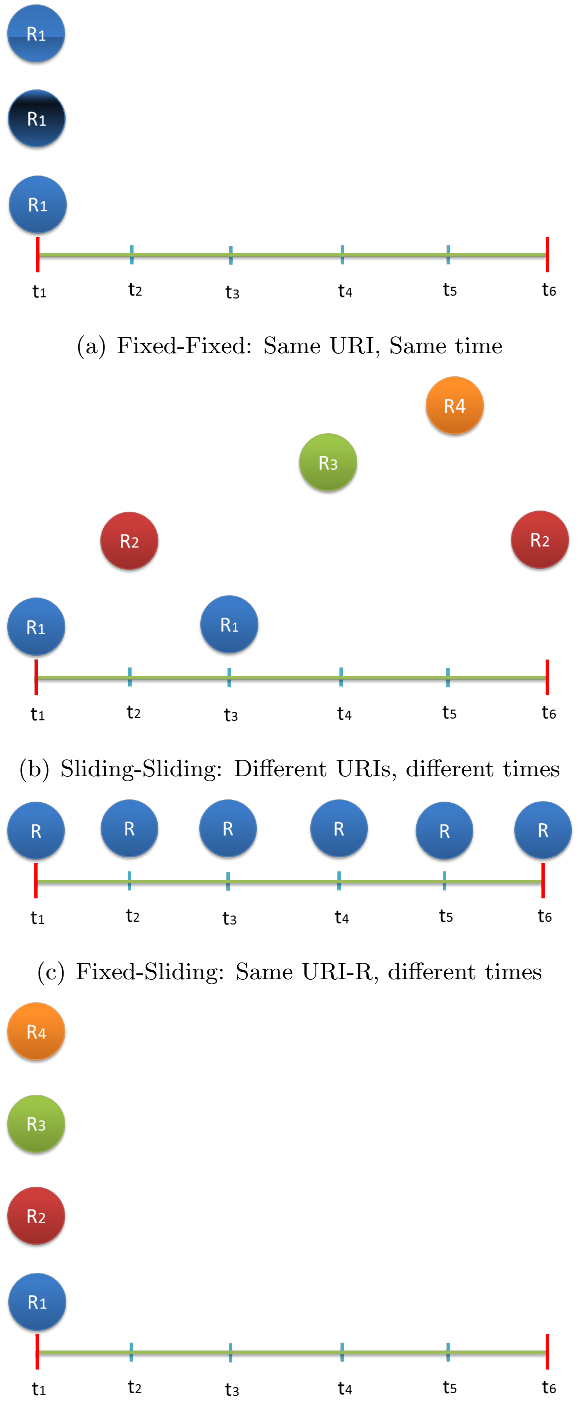
Look up URL in general Internet Archive web collection

* denotes when page was updated

Found 627 Captures between Feb 1, 2011 - Feb 6, 2014

2011	2012	2013	2014
28 pages	317 pages	270 pages	12 pages
Feb 1, 2011 *	Jan 2, 2012 *	Jan 1, 2013 *	Jan 26, 2014 *
Feb 1, 2011 *	Jan 9, 2012 *	Jan 2, 2013 *	Jan 27, 2014 *
Feb 2, 2011 *	Jan 16, 2012 *	Jan 3, 2013 *	Jan 28, 2014 *
Feb 4, 2011 *	Jan 23, 2012 *	Jan 4, 2013 *	Jan 29, 2014 *
Feb 5, 2011 *	Jan 30, 2012 *	Jan 5, 2013 *	Jan 30, 2014 *
Feb 7, 2011 *	Feb 22, 2012 *	Jan 6, 2013 *	Jan 31, 2014 *
Feb 9, 2011 *	Feb 23, 2012 *	Jan 7, 2013 *	Feb 1, 2014 *
Feb 11, 2011 *	Feb 24, 2012 *	Jan 8, 2013 *	Feb 2, 2014 *
Feb 11, 2011 *	Feb 25, 2012 *	Jan 9, 2013 *	Feb 3, 2014 *
Feb 13, 2011 *	Feb 26, 2012 *	Jan 10, 2013 *	Feb 4, 2014 *

FIG. 56: The archived collection has two dimensions: URI and time



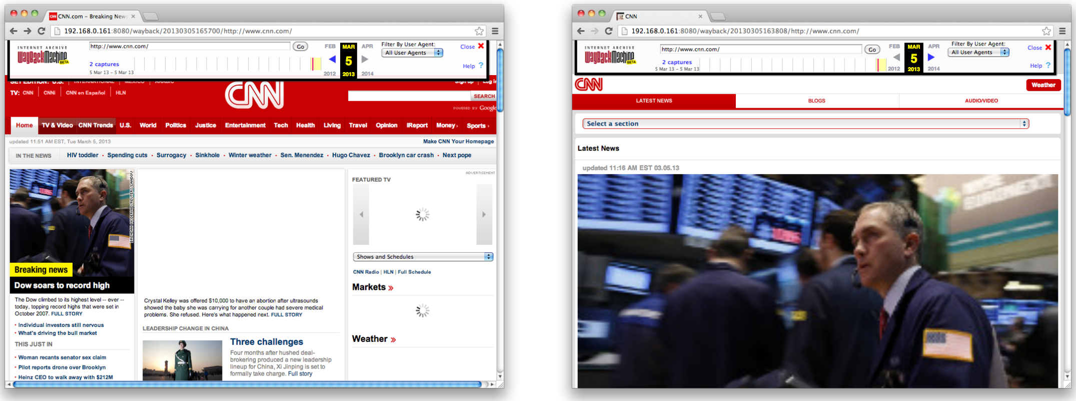
(a) Fixed-Fixed: Same URI, Same time

(b) Sliding-Sliding: Different URIs, different times

(c) Fixed-Sliding: Same URI-R, different times

(d) Sliding-Fixed: Different URIs, same time.

FIG. 57: There are different models for the story that can be created from the collection. The color maps to the unique URI-R.



192.168.0.161:8080/wayback/20130305165700/http://www.cnn.com/

The Dow climbed to its highest level -- ever -- today, topping record highs that were set in October 2007. FULL STORY

Crystal Kelley was offered $10,000 to have an abortion after ultrasounds showed the baby she was carrying for another couple had severe medical problems. She refused. Here's what happened next. FULL STORY

Four months after hushed deal-brokering produced a new leadership lineup for China, Xi Jinping is set to formally take charge. Full story

192.168.0.161:8080/wayback/20130305163808/http://www.cnn.com/

(a) The cnn.com memento when crawled with a desktop Mozilla user-agent accessed from a Mac.

(b) The cnn.com memento when crawled with an iPhone mozilla user-agent accessed from a Mac.

FIG. 58: Mementos differ based on the parameters influencing the representations at crawl/capture time and the devices used to access the mementos [168].

5.3 TYPES OF STORIES GENERATED FROM ARCHIVED COLLECTIONS

An archived collection has two dimensions. As we mentioned before, the collection is composed of a set of seed URIs and each seed has many copies through time (Figure 56). There may be multiple stories that convey different perspectives of the collection, such as the examples of Figures 12 - 20 of Chapter 1. We list four possible kinds of stories in Table 4. We name each story according to the change that happens to the URI and time. It is also possible that there are additional types of stories beyond those in Table 4, and we plan to investigate this in future work.

TABLE 4: Four basic story types (others may be possible).

		Time:	
		fixed	sliding
URIs:	fixed	differences in GeoIP, mobile, etc.	evolution of a single page (or domain) through time
	sliding	different perspectives at a point in time	broadest possible coverage of a collection

We present the definition for each story below, along with a model in Figure 57. The different colors in Figure 57 map to different URI-Rs. We use Memento terminology (URI-T, URI-M, and URI-R) in the definitions.

5.3.1 FIXED PAGE, FIXED TIME

Fixed Page, Fixed Time (FPFT) is defined as a different representation for the same Web site because of GeoIP, mobile, and other environmental factors (e.g., Figure 58) [168]. It is generated using the same URI at a specific point of time with differences in the representation. The model for this story is shown in Figure 57(a).

$$\begin{aligned}\text{Fixed Page, Fixed Time} &= (URI\text{-}M_i, URI\text{-}M_i, ..., URI\text{-}M_i),\ \text{where} \\ &\quad URI\text{-}M_i = URI\text{-}R@t_i \qquad (6)\end{aligned}$$

5.3.2 SLIDING PAGE, SLIDING TIME

Sliding Page, Sliding Time (SPST) is defined as the broadest possible coverage of a collection. It is generated using different URIs at different times.

$$\begin{aligned}\text{Sliding Page, Sliding Time} &= (URI\text{-}M_1, URI\text{-}M_2, ..., URI\text{-}M_k),\ \text{where} \\ &\quad URI\text{-}M_i = URI\text{-}R@t_i \text{ and } t_i \neq t_j \qquad (7)\end{aligned}$$

5.3.3 FIXED PAGE, SLIDING TIME

Fixed Page, Sliding Time (FPST) is defined as the evolution of a single page (or domain) through time Figure (57(d)). The possible scenario of this story is when a user wants to see how the story evolved over time from a specific Web site, e.g., `cnn.com`.

$$\begin{aligned}\text{Fixed Page, Sliding Time} &= (URI\text{-}M_1, URI\text{-}M_2, ..., URI\text{-}M_k),\ \text{where} \\ &\quad URI\text{-}R(URI\text{-}M_i) = URI\text{-}R(URI\text{-}M_j) \text{ and} \\ &\quad URI\text{-}M_i = URI\text{-}R@t_i \qquad (8)\end{aligned}$$

5.3.4 SLIDING PAGE, FIXED TIME

Sliding Page, Fixed Time (SPFT) is defined as different perspectives at a point in time. It is generated using different URI-Rs at nearly the same datetime.

$$\begin{aligned}\text{Sliding Page, Fixed Time} &= (URI\text{-}M_1, URI\text{-}M_2, ..., URI\text{-}M_k),\ \text{where}\\ & URI\text{-}R(URI\text{-}M_i) \neq URI\text{-}R(URI\text{-}M_j)\ \text{and}\\ & URI\text{-}M_i = URI\text{-}R@t_i \end{aligned} \tag{9}$$

Note that the Fixed-Fixed story can not be supported by the current capabilities of Web archives [168]. While Heritrix provides archivists the ability to modify the user-agent string to crawl different representations, such as mobile Web, archives currently do not provide users the ability to navigate representations by their environmental influences. Kelly et al. [168] proposed a method for identifying personalized representations in Web archives through a modified Wayback Machine to add environmental dimensions to browsing the past Web.

5.4 THE DARK AND STORMY ARCHIVES (DSA) FRAMEWORK

In this section, we describe the general methodology for addressing the research questions and constructing k archived pages that represent an extracted story from an Archive-It collection, arranging them in a narrative structure ordered by time (or any other type of story), then pulling them into existing storytelling tools or other visualizations, such as the examples in Figures 53 and 54.

The DSA framework can be divided into three main components (Figure 59):

1. Establish a baseline for the structure of human-generated stories (focusing on the popular ones with the most views) and the makeup of archived collections.

    - Determine the characteristics of the user-curated stories based on a user study of stories from Storify (Chapter 6).
    - Determine the characteristics of Archive-It collections by measuring the statistics of the collections such as the number of URIs, the number of mementos, the most used domains, etc. (Chapter 7).
    - Compare the created descriptive models of the created stories on social media and the collections in Archive-It (Chapter 7).

2. Reduce the candidate pool of archived pages.

    - Exclude the off-topic pages from the collection (Chapter 8).
        - Model TimeMap behavior in Web archives based on how the page's aboutness changes through time.

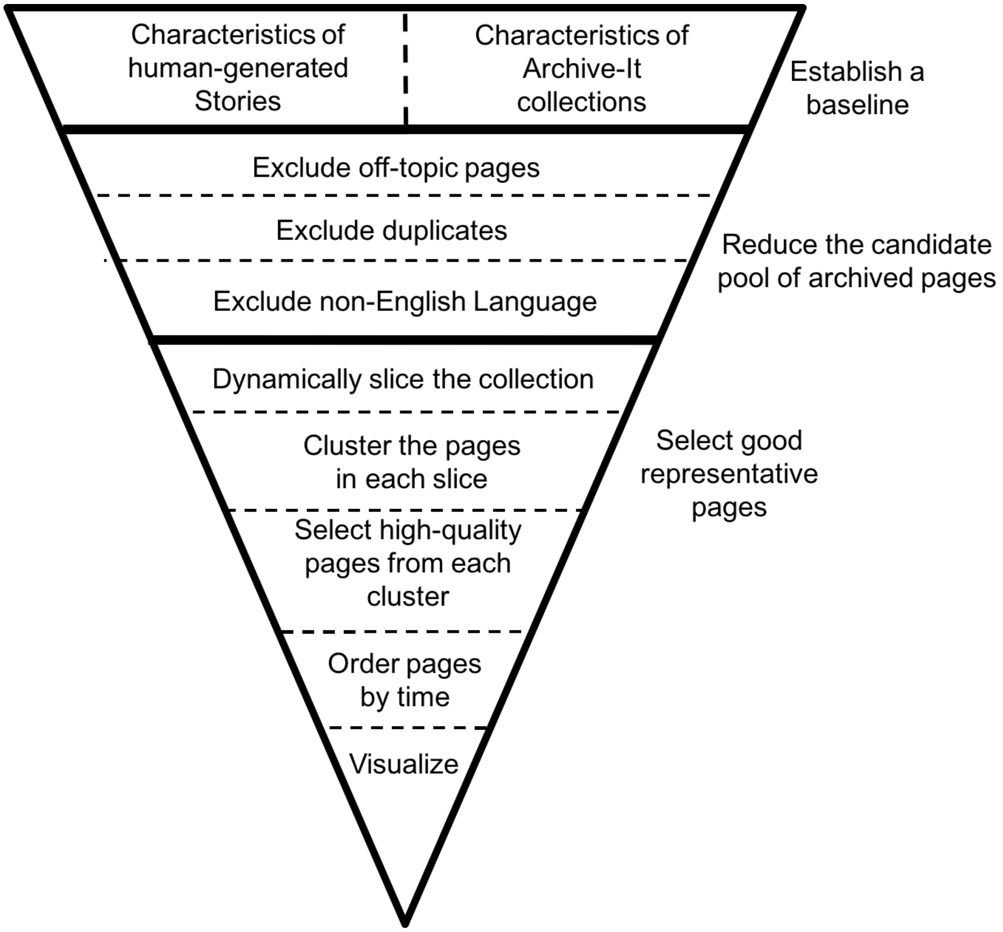

Characteristics of human-generated Stories
Characteristics of Archive-It collections
Establish a baseline
Exclude off-topic pages
Exclude duplicates
Exclude non-English Language
Reduce the candidate pool of archived pages
Dynamically slice the collection
Cluster the pages in each slice
Select high-quality pages from each cluster
Order pages by time
Visualize
Select good representative pages

FIG. 59: The main components of the Dark and Stormy Archives (DSA) framework.

    - Investigate different methods for determining when the page goes off-topic in individual TimeMaps.
    - Based on the best performing method, eliminate the off-topic pages.
  - Exclude the (near-)duplicate mementos of each TimeMap (Chapter 9).
  - Exclude the non-English language mementos (Chapter 9).

3. Select good representative pages for each story (Chapter 9).

   - Slice the collection dynamically.
   - Cluster the pages in each time slice.
   - Evaluate and select the best representative page from each cluster based on multiple quality metrics.
   - Identify the different notions of time for each page.

- Put the selected pages in chronological order.
- Extract the metadata of the selected pages.
- Visualize the pages by leveraging storytelling tools, such as Storify.

5.5 SUMMARY

The output of our work is a framework that automatically creates stories out of Archive-It collections. Our goal is to provide users with a tool that allows them to get many perspectives about the collection and also how the story of the collection has evolved over time. We leverage narrative visualizations and storytelling tools, such as Storify, to visualize the created stories and demonstrate how they have evolved over time. Furthermore, we provide collection curators with tools that allow the detection of the off-topic Web pages in the collection, as specified in Chapter 8.

In this chapter, we provided a conceptual model for the framework of automatically generating stories out of archived collections along with the definitions of the types of stories that can be generated. The following chapters handle each step of the framework starting from establishing a baseline (Chapters 6 and 7) until applying the characteristics of human-generated stories on the stories and selecting the best representative pages to leverage them with Storify (Chapter 9). We will evaluate the automatically generated stories in Chapter 9.

CHAPTER 6

CHARACTERISTICS OF SOCIAL MEDIA STORIES

Since the stories in Storify are created by humans, we model the structural characteristics of these stories, with particular emphasis on "popular" stories (i.e., the top 25% of views, normalized by time available on the Web) [19, 21]. In this chapter, we answer the following questions:

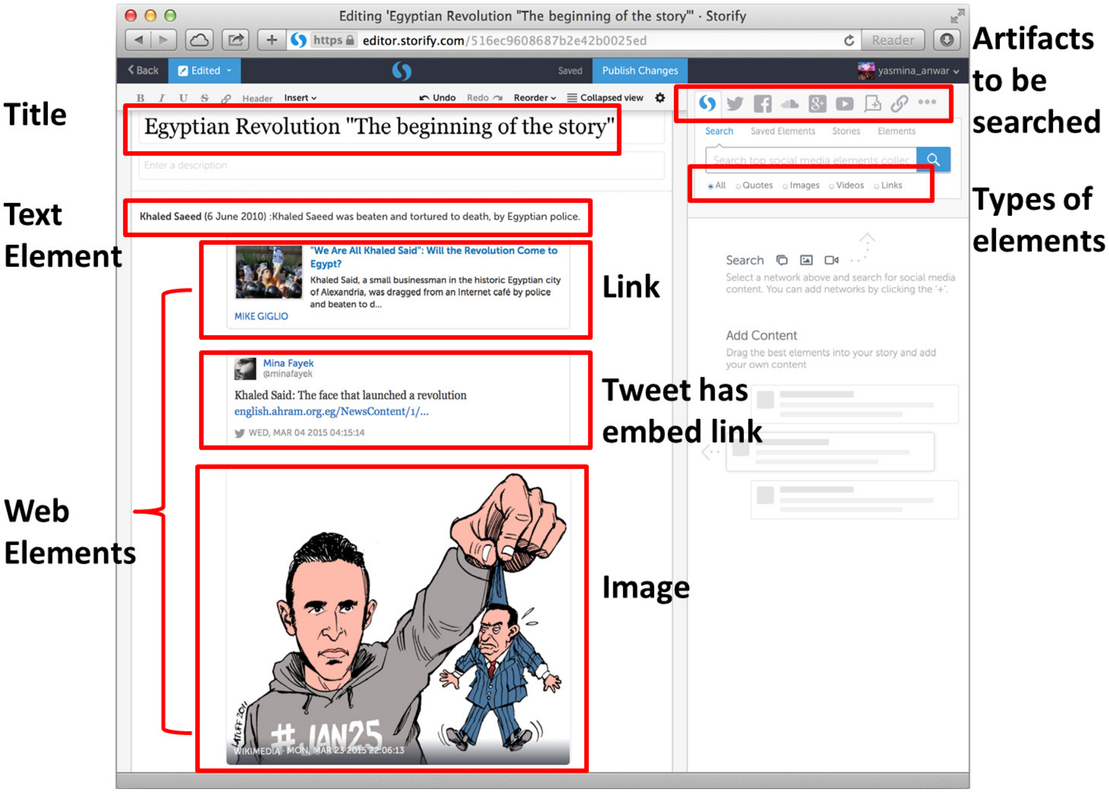
- What is the length of the human-generated stories?
- What are the types of resources used in these stories?
- What are the most frequently used domains in the stories?
- What is the editing time of the stories?
- Is there a relationship between the timespan and the features of the story?
- Is there a relationship between the popularity of the stories and the number of elements?
- What differentiates the popular stories?
- How many of the resources in these stories disappear every year?
- Can we find these missing resources in the archives?

To answer these questions, we investigated 14,568 stories from Storify, comprising 1,251,160 individual resources.

6.1 CONSTRUCTING THE DATASET

As we mentioned earlier, Storify provides a graphical interface for selecting URIs of Web resources and arranging the resulting snippets and previews (see Figure 60), with a special emphasis on social media (e.g., Twitter, Facebook, YouTube, Instagram). We name these previews of Web resources "Web elements", and the annotations Storify allows on these previews we name "text elements". To investigate the characteristics of human created stories, we created the dataset by querying the Storify Search API[1] with the most

[1] `http://dev.storify.com/api/`


Title
Text Element
Web Elements
Link
Tweet has embed link
Image
Artifacts to be searched
Types of elements
Editing 'Egyptian Revolution "The beginning of the story"' · Storify
https editor.storify.com/516ec9608687b2e42b0025ed
Reader
Back
Edited
Saved
Publish Changes
yasmina_anwar
Header
Insert
Undo
Redo
Reorder
Collapsed view
Search
Saved Elements
Stories
Elements
Search top social media elements collec
All
Quotes
Images
Videos
Links
Search
Select a network above and search for social media content. You can add networks by clicking the '+'.
Add Content
Drag the best elements into your story and add your own content
Enter a description
#JAN25
WIKIMEDIA · MON, MAR 23 2015 22:06:13

Egyptian Revolution "The beginning of the story"

Khaled Saeed (6 June 2010) :Khaled Saeed was beaten and tortured to death, by Egyptian police.

"We Are All Khaled Said": Will the Revolution Come to Egypt?
Khaled Said, a small businessman in the historic Egyptian city of Alexandria, was dragged from an Internet café by police and beaten to d...
MIKE GIGLIO

Mina Fayek
@minafayek
Khaled Said: The face that launched a revolution
english.ahram.org.eg/NewsContent/1/...
WED, MAR 04 2015 04:15:14

FIG. 60: An example of creating a story on Storify shows the Storify-defined categories for resources of the stories.

frequent 1000 English keywords issued to Yahoo[2]. This set of available search keywords allowed us to gather sets of stories about many different topics. This was especially useful since we do not know the ranking algorithm used by Storify search.

We retrieved 400 results for each keyword, resulting in a total of 145,682 stories downloaded in the JavaScript Object Notation (JSON) format [48]. We created the dataset in February 2015 and only considered stories authored in 2014 or earlier, resulting in 37,486 stories. We eliminated stories with zero or one elements or zero views, resulting in 14,568 unique stories authored by 10,199 unique users and containing a total of 1,251,160 Web and text elements.

6.2 GENERAL CHARACTERISTICS OF HUMAN-GENERATED STORIES

Figure 61 contains the distribution of the number of views of the stories, the number of Web elements, the number of text elements, and the number of subscribers. We notice

[2] `http://webscope.sandbox.yahoo.com/catalog.php?datatype=l`

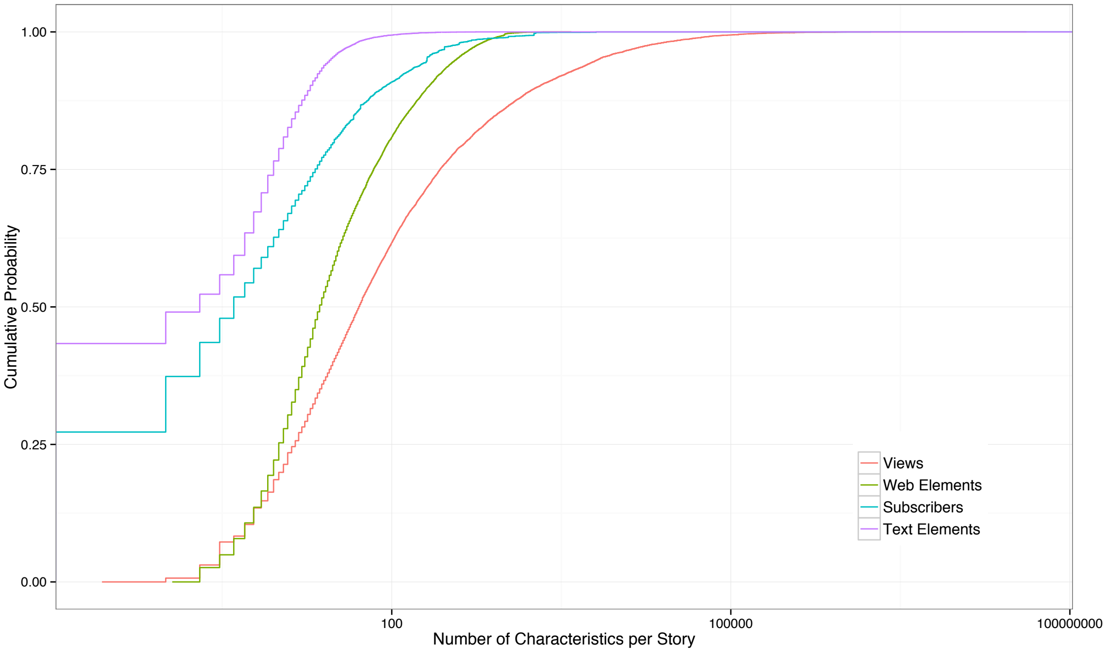


FIG. 61: Distribution of the characteristics of the 14,568 Storify stories analyzed.

that around 48% of the stories do not have any text elements. This indicates that only about half of the stories are annotated with descriptive text.

For a closer look at the features of the stories, we present the distribution percentiles along with means of story views, Web and text elements, and number of subscribers for the story authors in Table 5. We show the distribution percentiles along with means because the distribution of the data is long-tailed. The editing time is the time interval (in hours) in which users edit their stories and is calculated by taking the difference between the story creation-date and last-modified date. The median for all stories is 23 Web elements and 1 text element, and 44% of the stories have no text elements at all. Due to the large range of values, we believe median is a better indicator of typical values.

6.2.1 WHAT KIND OF RESOURCES ARE IN STORIES?

Using the Storify-defined categories reflected in the Storify user interface (Figure 60), the 1,251,160 elements consist of 70.8% links, 18.4% images, 8.1% text, 2.0% videos, and 0.7% quotes. Text elements are relatively rare, meaning that few users choose to annotate the Web elements in their story.

TABLE 5: Distribution of features of the stories in the dataset. Editing time is measured in hours.

Features	**Views**	**Web elements**	**Text elements**	**Subscribers**	**Editing Time**
25th percentile	14	10	0	0	0.18
50th percentile	51	23	1	4	3
75th percentile	268	69	9	21	120
90th percentile	1949	210	19	85	1,747
Maximum	11,284,896	2,216	559	1,726,143	36,111
Mean	3,790	80	8	286	855
Std. Dev.	99,226	158	18	20,220	2,982

6.2.2 WHAT DOMAINS ARE USED IN STORIES?

The Web elements in Storify stories represent 91.95% (1,150,399 out of 1,251,160) of all the resources. To analyze the distribution of domains in stories, we canonicalized the domains (e.g., `www.cnn.com` → `cnn.com`) and dereferenced all shortened URIs (e.g., `t.co`, `bit.ly`) to the URIs of the final locations. This resulted in 25,947 unique domains in the 14,568 unique stories.

Figure 62 shows the relationship between the frequency of the domains and the number of stories they appeared in. For example, the rightmost dot at the top of the graph represents the most frequent domain in the stories (`twitter.com`), which also appeared in the largest number of stories. This domain appears almost 1,000,000 times in over 10,000 different stories. We conclude from the graph that the most frequent domains are often used in the majority of stories.

Table 6 contains the top 25 domains of the resources ordered by their frequency. The list of top 25 domains represents 92.3% of all resources. The table also contains the global rank of the domains according to Alexa[3] as of March 2015. We see from the table that Web elements from `twitter.com` appeared 943,859 times in 10,914 stories, comprising over 82% of all Web elements. Note that `plus.google.com` has rank one because Alexa does not differentiate `plus.google.com` from `google.com`. We manually categorized these domains in a more fine-grained manner than Storify provides with its "links, images, text, videos, quotes" descriptions.

[3] `http://www.alexa.com/`

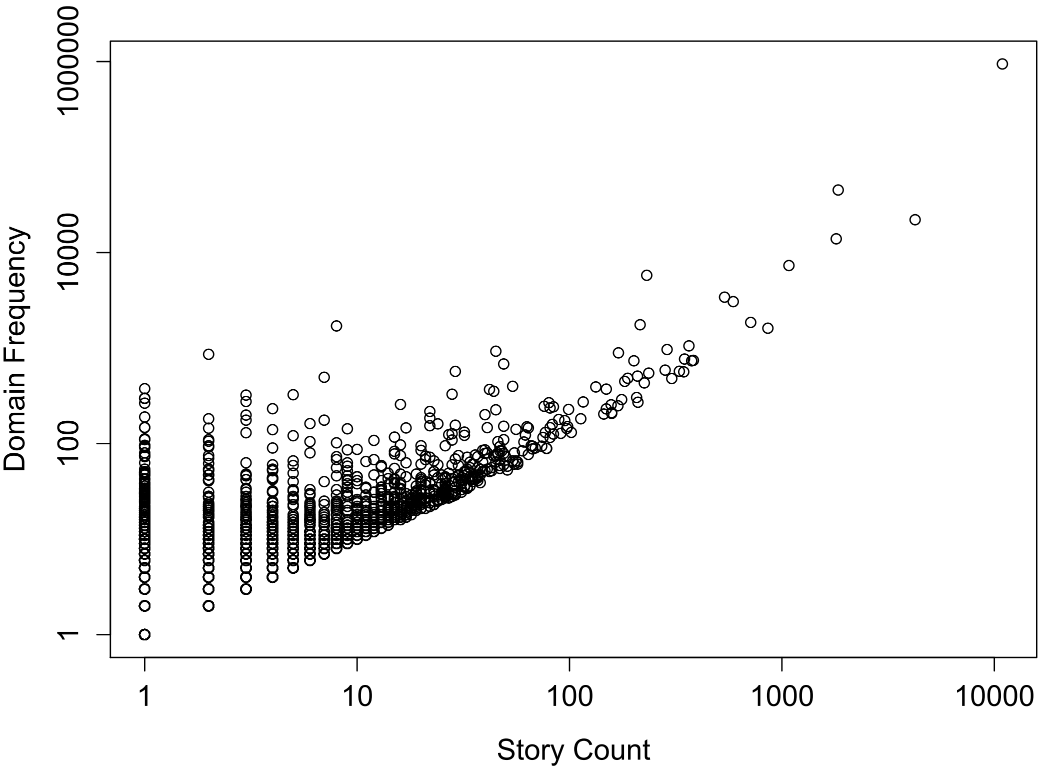


FIG. 62: The relationship between the frequency of the domains in Storify stories and the number of stories in which those domains appear.

Although the top 25 list of domains appearing in the stories is dominated by globally popular Web sites (e.g., Twitter, Instagram, YouTube, Facebook), the long-tailed distribution results in the presence of many globally lesser known sites. In Section 6.2.3, we investigate the correlation between Alexa global rank and rank within Storify.

We also presented the list of top domains based on the count of stories in which they were used (Table 7). We notice that the two lists are similar. We also can see from Table 7 that `storify.com` appeared in the highly ranked domains across the stories, which means many stories refer to other stories in Storify.

The Embedded Resources of twitter.com

Since Twitter is the most popular domain (> 82% of Web elements), we investigate if the tweets have embedded resources of their own. For example, Figure 63 shows a tweet in a Storify story that contains an image from Twitter. Furthermore, other tweets may

TABLE 6: The top 25 domains based on the frequency of appearance in Storify stories. Alexa global rank was retrieved in 2015-03.

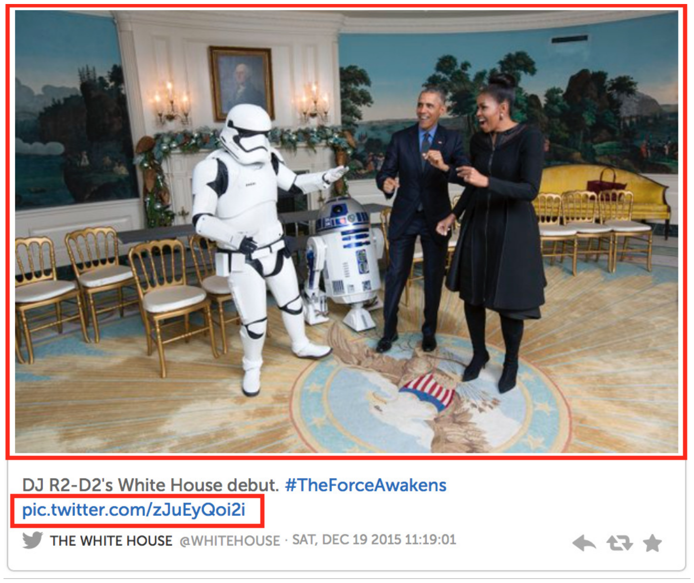
Domain	**Frequency**	**Percentage of Domains**	**Story Count**	**Alexa Global Rank**	**Category**
twitter.com	943,859	82.05%	10,914	8	Social media
instagram.com	45,188	3.93%	1,841	25	Photos
youtube.com	22,076	1.92%	4,238	3	Videos
facebook.com	13,930	1.21%	1,802	2	Social media
flickr.com	7,317	0.64%	1,079	126	Photos
patch.com	5,783	0.50%	231	2,096	News
plus.google.com	3,413	0.30%	537	1	Social media
tumblr.com	3,066	0.27%	590	31	Blogs
blogspot.com	1,857	0.16%	713	18	Blogs
imgur.com	1,756	0.15%	215	36	Photos
coolpile.com	1,706	0.15%	8	149,281	Entertainment
wordpress.com	1,615	0.14%	859	33	Blogs
giphy.com	1,055	0.09%	365	1,604	Photos
bbc.com	966	0.08%	288	156	News
lastampa.it	927	0.08%	45	2,440	News
pinterest.com	892	0.08%	170	32	Photos
softandapps.info	861	0.07%	2	160,980	News
photobucket.com	768	0.07%	348	341	Photos
nytimes.com	744	0.06%	383	97	News
soundcloud.com	736	0.06%	201	167	Audio
wikipedia.org	736	0.06%	376	7	Encyclopedia
repubblica.it	682	0.06%	49	439	News
theguardian.com	588	0.05%	282	157	News
huffingtonpost.com	572	0.05%	329	93	News
punto-informatico.it	570	0.05%	29	42,955	News

TABLE 7: The top 25 domains based on the number of stories they appear in (Story Count). The percentage of stories is out of 14,568. Alexa global rank was retrieved in 2015-03.

Domain	Story Count	Percentage of Stories	Frequency	Alexa Global Rank	Category
twitter.com	10,914	74.92%	943,859	8	Social media
youtube.com	4,238	29.09%	22,076	3	Videos
instagram.com	1,841	12.64%	45,188	25	Photos
facebook.com	1,802	12.37%	13,930	2	Social media
flickr.com	1,079	7.41%	7,317	126	Photos
wordpress.com	859	5.90%	1,615	33	Blogs
blogspot.com	713	4.89%	1,857	18	Blogs
tumblr.com	590	4.05%	3,066	31	Blogs
plus.google.com	537	3.69%	3,413	1	Social media
nytimes.com	383	2.63%	744	97	News
wikipedia.org	376	2.58%	736	7	Encyclopedia
giphy.com	365	2.51%	1,055	1,604	Photos
photobucket.com	348	2.39%	768	341	Photos
upload.wikimedia.org	345	2.37%	564	200	Encyclopedia
huffingtonpost.com	329	2.26%	572	93	News
cnn.com	303	2.08%	480	76	News
bbc.com	288	1.98%	966	156	News
theguardian.com	282	1.94%	588	157	News
google.com	236	1.62%	547	1	Search
patch.com	231	1.59%	5,783	2,096	News
washingtonpost.com	225	1.54%	432	218	News
imgur.com	215	1.48%	1,756	36	Photos
foxnews.com	210	1.44%	271	215	News
storify.com	209	1.43%	509	3,237	Social network
forbes.com	207	1.42%	304	164	News

TABLE 8: The 10 most frequent domains in the embedded resources of the tweets.

Domain	Percentage	Category
twimg.com	46.17%	Images
instagram.com	4.28%	Images
youtube.com	2.82%	Videos
linkis.com	2.04%	Media sharing
facebook.com	1.40%	Social Media
wordpress.com	0.61%	Blogs
vine.co	0.53%	Videos
blogspot.com	0.52%	Blogs
storify.com	0.49%	Social Network
bbc.com	0.44%	News



DJ R2-D2's White House debut. #TheForceAwakens
pic.twitter.com/zJuEyQoi2i
THE WHITE HOUSE @WHITEHOUSE · SAT, DEC 19 2015 11:19:01

FIG. 63: A tweet in Storify has an image as an embedded resource. Note that the text of the tweet includes the URI of the image.

contain links or videos. This captures the behavior of users including tweets in the stories because the tweets are surrogates for embedded content. We randomly sampled 5% of the Twitter resources (47,512 URIs). Of the sampled tweets in the stories, 32% (15,217) have embedded resources, of which there are 14,616 unique URIs. Of the 15,217, 46% are photos from `twitter.com` (hosted at `twimg.com`). Table 8 contains the 10 most frequent domains for the embedded resources, which represent 61.6% of all the URIs embedded in tweets. Again, we see that some Storify stories (0.49%) point to other stories in Storify.

6.2.3 CLASSIFICATION OF THE RESOURCES BASED ON THE TLD

Table 9 presents the distribution of Top Level Domains (TLDs) for the URIs that were used in Storify stories (only the top 10 are shown). The table shows that the most used TLD is .com by far. Note that .cat is the TLD for a Catalan site (`http://www.aragirona.cat/`). The top 10 list represents 98.92% of all resources in Storify stories.

We calculate the Kendall's Tau correlation (τ_{sf}) between the top n domains in Storify stories based on their frequency (for example, the list of the top 25 domains in Table 6) and their Alexa global rank. We also checked the Kendall's Tau correlation (τ_{sc}) between the top n domains used in the most number of stories (for example, the list of top 25 domains in Table 7) and their Alexa global rank.

TABLE 9: The top 10 TLDs of the resources.

TLD	**Percentage**
.com	96.48%
.org	0.64%
.it	0.52%
.uk	0.34%
.net	0.32%
.de	0.21%
.es	0.11%
.info	0.11%
.fr	0.10%
.cat	0.09%

TABLE 10: The Kendall's Tau correlation between the n most frequent domains in the stories and their Alexa global Rank (τ_{sf}) and between the top n domains that have the most number of stories and Alexa global rank (τ_{sc}).

n	10	15	25	50	100
τ_{sf}	0.1555	**0.4476**	**0.3372**	**0.3194**	**0.2485**
τ_{sc}	0.1556	0.3524	**0.4107**	**0.4260**	**0.4639**

The results are shown in Table 10. Statistically significant ($p \leq 0.05$) correlations are bolded. The highest correlation we found between Alexa global rank and top domains based on frequency was 0.45 for the top 15 domains. The highest correlation between Alexa global rank and top domains based on number of stories was 0.46 for the top 100 domains. From the results, we notice that most of the time the highly ranked real-world resources, such as `twitter.com`, are correspondingly the most used in human-generated stories.

This is interestingly in contrast with Zhong et al. [347], which found that the most frequent sites on Pinterest had low Alexa global ranks. This is possibly due to the different nature of the usage of both sites. In Pinterest, users pin photos or videos of interest to create theme-based image/video collections such as hobbies, fashion, and events. The most used subject areas by Pinterest users are food and drinks, décor and design, and apparel and accessories [118]. Most of the pins on Pinterest come from blogs or are uploaded by users. In Storify, people tend to use social media and Web resources to create their narratives about events or news.

TABLE 11: The percentage of the stories based on the editing interval along with the median of Web elements, text elements, and views. The percentage is out of 15,568 stories.

Intervals	Percentage	Median Web elements	Median text elements	Median views
0-60 seconds	14.0%	15	0	23
1-60 minutes	26.7%	19	0	53
1-24 hours	23.4%	25	5	110
1-7 days	13.5%	26	7	78
1-4 weeks	8.4%	26	9	80
1-12 months	10.9%	38	2	129
1-4 years	3.1%	56	15	156

6.2.4 WHAT IS THE MEAN EDITING TIME FOR STORIES?

Table 11 shows the percentage of the stories with editing times in various time intervals. The table also shows the corresponding features of the stories, divided by their editing time. We normalized the number of views by the age of the story (dataset collection date − story creation date). The first two intervals ($<$ 1 hour) represent stories that were created, modified, and then published with no continuing edits.

We see that the majority of the stories in the dataset were created and edited in the span of one day. There are 14% of stories that have been updated over a long period of time, with the longest editing time in our dataset covering more than four years and with more than 13,000 views. Curiously, this story had only 33 Web elements and 51 total elements. Although the story with the longest editing time did not have the largest number of elements, from Table 11 we can see that based on the median number of elements in each interval there is a nearly linear relationship between the editing time length of the story and the number of elements.

6.2.5 DECAY OF WEB ELEMENTS

In this section, we investigate how many resources in the stories are missing from the live Web and how many are available in public Web archives. We used Memento to check the existence in the archives. We checked the live Web and public Web archives for 265,181 URIs (202,452 URIs from the Web elements of stories + 47,512 randomly sampled tweet URIs + 15,217 URIs of embedded resources in those tweets), in which there are 253,978 unique URIs. Here we further examine the results for the most frequent five domains in the stories: `twitter.com`, `instagram.com`, `youtube.com`, `facebook.com`, and `flickr.com`.

TABLE 12: The existence of the resources on the live Web (on the left) and in the archives (on the right). Available represents the requests which ultimately return HTTP 200, while missing represents the requests that return HTTP 4xx, HTTP 5xx, HTTP 3xx to others except 200, timeouts, and soft 404s. Total is the total unique URIs from each domain.

	Existence on live Web			**Found in archives**		
Resources	Available	Missing	Total	Of the available	Of the missing	Total
Twitter	95.5%	4.5%	47,385	0.9%	3.4%	477
Instagram	86.6%	13.4%	43,396	0.3%	0.07%	103
Youtube	99.3%	0.7%	19,809	16.0%	0.75%	3,140
Facebook	95.2%	4.8%	12,793	0.6%	0.49%	80
Flickr	95.6%	4.4%	6,859	0.4%	0.0%	25
others	82.1%	17.9%	109,120	26.8%	15.5%	27,033
Twitter resources	90.1%	9.9%	14,616	8.0%	14.1%	1,257

Existence on the Live Web

We checked the existence of the 253,978 unique URIs on the live Web. We also checked the pages that give "soft 404s", which return HTTP 200, but do not actually exist [40]. The left two columns of Table 12 contain the results of checking the status of the Web pages on the live Web. Of all the unique URIs, 11.8% are missing on the live Web. The table also contains the results of the five most frequent domains and all other URIs. We also included the results of checking the existence of Twitter embedded resources at the bottom of the table. From the table, we conclude that the decay rate of social media content is lower than the decay rate of the regular Web content and Web sites.

Existence on the Live Web as a Function of Time

We measured the decay of the resources of Storify stories in time by measuring the percentage of the missing resources in the stories over time. For this experiment, we used the 249,964 (all the URIs excluding twitter embedded resources) resources in 14,513 stories to check the rate of the decay in the stories.

We found that 40.8% of the stories contain missing resources with a mean value of 10.3% of the elements missing per story. Figure 64 contains the distribution of the creation date of stories in our dataset in each year and the percentage of the missing resources in each corresponding year. From the graph, we can infer a nearly linear decay rate of resources through time: the resources disappear at rate of 30% the first year, 20% the second year, then the rate decreases steadily the last three years until it reaches 9.6% for the last year. This finding is close to the findings by SalahEldeen and Nelson [276], in which they found

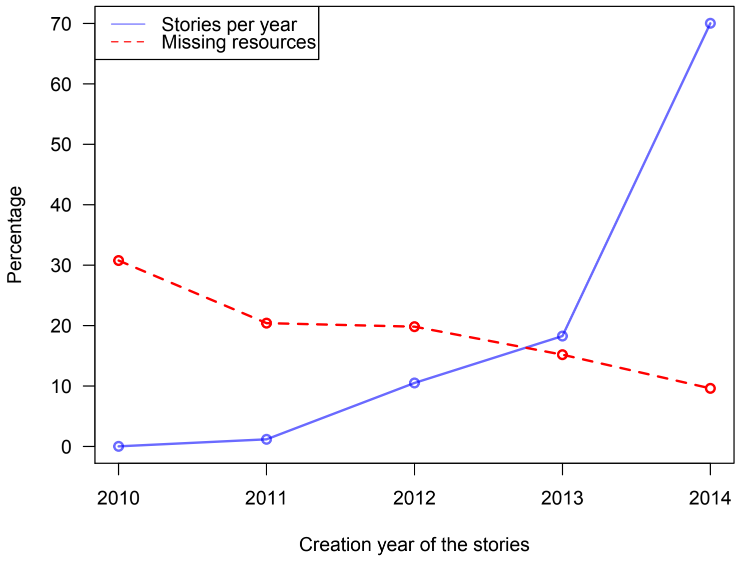


FIG. 64: The distribution of the stories per year and the decay rate of the resources in these stories through time.

that there is a nearly linear relationship between time of sharing the resources and the percentage of resources lost from the live Web, with a rate of 11% the first year and 7% for each following year.

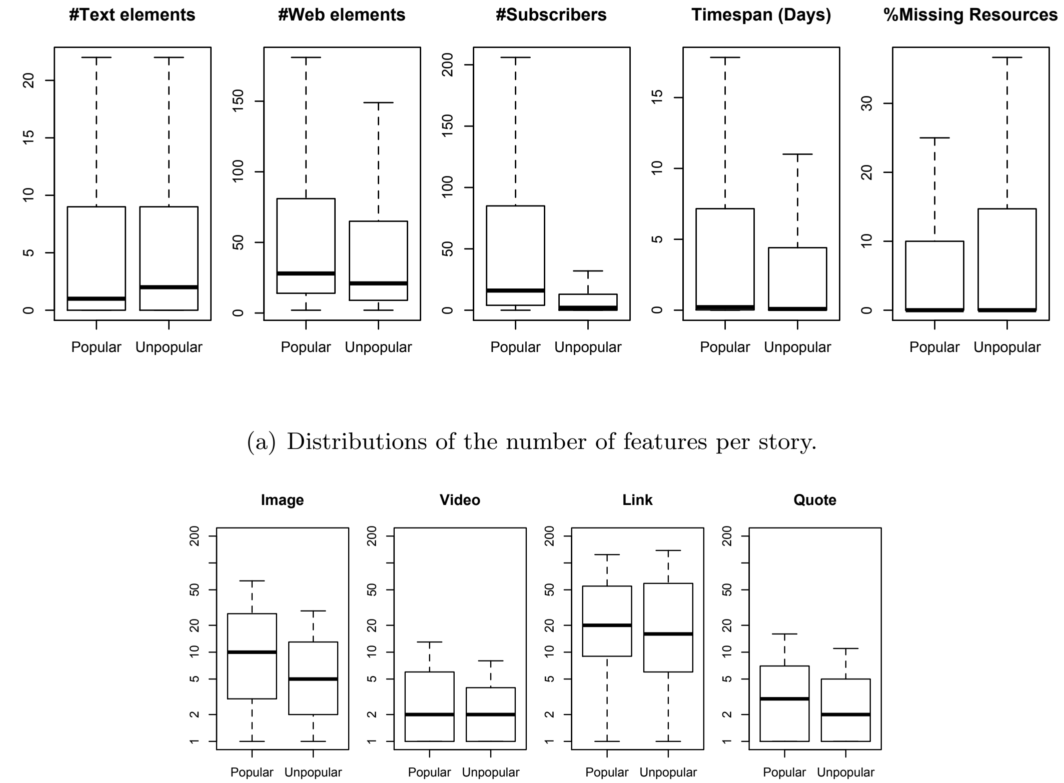
Existence in the Archives

We checked the 253,978 resources for existence in general Web archives in March 2015. The existence in the Web archives was tested by querying a Memento Aggregator[4].

The right-most columns of Table 12 contain the percentage of the URIs found in the Web archives out of the missing and the available URIs on the live Web. In total, 12.6% of the URIs were found in the public Web archives. Of the missing resources (29,964), 11% were found in public Web archives. From Table 12, we notice that social media is not as well-archived as the regular Web. Facebook uses robots.txt to block Web archiving by the Internet Archive[5], but the other sites do not have this restriction.

[4] `http://timetravel.mementoweb.org/guide/api/`, which provided results from 12 different public Web archives.

[5] See: `https://archive.org/about/faqs.php#14`



(a) Distributions of the number of features per story.

(b) Distributions of the number of elements per story.

FIG. 65: Characteristics of popular and unpopular stories.

6.3 WHAT DOES A POPULAR STORY LOOK LIKE?

In this section, we establish structural features for what differentiates popular stories from normal stories for building a baseline for the stories we will automatically create from the archives. We divided the stories into popular and unpopular stories based on their number of views, normalized by the amount of time they were available on the Web. We consider as popular the top 25% of stories (3,642 stories) based on the number of views (over 377 views/year).

6.3.1 FEATURES OF THE STORIES

We considered the distributions of several features of the stories: number of Web elements, the number of text elements, and the editing time. We also check if there is a relationship between the popular stories and the relative number of subscribers. Furthermore, we test if popular stories are different from the unpopular stories using the Kruskal-Wallis test [191], which allows comparing two or more samples that are independent and have different sample sizes.

We found that at the $p \leq 0.05$ significance level, the popular and the unpopular stories are different in terms of the following features: number of Web elements, text elements, editing time, and subscribers. Figure 65(a) shows that popular stories tend to have more Web elements (medians of 28 vs. 21) and a longer editing time (5 hours vs. 2 hours) than the unpopular stories. The number of elements in the popular stories is between 2 to 1950 Web elements with $median = 28$ and $mode = 10$, and the number of text elements ranges from 0 to 559 with $median = 1$ and $mode = 0$. The popular stories tend to have longer editing time intervals than the unpopular stories. For the popular stories, 38% have an editing time of at least one day, while only 35% of the unpopular stories have this feature. The maximum editing time in the popular stories is 4.1 years, while it is 3.5 years for unpopular stories.

There is a large difference between the number of subscribers for authors of popular stories than for those of unpopular stories. The authors of popular stories have min/median/max values of 0/16/1,726,143 subscribers, while the authors of unpopular stories have 0/2/2,469 subscribers.

6.3.2 THE TYPE OF ELEMENTS

Figure 65(b) shows the distributions for the popular and the unpopular stories for each element type. The popular stories tend to have more images than the unpopular stories. The median number of images in popular stories is 10, while it is 5 in the unpopular stories.

For videos, the median is 2 for both popular and unpopular. We found that the median number of the links in popular stories (20 links) is higher than the unpopular stories (16 links). We also test if the types of elements used in popular stories are different from the unpopular stories using the Kruskal-Wallis test and found that $p \leq 0.05$ for the distributions of each of the elements (images, videos, links, and quotes).

6.3.3 DO POPULAR STORIES HAVE A LOWER DECAY RATE?

We checked the decay rate of the popular and the unpopular stories to investigate if there is a relationship between popularity and lower decay rate. We found that for the popular stories, 11.0% of the resources were missing, while 12.8% of the resources were missing for unpopular stories. Figure 65(a) contains the distribution of the percentage of missing resources per story in popular and unpopular stories. It shows that the resources of the popular stories have lower decay than the resources of the unpopular. A reason could be that the popular stories are edited, and edits could be fixing broken links. The 75th percentile of the decay rate per popular story is 10% of the resources, while it is 15% in the unpopular stories.

6.4 SUMMARY

We presented the structural characteristics of human-generated stories on Storify, with particular emphasis on “popular” stories (i.e., the top 25% of views, normalized by time available on the Web) [19, 21]. To answer the research questions that were listed earlier, we analyzed 14,568 stories from Storify comprising 1,251,160 elements. We found that popular stories have a min/median/max values of 2/28/1,950 elements, with the unpopular stories having 2/21/2,216 elements. Popular stories have a median of 12 multimedia resources (the unpopular stories have a median of 7), 38% receiving continuing edits (as opposed to 35%), and only 11% of Web elements are missing on the live Web (as opposed to 13%). The authors of popular stories have min/median/max values of 0/16/1,726,143 subscribers, while the authors of unpopular stories have 0/2/2,469 subscribers. We found that there is a nearly linear relationship between the editing time of the story and the number of Web elements. We found that `twitter.com` dominates the Web resources of Storify stories. We also found that only 11% of the missing resources could be found in public Web archives.

Studying human-generated stories in Storify helped us to profile different kinds of stories by examining the typical length (in terms of the number of resources included), time frames covered, structural metadata (e.g., PageRank, images and video, social media vs. news) and other features. We model the structural characteristics of these stories, with particular emphasis on “popular” stories. For example, we generate stories automatically from archived collections with a typical length close to 28 (more or less based on the collection size).

In Chapter 7, we will investigate the characteristics of the archived collections using a dataset from Archive-It for specifying what can be applied in the DSA framework for generating stories from these collections. The structural characteristics of human-generated stories, such as the number of elements and the distribution of domains, will provide us with a template with which to evaluate our automatically generated stories.

CHAPTER 7

CHARACTERISTICS OF ARCHIVE-IT COLLECTIONS

Since archived collections will be our source for creating stories, we want to understand the characteristics of these collections and determine the most used resources in the archived collections. We quantified the collections in terms of the mean and median number of URIs in a collection, the typical crawl depth and breadth, etc. [21]. We built a baseline of what is inside the archived collections, based on the analysis of 3,109 collections with 305,522 seed URIs, for clarifying the intended framework of our archival summaries characteristics. In this chapter, we investigate the following questions:

- What is the mean and median number of URI-Rs in a collection?
- What is the mean number of mementos per seed URI in a collection?
- What are the types of resources used in these collections?
- What are the most frequent domains in the collections?
- What is the timespan of the collections?
- What are the similarities and differences between the Storify stories and Archive-It collections?

7.1 CHARACTERISTICS OF ARCHIVED COLLECTIONS

In this section, we check the population of Archive-It collections for better understanding the characteristics of the collections we intend to summarize.

7.1.1 ARCHIVE-IT COLLECTIONS

As of November 2015, we obtained the IDs of the whole population of Archive-It collections from the front-end interface of Archive-It. We excluded the collections that we knew were created automatically (the seed URIs have been extracted automatically from the Web), and also collections with no data. We kept collections with one URI because they have mementos. The number of remaining collections is 3,109, comprising 305,522 seed URIs. The total number of mementos for all the collections is 2,385,397. We downloaded

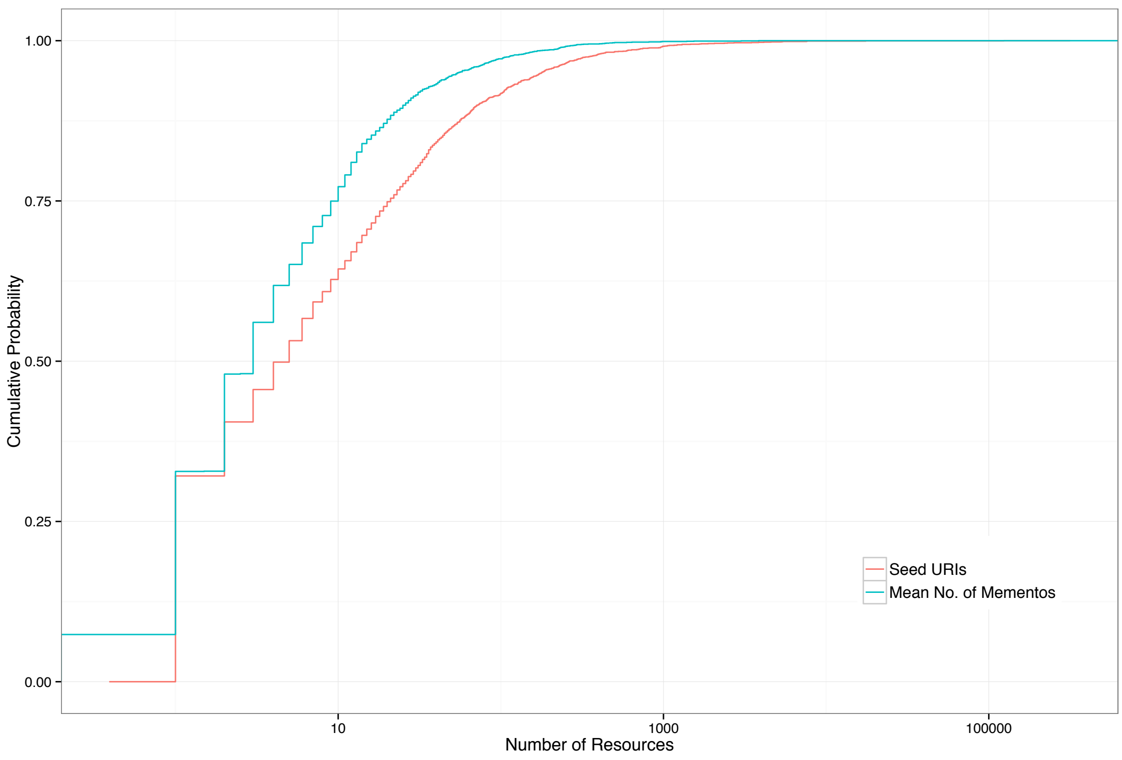


FIG. 66: The distribution of the number of seed URIs and the mean number of mementos per seed in Archive-It collections.

the metadata of all seed URIs in November 2015. For each seed URI, we obtained its first crawling date, last crawling date, and number of mementos.

7.1.2 GENERAL CHARACTERISTICS

Table 13 shows the characteristics of Archive-It collections in terms of the number of seed URIs, the mean number of the mementos per seed, and timespan, which is the range of time period over which the Web pages have been archived. The mean number of seed URIs in Archive-It collections is 98 URIs, and the median is 5 URIs. The mean number of mementos is 17 mementos per seed URI, and the mean timespan is 21 months. Figure 66 contains the distribution of the number of seed URIs and the mean number of mementos per seed in each collection.

The largest collection in terms of the number of seed URIs is the "Government of Canada Publications"[1] collection that archives Canadian governmental pages, created by

[1] https://archive-it.org/collections/3572/

TABLE 13: Distribution of features of Archive-It collections. Timespan is measured in days.

Features	**Seed URIs**	**Mementos**	**Timespan**
25th percentile	1	1	0
50th percentile	5	3	154
75th percentile	21	9	973
90th percentile	73	26	1,791
Maximum	123,600	3,848	6,945
Mean	98	17	628
Std. Dev.	2,260	106	921

the Canadian Government Information PLN Web Archive[2]. It contains 123,647 URIs with a span of 2 years (2013-2015) and a mean of 2 mementos for each URI. The largest timespan in the collections is 19 years (from 1996 until 2015) for only 21 seed URIs. The start date of crawling for multiple collections is before the existence of Archive-It in 2006. This is possible because some organizations imported previously archived pages to initialize their collections.

7.1.3 WHAT DOMAINS ARE USED IN COLLECTIONS?

Canonicalizing the domains of 305,522 URIs resulted in 57,640 unique domains in the 3,109 collections. Figure 67 shows the relationship between the frequency of domains and the number of collections they appeared in. For example, the dot at the top of the graph represents the most frequent domain, which appears over 100,000 times in only 4 different collections. We notice that multiple domains in Archive-It collections have a high frequency, but appear in only a few collections. This is because some collections are devoted to archiving specific domains.

Table 14 contains the top 25 domains of the resources ordered by their frequency. The list of top 25 domains represents 66.1% of all the resources. The table also contains the global rank of the domains according to Alexa as of March 2015. We also added our manual categorization for the domains. We notice that the most used domain is `publications.gc.ca` from the "Government of Canada Publications" collection, which contains the largest number of URI-Rs. We added the collection counts to the table to reflect the global rank of the domains across the collections. We notice that the first ranked domain based on the frequency of the domains appeared in only four collections. The table also shows that most of domains in the top list are for government and education Web sites.

[2] `https://archive-it.org/organizations/700/`

TABLE 14: The top 25 domains based on the frequency of appearance in Archive-It. The percentage is the frequency of the domain out of 305,522. Alexa global rank was retrieved in 2015-11.

Domain	**Frequency**	**Percentage**	**Collection Count**	**Alexa Global Rank**	**Category**
publications.gc.ca	123,604	40.46%	4	192,814	Government
youtube.com	21,838	7.15%	337	3	Videos
mtholyoke.edu	7,632	2.50%	3	34,718	Education
nsa.gov	7,625	2.50%	5	49,313	Government
blogspot.com	6,072	1.99%	305	38	Blogs
nsf.gov	5,312	1.74%	3	15,613	Government
facebook.com	5,268	1.72%	480	2	Social media
hem.bredband.net	4,582	1.50%	1	367,103	Company
wikipedia.org	4,405	1.44%	93	7	Encyclopedia
twitter.com	3,089	1.01%	460	9	Social media
nlm.nih.gov	2,030	0.66%	20	196	Government
wayback.archive-it.org	1,791	0.59%	4	133,005	Archive
wordpress.com	1,471	0.48%	276	36	Blogs
vimeo.com	1,354	0.44%	44	186	Blogs
uwrf.edu	1,218	0.40%	2	157,000	Education
pubs.pembina.org	1,196	0.39%	1	709,328	Education
hhs.gov	579	0.19%	15	8,641	Government
globe.gov	462	0.15%	2	559,353	Government
flickr.com	460	0.15%	132	159	Education
netfiles.uiuc.edu	429	0.14%	2	17,442	Education
orgsync.com	356	0.12%	6	12,450	Company
nytimes.com	330	0.11%	69	97	News
tumblr.com	328	0.11%	102	43	Blogs
baylor.edu	274	0.09%	12	17,643	Education
rochester.edu	254	0.08%	12	9,093	Education

TABLE 15: The top 25 domains based on the number of Archive-It collections they appear in. The percentage is the number of collections the domain appeared in out of 3,109. Alexa global rank was retrieved in 2015-11.

Domain	**Collection Count**	**Percentage**	**Frequency**	**Alexa Global Rank**	**Category**
facebook.com	480	15.44%	5,268	2	Social media
twitter.com	460	14.80%	3,089	9	Social media
youtube.com	337	10.84%	21,838	3	Videos
blogspot.com	305	9.81%	6,072	38	Blogs
wordpress.com	276	8.88%	1,471	36	Blogs
flickr.com	132	4.25%	460	159	Photos
tumblr.com	102	3.28%	328	43	Blogs
wikipedia.org	93	2.99%	4,405	7	Encyclopedia
ok.gov	92	2.96%	141	24,315	Government
instagram.com	78	2.51%	203	24	Photos
nytimes.com	69	2.22%	330	97	News
sites.google.com	69	2.22%	194	1	Wikipedia
tn.gov	53	1.70%	153	13,494	Government
bbc.com	52	1.67%	183	100	News
slco.org	52	1.67%	74	100,152	Government
cnn.com	51	1.64%	147	75	News
sfgov.org	50	1.61%	61	101,777	Government
huffingtonpost.com	47	1.51%	149	122	News
tennessee.gov	46	1.48%	57	175,859	Government
yahoo.com	45	1.45%	85	5	Search
vimeo.com	44	1.42%	1,354	186	Videos
weebly.com	43	1.38%	142	252	Company
typepad.com	39	1.25%	70	1,126	Blogs
washingtonpost.com	36	1.16%	180	198	News
pinterest.com	36	1.16%	55	30	Photos

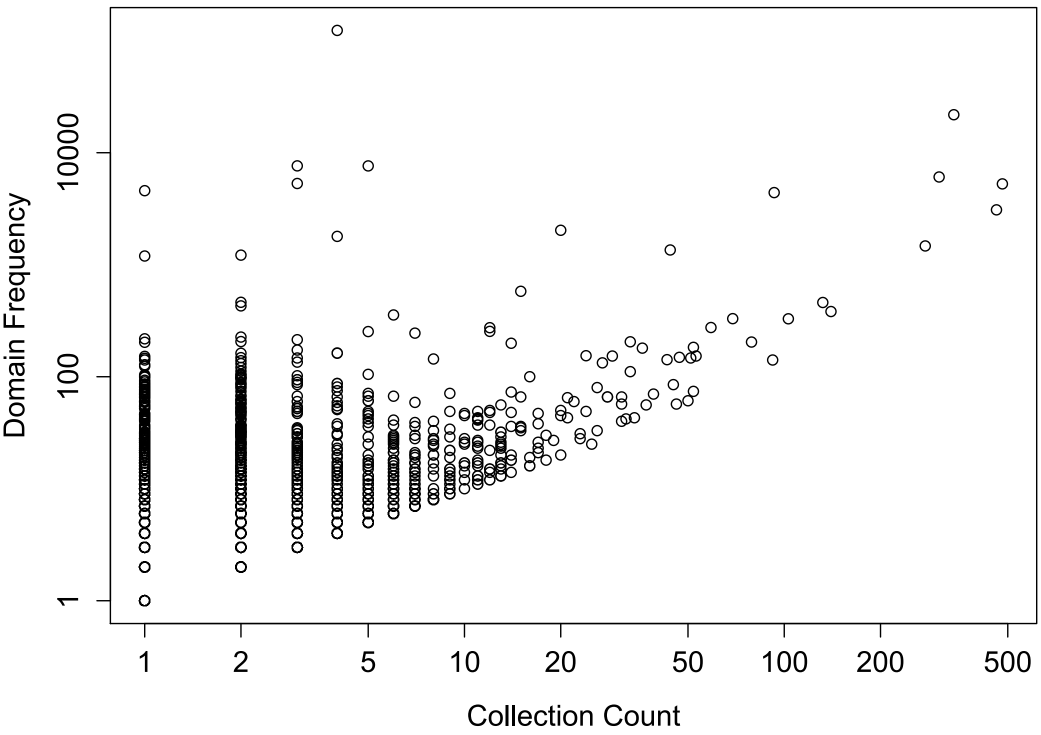


FIG. 67: The relationship between the frequency of the domains in Archive-It collections and the number of collections in which those domains appear.

There are also blogs and social media Web sites, such as `facebook.com` and `twitter.com`. Table 14 also shows that some collections use archived URIs in their seed list. The domain `wayback.archive-it.org` is ranked 12th based on its frequency and appeared in four collections.

Table 15 shows the top 25 domains based on the number of collections that they appeared in. It is clear from the table that the top list of domains based on the number of collections they appeared in is different from the top domains based on the frequency. Note that `sites.google.com` has rank one because Alexa does not differentiate `sites.google.com` from `google.com`. In Section 7.1.5, we investigate the correlation between the rank of the domains within Archive-It collections and their Alexa global rank.

7.1.4 CLASSIFICATION OF SEED URIS BASED ON THE TLD

TABLE 16: The top 10 TLDs of the resources.

TLD	**Percentage**
.ca	41.96%
.com	23.73%
.edu	9.77%
.org	8.50%
.gov	8.21%
.net	2.24%
.us	0.70%
.uk	0.61%
.de	0.38%
.fr	0.31%

TABLE 17: The Kendall's Tau between the most frequent n domains in the stories and their Alexa global rank (τ_{af}) and between the top n domains that have the most number of collections and Alexa global rank (τ_{ac}).

n	10	15	25	50	100
τ_{af}	-0.2000	0.0286	-0.0467	0.0008	**0.1741**
τ_{ac}	0.4222	**0.4857**	**0.4174**	**0.4399**	**0.3180**

Table 16 presents the distribution of TLDs for the seed URIs in Archive-It collections (only the top 10 are shown). The top 10 list represents 97.8% of the TLDs in the collections. It can be noticed that most of the URIs are for the .ca, .com, .edu, .org, .gov, .net, .us, .uk, and .de domains. The .ca comes from the `publications.gc.ca`, which dominates the top 25 most frequent domains. We notice that there are many governmental, organizational, and educational sites in the collections.

7.1.5 CORRELATION OF GLOBAL AND ARCHIVE-IT POPULARITY

Table 17 shows Kendall's Tau correlation τ_{af} for the most frequent n domains in Archive-It collections and their Alexa global rank. It also shows Kendall's Tau correlation τ_{ac} for the top n domains based on the largest number of collections and their Alexa global rank. Statistically significant ($p \leq 0.05$) correlations are bolded. The table shows that the correlation between the most frequent n domains and their Alexa global rank is very low. The highest correlation between the most frequent n domains and the Alexa global rank is 0.17 for the list of the 100 domains. This may be due to the nature of the collections and the purpose for which they are created. Most of the collections are

TABLE 18: The distributions of the number of collections in each time interval.

		Seed URIs			**Ave. no. of URI-Ms/Seed**		
Intervals	**Percentage**	**Mean**	**Median**	**Max.**	**Mean**	**Median**	**Max.**
URI with no captures	6.80%	7	1	412	0	0	0
< 1 day	21.00%	24	1	7,619	1.1	1	4.8
1-7 days	4.90%	101	5	7,590	2.6	2	10.1
1-4 weeks	4.60%	28	12	495	3.7	2.7	29.8
1-12 months	19.90%	66	10	5,309	10.9	3.4	277.6
1-4 years	25.40%	242	6	123,648	16.6	5	594.5
> 4 years	17.30%	69	10	2,365	59.5	13.7	3848

explicitly centered around topics. Furthermore, some collections archive specific domains (e.g., `publications.gc.ca`). Many of these domains are not high ranked globally, but the collections they appeared in have a large number of seed URIs, which results in high frequency for these domains.

Although the frequency of domains does not correlate with the globally high ranked domains, the top list of the domains based on the number of collections they appeared in highly correlates with the global rank of these domains. For most of the top n domains across Archive-It collections $\tau_{ac} > 0.4$. The highest correlation is 0.49 for the list of 15 domains.

7.1.6 WHAT IS THE MEAN TIMESPAN FOR DIGITAL COLLECTIONS?

Table 18 shows the percentage of the collections that have been crawled in each time interval. The table also shows the corresponding features of the collection in terms of the number of seed URIs and the mean number of mementos per seed. Note that the timespan of the collection is different from the editing time of Storify stories.

The first row contains collections with 0 mementos as of November 2015. About 20% of these collections have been created recently and their crawling date started after we captured the metadata of the collections. Among these collections, the collection with the largest number of URIs in this category ("Cal Poly University Web Archive"[3]) has 412 seed URIs.

We see that the majority of the collections have a long timespan, meaning that they have been crawled over the span of years. There are 17% of the collections with a span of

[3] `https://archive-it.org/collections/6191/`

more than four years. The collection with the longest timespan of 19 years has URIs that were crawled before Archive-It existed.

From Table 18, we notice that there is a linear relationship between the mean number of mementos per seed in the collection and the timespan of the collection. The mean number of mementos per seed URIs increases with an increase in the timespan of the collection. The mean number of mementos in the span of 4 years (or more) is 60 mementos per seed, and goes down 70% to be 17 mementos per seed in the span of 1-4 years.

7.1.7 THE DECAY RATE IN ARCHIVE-IT COLLECTIONS

In Chapter 4, we found that most people come to the Web archives because they did not find the pages on the live Web [17]. We extracted 293,883 unique seed URIs from Archive-It collections and checked their existence on the live Web. We found that 8.3% (24,521 out of 293,883) of the seed URIs in Archive-It collections are missing from the live Web. Missing represents the requests that return HTTP 4xx, HTTP 5xx, HTTP 3xx to others except 200, timeouts, and soft 404s. Note that 42% of the seed URIs belong to the "Government of Canada Publications" collection, which is devoted to archiving governmental publication documents that are well preserved by the Canadian government. We measured the loss for this collection and found that only 0.1% (102 URIs out of 122,948 unique URIs) of the documents are missing. For these kind of collections, we expect that if the domain is lost or unavailable for any reason [40, 222], all the 122,948 URIs might disappear. Excluding the "Government of Canada Publications" collection, the decay rate for the rest of the collections is 14.3% (24,419 out of 170,935 unique URIs).

We also found that 58.7% (1825 out of 3109) of the collections contain seed URIs that had disappeared from the live Web. Of these, 22.5% (410 out of 1825) have 100% loss of their seed URIs from the live Web.

7.2 ARCHIVE-IT COLLECTIONS VERSUS STORIFY STORIES

In this section, we contrast the general characteristics of human-generated stories from Storify that were presented in Chapter 6 and human-curated collections from Archive-It. Figures 62 in Chapter 6 and Figure 67 show that the most frequent domains in Storify appeared in the majority of stories, while many of the most frequent domains in Archive-It appeared in few collections. For example, the most frequent domain in Storify (`twitter.com`), which is represented in Figure 62 by the rightmost dot, appeared almost 1,000,000 times in the largest number of stories (over 10,000 stories). On the other hand, the most frequent domain in Archive-It collections (`publications.gc.ca`), which is represented by the dot on the top left of Figure 67, appeared over 100,000 times in only four collections.

The difference in the nature of the domains could be due to the difference of who is creating the collection: regular users (Storify), or librarians employed by government, museums, etc. (Archive-It).

Also, the most frequent domains in the stories have a higher correlation with the Alexa global rank than the most frequent domains in the archived collection as shown in Tables 10 and 17. For most of the n values in Table 10 in Chapter 6, there is a high correlation between the most frequent n domains in the stories and their Alexa global Rank (τ_{sf}) The τ_{sf} at $n = 15$ is 0.45, while in Archive-It collections, the list of the most frequent 15 domains and their Alexa global rank (τ_{af}) are statistically independent (Table 17). The largest value of the τ_{af} is 0.17 at $n = 100$.

Additionally, Tables 9 and 16 show that the list of TLDs in Storify is dominated by .com, which represents 96.5% of the resources, while it represents only 23% in Archive-It collections. The list of TLDs in Archive-It collections contains a significant existence for .gov and .edu domains. That is because many collections are devoted to archiving governmental pages (e.g., all Web pages published by the state of California) and memory organizations like libraries and museums, but many of the collections are explicitly centered around topics in arts and humanities, politics, spontaneous events, and blogs and social media.

For the decay rate, 11.8% of Storify resources do not exist on the live Web, while 8.3% of Archive-It URIs are missing. Although the decay rate in Storify stories is larger than the decay rate of Archive-It collections, the percentage of the affected collections (58.7%) is larger than the percentage of the affected stories (40.8%). Furthermore, the mean value of the missing elements per story is 10.3%, although the mean value of the missing seed URIs per collection is 42%.

To conclude, the resources that are used in Storify stories are different from the resources in Archive-It collections. In summarizing a collection, we can only choose from what is archived. So if there are no tweets in the collection, `twitter.com` will not be the most common domain in the generated stories. Although some content in Storify stories will not be applicable (e.g., `twitter.com` is popular in Storify, but mostly missing in Archive-It collections), some other characteristics will be applicable, such as the number of resources. Accordingly, our choices of what to select from the collection needs to be informed by what constitutes a “popular” story.

7.3 SUMMARY

We presented a baseline for the characteristics of archived collections based on qualifying the whole population of Archive-It collections [21]. We checked the resources in these collection, the timespan of the collections, the average number of URI-Ms per seed, and so on. The most frequent domain in Archive-It collections is `publications.gc.ca`, which appeared over 100,000 times in only four collections. Furthermore, the most frequent domains in the Archive-It collections have very low correlation with the Alexa global rank. The list of TLDs in Archive-It collections contains a significant existence for .gov and .edu domains. For the decay rate, 8.3% of Archive-It URIs are missing from the live Web with 42% mean value of the missing elements per collection.

We found that some characteristics of human-generated stories may not be possible to apply because the nature of the resources in the stories is different from what compose the collections. For example, we found that `twitter.com` is popular in Storify, but mostly is missing in Archive-It. Our choices of what to select from the collection needs will be informed by what constitutes a "popular" story. For example, we will use the median number of resources in the popular stories ($k = 28$) as a default value for the number of resources in the automatically generated stories, as will be explained in Chapter 9.

CHAPTER 8

DETECTING OFF-TOPIC PAGES IN WEB ARCHIVES

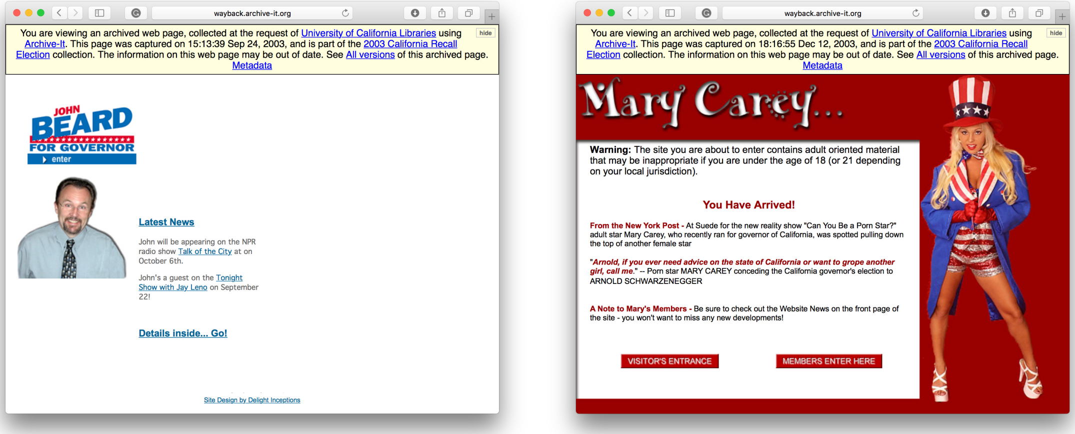
As we declared in Chapter 2, Archive-It provides their partners with tools that allow them to build themed collections of archived Web pages hosted on Archive-It's machines. This is done by the user manually specifying a set of seed URIs that should be crawled periodically. Archive-It has deployed tools that allow a collection's curators to perform quality control on their crawls, as shown in Figure 24(b) of Chapter 2. However, the tools are currently focused on issues such as the mechanics of HTTP (e.g., how many HTML files vs. PDFs, how many HTTP 404 responses) and domain information (e.g., how many .uk sites vs. .com sites). Currently, there are no content-based tools that allow curators to detect when seed URIs go off-topic.

In this chapter, we introduce different approaches for detecting off-topic pages in individual TimeMaps (Section 8.4.2). Those predicted off-topic pages will be presented to the collection's curator for possible elimination from the collection or cessation of crawling. For evaluating the proposed methods, we built our gold standard dataset from three Archive-It collections, then we employ the following performance measurements: accuracy, F_1 score values, and area under the ROC curve (AUC) (Sections 8.5). We evaluate the performance of the best performed method on several Archive-It collections (Section 8.6).

8.1 MOTIVATING EXAMPLES

We can define off-topic pages as the Web pages that have changed through time to move away from the initial scope of the page. There are multiple reasons for pages to go off-topic, such as hacking, loss of account, domain expiration, owner deletion, or server/service discontinued [222]. Expired domains should return a 404 HTTP status that will be caught by Archive-It quality control methods. However, some expired domains may be purchased by spammers who desire all the incoming traffic that the site accrued while it was "legitimate" (see Figure 68). In this case, the Web page returns a 200 HTTP response but with unwanted content [40].

There are also many cases in which the archived page redirects to another page which is not relevant but still not spam. In Figure, 69 the Facebook page contained relevant content in the beginning (Figure 69(a)), then later redirects to the homepage of Facebook as shown in Figure 69(b). The example in Figure 69 shows how a page in a collection goes off-topic, even though the particular Web site has not been lost.



You are viewing an archived web page, collected at the request of University of California Libraries using Archive-It. This page was captured on 15:13:39 Sep 24, 2003, and is part of the 2003 California Recall Election collection. The information on this web page may be out of date. See All versions of this archived page. Metadata

Latest News

John will be appearing on the NPR radio show Talk of the City at on October 6th.

John's a guest on the Tonight Show with Jay Leno on September 22!

Details inside... Go!

Site Design by Delight Inceptions

You are viewing an archived web page, collected at the request of University of California Libraries using Archive-It. This page was captured on 18:16:55 Dec 12, 2003, and is part of the 2003 California Recall Election collection. The information on this web page may be out of date. See All versions of this archived page. Metadata

Warning: The site you are about to enter contains adult oriented material that may be inappropriate if you are under the age of 18 (or 21 depending on your local jurisdiction).

You Have Arrived!

From the New York Post - At Suede for the new reality show "Can You Be a Porn Star?" adult star Mary Carey, who recently ran for governor of California, was spotted pulling down the top of another female star

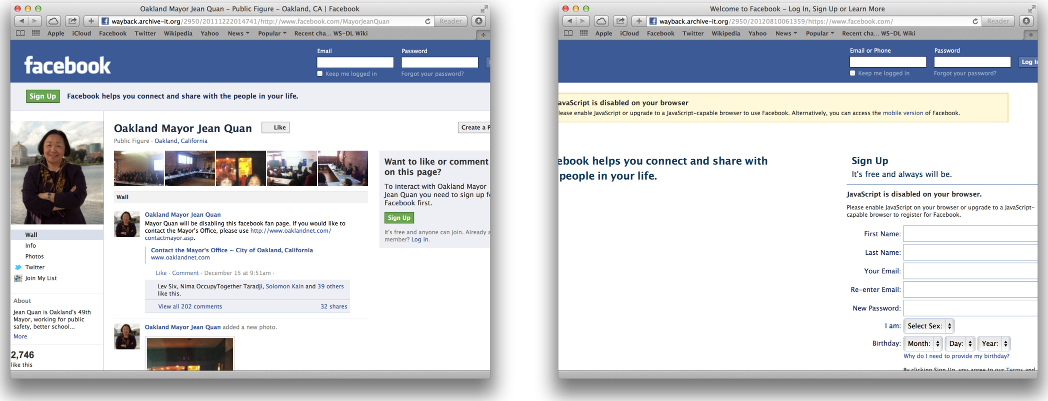
"*Arnold, if you ever need advice on the state of California or want to grope another girl, call me.*" -- Porn star MARY CAREY conceding the California governor's election to ARNOLD SCHWARZENEGGER

A Note to Mary's Members - Be sure to check out the Website News on the front page of the site - you won't want to miss any new developments!

(a) Sept. 24, 2003: johnbeard4gov.com was for a California gubernatorial candidate.

(b) Dec. 12, 2003: johnbeard4gov.com became spam.

FIG. 68: Example of johnbeard4gov.com in the 2003 California Recall Election collection that went off-topic.


Oakland Mayor Jean Quan – Public Figure – Oakland, CA | Facebook
wayback.archive-it.org/2950/20111222014741/http://www.facebook.com/MayorJeanQuan
Reader
Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki
facebook
Email
Password
Keep me logged in
Forgot your password?
Sign Up
Like
Create a P
Wall
Info
Photos
Twitter
Join My List
Welcome to Facebook – Log In, Sign Up or Learn More
wayback.archive-it.org/2950/20120810061359/https://www.facebook.com/
Reader
Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki
Email or Phone
Password
Log In
Keep me logged in
Forgot your password?

Facebook helps you connect and share with the people in your life.

Oakland Mayor Jean Quan

Public Figure · Oakland, California

Wall

Oakland Mayor Jean Quan
Mayor Quan will be disabling this facebook fan page. If you would like to contact the Mayor's Office, please use http://www.oaklandnet.com/contactmayor.asp.

Contact the Mayor's Office ~ City of Oakland, California
www.oaklandnet.com

Like · Comment · December 15 at 9:51am ·

Lev Six, Nima OccupyTogether Taradji, Solomon Kain and 39 others like this.

View all 202 comments 32 shares

Oakland Mayor Jean Quan added a new photo.

Want to like or comment on this page?

To interact with Oakland Mayor Jean Quan you need to sign up f Facebook first.

Sign Up

It's free and anyone can join. Already a member? Log in.

About

Jean Quan is Oakland's 49th Mayor, working for public safety, better school...

More

2,746
like this

avaScript is disabled on your browser
lease enable JavaScript or upgrade to a JavaScript-capable browser to use Facebook. Alternatively, you can access the mobile version of Facebook.

ebook helps you connect and share with people in your life.

Sign Up
It's free and always will be.

JavaScript is disabled on your browser.

Please enable JavaScript on your browser or upgrade to a JavaScript-capable browser to register for Facebook.

First Name:
Last Name:
Your Email:
Re-enter Email:
New Password:
I am: Select Sex:
Birthday: Month: Day: Year:
Why do I need to provide my birthday?

By clicking Sign Up, you agree to our Terms and

(a) Dec. 22, 2011: Facebook page was relevant to the Occupy collection.

(b) Aug. 10, 2012: URI redirects to www.facebook.com.

FIG. 69: Example of a Facebook page from the Occupy Movement collection that went off-topic.

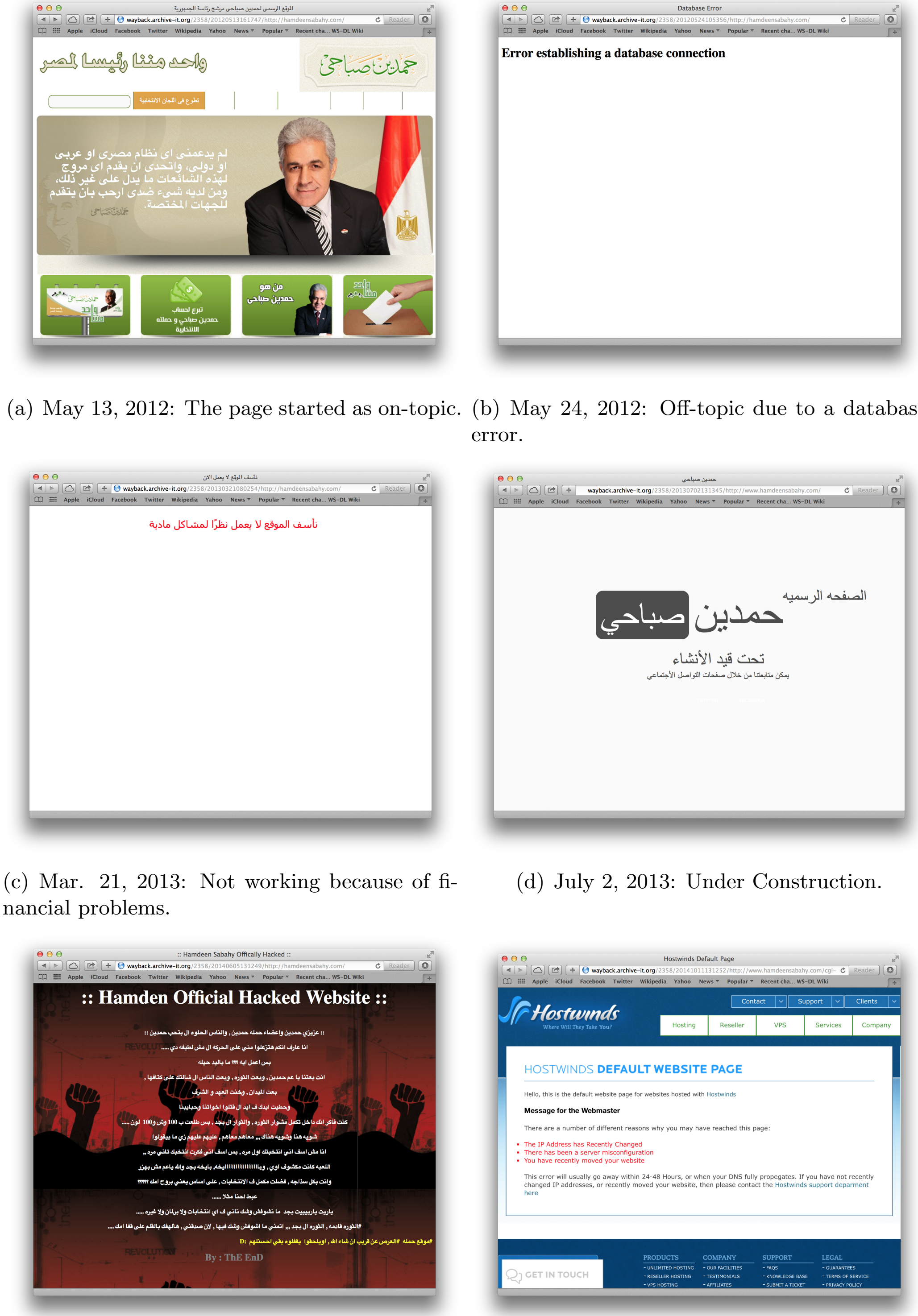
TABLE 19: Description of the Archive-It collections.

Collection Name	**Occupy Movement 2011/2012**	**Egypt Revolution and Politics**	**Human Rights**
Collection ID	2950	2358	1068
Curator	Internet Archive Global Events	American University in Cairo	Columbia University Libraries
Timespan	2011/12/03 - 2012/10/09	2011/02/01 - 2013/04/18	2008/05/15 - 2013/03/21
Total URI-Rs	728	182	560
Total URI-Ms	21,268	18,434	6,341

Figure 70 shows a scenario of a page that goes off-topic for many different reasons. In May 2012, `hamdeensabahy.com` Web page, which belonged to a famous politician and a candidate in Egypt's 2012 presidential election, was originally relevant to the "Egypt Revolution and Politics" collection (Figure 70(a)). Then, the page went back and forth between on-topic and off-topic many times for different reasons. Note that there are on-topic pages between the off-topic ones in Figure 70. In the example, the page went off-topic because of a database error on May 24, 2012 (Figure 70(b)), then it returned on-topic again. After that, the page went off-topic because of financial issues (Figure 70(c)). The page continued off-topic for a long period (from March 27, 2013 until July 2, 2013) because the site was under construction (Figure 70(d)). The page went on-topic again for a period of time, then the site was hacked (Figure 70(e)), and then the domain was lost by late 2014 (Figure 70(f)).

The Web page `hamdeensabahy.com` has 266 mementos. Of these, over 60% are off-topic. While it might be useful for historians to track the change of the page in Web archives (possibly the hacked version is a good candidate for historians), the 60% off-topic mementos such as the ones in Figures 70(b) - 70(f) do not contribute to the Egypt Revolution collection in the same way that the on-topic archived Web site in Figure 70(a) does.

Although the former can be kept in the IA's general Web archive, they are candidates to be purged from the Egyptian Revolution collection. Even if the pages are kept in the collection, we exclude them from consideration for generating stories (Chapter 9).


الموقع الرسمى لحمدين صباحى مرشح رئاسة الجمهورية
wayback.archive-it.org/2358/20120513161747/http://hamdeensabahy.com/
واحد مننا رئيسا لمصر
حمدين صباحي
تطوع فى اللجان الانتخابية
تبرع لحساب
حمدين صباحي و حملته
الانتخابية
من هو
حمدين صباحي

لم يدعمنى اى نظام مصرى او عربى او دولى، واتحدى ان يقدم اى مروج لهذه الشائعات ما يدل على غير ذلك، ومن لديه شىء ضدى ارحب بان يتقدم للجهات المختصة.

(a) May 13, 2012: The page started as on-topic.

Database Error
wayback.archive-it.org/2358/20120524105356/http://hamdeensabahy.com/

Error establishing a database connection

(b) May 24, 2012: Off-topic due to a database error.

نأسف الموقع لا يعمل الان
wayback.archive-it.org/2358/20130321080254/http://hamdeensabahy.com/

نأسف الموقع لا يعمل نظرًا لمشاكل مادية

(c) Mar. 21, 2013: Not working because of financial problems.

حمدين صباحى
wayback.archive-it.org/2358/20130702131345/http://www.hamdeensabahy.com/

الصفحه الرسميه
حمدين صباحي

تحت قيد الأنشاء

يمكن متابعتنا من خلال صفحات التواصل الأجتماعي

(d) July 2, 2013: Under Construction.

:: Hamdeen Sabahy Offically Hacked ::
wayback.archive-it.org/2358/20140605131249/http://hamdeensabahy.com/

:: Hamden Official Hacked Website ::

:: عزيزي حمدين واعضاء حمله حمدين , والناس الحلوه ال بتحب حمدين ::
انا عارف انكم هتزعلوا مني على الحركه ال مش لطيفه دي .....
بس اعمل ايه ؟؟؟ ما باليد حيله
انت بعتنا يا عم حمدين , وبعت الثوره , وبعت الناس ال شالتك على كتافها ,
بعت الميدان , وخنت العهد و الشرف
وحطيت ايدك ف ايد ال قتلوا اخواتنا وحبايبنا
كنت فاكر انك داخل تكمل مشوار الثوره , والثوار ال بجد , بس طلعت ب 100 وش و100 لون .....
شويه هنا وشويه هناك ,,, معاهم معاهم , عليهم عليهم زي ما بيقولوا
انا مش اسف اني انتخبتك اول مره , بس اسف اني فكرت انتخبك تاني مره ,,
اللعبه كانت مكشوف اوي , وبااااااااااااااايخه بايخه بجد والله ياعم مش بهزر
وانت بكل سذاجه , فضلت مكمل ف الانتخابات , على اساس يعني بروح امك ؟؟؟؟؟
عبط احنا مثلا ......
ياريت ياربيييت بجد ما نشوفش وشك تاني ف اي انتخابات ولا برلمان ولا غيره .....
#الثوره قادمه , الثوره ال بجد ,,, اتمني ما اشوفش وشك فيها , لان صدقني , هالهفك بالقلم على قفا امك ....
#موقع حمله #العرص عن قريب ان شاء الله , اويلحقوا يقفلوه بقي احسنلهم :D

By : ThE EnD

(e) June 5, 2014: The site has been hacked.

Hostwinds Default Page
wayback.archive-it.org/2358/20141011131252/http://www.hamdeensabahy.com/cgi-
Hostwinds
Where Will They Take You?
Contact
Support
Clients
Hosting
Reseller
VPS
Services
Company

HOSTWINDS **DEFAULT WEBSITE PAGE**

Hello, this is the default website page for websites hosted with Hostwinds

Message for the Webmaster

There are a number of different reasons why you may have reached this page:

- The IP Address has Recently Changed
- There has been a server misconfiguration
- You have recently moved your website

This error will usually go away within 24-48 Hours, or when your DNS fully propegates. If you have not recently changed IP addresses, or recently moved your website, then please contact the Hostwinds support deparment here

(f) Oct. 10, 2014: The domain has expired.

FIG. 70: A site for one of the candidates for Egypt's 2012 presidential election. Many of the captures of hamdeensabhay.com are not about the Egyptian Revolution. Later versions show an expired domain (as does the live Web version).

TABLE 20: The results of manually labeling the collections.

Collection	**Occupy Movement 2011/2012**	**Egypt Revolution and Politics**	**Human Rights**
Sampled URI-Rs	255 (35%)	136 (75%)	198 (35%)
Sampled URI-Ms	6,570	6,886	2,304
Off-topic URI-Ms	458 (7%)	384 (9%)	94 (4%)
URI-Rs with off-topic URI-Ms	67 (26%)	34 (25%)	33 (17%)

8.2 DATASET

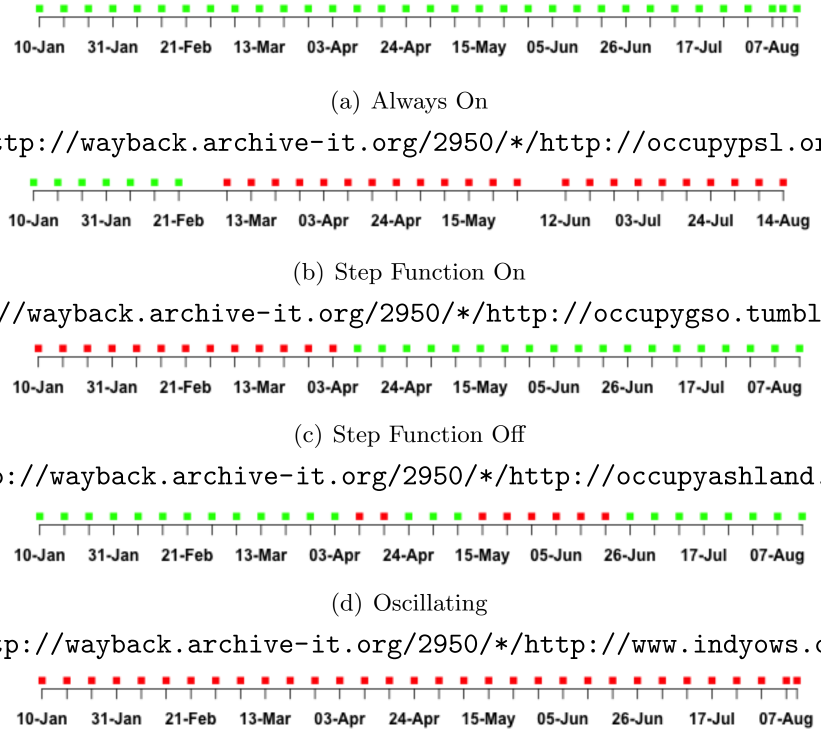
In this section we describe our gold standard dataset. We evaluate our techniques using the ODU mirror of Archive-It's collections. ODU has received a copy of the Archive-It collections (in the form of WARC files) through April 2013. The three collections in our dataset differ in terms of the number of URI-Rs, number of URI-Ms, and timespan, which is the range of time over which the Web pages have been archived. Next, we will describe the three collections that we constructed our samples from, then we will present the results of manually labeling the samples.

The "Occupy Movement 2011/2012" collection was built over a period of 10 months between Dec. 2011 - Oct. 2012 by Archive-It. This collection covers the Occupy Movement protests and the international branches of the Occupy Wall Street movement around the world. This collection contained 728 seed URIs and a total of 21,268 mementos.

The "Egypt Revolution and Politics" collection was started in Feb. 2011 and is still ongoing. This collection covers the January 25th Egyptian Revolution and Egyptian politics. It contains different kinds of Web sites (e.g., social media, blogs, news, etc.) that have been collected by the American University in Cairo. As of April 2013, this collection contained 182 seed URIs and a total of 18,434 mementos.

The "Human Rights" collection was started in May 2008 by Columbia University Libraries and is still ongoing. The Human Rights collection covers documentation and research about human rights that have been created by non-governmental organizations, national human rights institutions, and individuals. As of April 2013, this collection contained 560 seed URIs and a total of 6,341 mementos.

Table 19 provides the details of the three collections. The timespan in the table represents the range of the crawls for the ODU mirror which ends in April 2013. The collections contain pages in different languages, including English, Arabic, French, Russian, and Spanish.



(a) Always On
`http://wayback.archive-it.org/2950/*/http://occupypsl.org`

(b) Step Function On
`http://wayback.archive-it.org/2950/*/http://occupygso.tumblr.com`

(c) Step Function Off
`http://wayback.archive-it.org/2950/*/http://occupyashland.com`

(d) Oscillating
`http://wayback.archive-it.org/2950/*/http://www.indyows.org`

(e) Always Off
`http://wayback.archive-it.org/2950/*/http://occupy605.com`

FIG. 71: Example showing different behaviors for TimeMaps (green=on-topic, red=off-topic).

We randomly sampled 589 URI-Rs from the three collections (excluding URI-Rs with only one memento). Together, the sampled URI-Rs had over 18,000 URI-Ms, so for each of the sampled URI-Rs, we randomly sampled from their URI-Ms. This resulted in our manually labeling 15,760 mementos as on-topic or off-topic. We labeled the URI-M as off-topic if the content in the URI-M was no longer relevant to the content in the URI-R@t_0, which is assumed to be relevant to the topic of the collection.

Table 20 contains the results of manually labeling the sampled data of each collection. We sampled from 35% of the seed URIs of each collection, except for the Egypt Revolution collection; it has fewer URIs than the other two collections, so we sampled from 75% of its URIs. The labeled gold standard dataset is available for download at `https://github.com/yasmina85/OffTopic-Detection`.

We found that 24% of the TimeMaps we sampled contain off-topic pages. Detecting these pages automatically for the collection curator will not only avoid diluting the value

TABLE 21: The statistics of TimeMap behaviors in archived collections.

TimeMap Behavior	**Occupy Movement 2011/2012**	**Egypt Revolution and Politics**	**Human Rights**
Always On	73.7%	75.0%	83.3%
Step Function On	11.4%	11.0%	7.6%
Step Function Off	1.2%	0.7%	0.0%
Oscillating	13.3%	12.5%	9.1%
Always Off	0.4%	0.7%	0.0%

of their collections, but also will save the time required for a manual check of the relevance of the URIs and save the storage required for these pages.

8.3 TIMEMAP BEHAVIOR

Many studies have been performed on the key aspects of document "aboutness", such as the page's title [179], tags [177], lexical signatures [181], etc. Section 8.4.2 enumerates different methods we explored to distill a page's aboutness and quantify how this aboutness changes through time. Here, we define five general classes of TimeMaps based on how a page's aboutness changes through time. Table 21 shows the percentage of each type of TimeMap present in our three manually labeled collections.

As defined in Chapter 5, an Archive-It collection (C) is a set of seed URIs collected by the users from the Web (W), where $C \subset W$ (Equation 5). Each seed URI ($URI\text{-}Rs$) has many different mementos ($URI\text{-}Ms$), and a set of mementos for a seed URI composes a TimeMap ($URI\text{-}T$).

We define $URI\text{-}R@t$ to be: on-topic, if $aboutness(URI\text{-}R@t) \approx aboutness(URI\text{-}R@t_0)$ and off-topic, if $aboutness(URI\text{-}R@t) \not\approx aboutness(URI\text{-}R@t_0)$, where $URI\text{-}R@t_0$ is relevant to C.

For the gold standard dataset (Section 8.1), we manually assess if a memento is relevant to C. We empirically observed five classes of TimeMaps based on the page's aboutness.

Always On: This is the ideal case, in which the page does not go off-topic (Figure 71(a)):

$\forall t\ aboutness(URI\text{-}R@t) \approx aboutness(URI\text{-}R@t_0),$ and $URI\text{-}R@t_0$ is relevant to C.

This is the majority case in the gold standard dataset, with at least 74% of the TimeMaps always on-topic (Table 21).

Step Function On: $URI\text{-}R@t_0$ is on-topic, but then at some t goes off-topic and continues $\forall t$ (Figure 71(b)):

$\forall t \geq i$, where $i \geq 1$, and i is an integer, $aboutness(URI\text{-}R@t) \not\approx aboutness(URI\text{-}R@t_0)$, where $URI\text{-}R@t_0$ is relevant to C.

We found that 8-11% of the TimeMaps are Step Function On.

Step Function Off: $URI\text{-}R@t_0$ is off-topic, but then at some t goes on-topic and continues $\forall t$ (Figure 71(c)):

$\forall t \geq i$, where $i \geq 1$, and i is an integer, $aboutness(URI\text{-}R@t) \not\approx aboutness(URI\text{-}R@t_0)$, where $URI\text{-}R@t_0$ is *not* relevant to C.

The case when the TimeMap starts with an off-topic memento then goes on-topic is very rare. We found that only 0-1% TimeMaps are Step Function Off. This case violates our assumption that the $URI\text{-}R@t_0$ is relevant to C. In our gold standard dataset, we manually shifted the first memento to be the first memento relevant to the collection.

Oscillating: The aboutness of pages changes between on-topic and off-topic more than once (Figure 71(d)):

$\exists t$ where$(URI\text{-}R@t) \quad \not\approx \quad aboutness(URI\text{-}R@t+i)$ and $aboutness(URI\text{-}R@t) \quad \approx \quad aboutness(URI\text{-}R@t-j)$ where $i, j \geq 0$ and i, j are integers.

We found that 9-13% of the TimeMaps are Oscillating between on-topic and off-topic.

Always Off: This is the most challenging case, where all the mementos are off-topic (Figure 71(e)):

$\forall\ t$, $URI\text{-}R@t$ is *not* relevant to C.

We manually identified these cases (totaling 3 seed URIs) and excluded these from the gold standard dataset. This situation can arise if seed URIs were included by accident, or if their content changed (e.g., site shutdown) in the interval between when the seed URI was identified and when the crawling began.

8.4 RESEARCH APPROACH

In this section, we explain the methodology for preparing the dataset and then the methodology for applying different measures to detect the off-topic pages.

8.4.1 DATASET PREPROCESSING

We applied the following steps to prepare the gold standard dataset:

1. Obtain the seed list of URIs from the front-end interface of Archive-It.

2. Obtain the TimeMaps of the seed URIs from the CDX file[1].

3. Extract the HTML of the mementos from the WARC files (locally hosted at ODU).

[1] `http://archive.org/web/researcher/cdx_file_format.php`

4. Extract the text of the page using the Boilerpipe library [184].
5. Extract terms from the page, using scikit-learn [257] to tokenize, remove stop words, and apply stemming.

8.4.2 METHODS FOR DETECTING OFF-TOPIC PAGES

In this section, we use different similarity measures between pages to detect when the $aboutness(URI\text{-}R)$ over time changes and to define a threshold that separates the on-topic and the off-topic pages.

Cosine similarity

Cosine similarity [217] is one of the most commonly used similarity measures to solve different problems in IR and text mining, such as text classification and categorization, question answering, document filtering, etc. Cosine similarity measures the cosine of the angle between two vectors (d_1 and d_2) by taking the dot product between them [292, 280]:

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\| d_1 \| \| d_2 \|} \tag{10}$$

After text preprocessing, we calculated the TF-IDF for mementos, then we applied cosine similarity to compare the $aboutness(URI\text{-}R@t_0)$ with $aboutness(URI\text{-}R@t)$ by calculating the similarity between the mementos.

Jaccard similarity coefficient

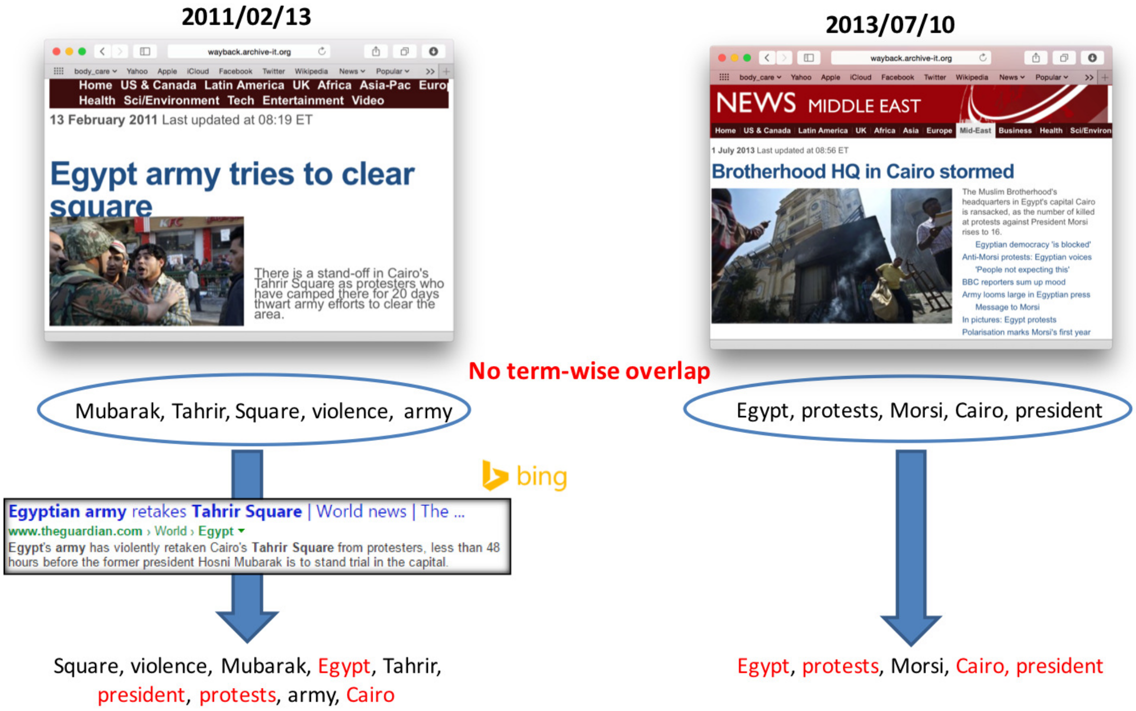
The Jaccard similarity coefficient measure is the size of the intersection of two sets divided by the size of their union [217]. The Jaccard between set A and set B is formulated as following:

$$J(A, B) = \frac{A \cap B}{A \cup B} \tag{11}$$

After preprocessing the text (result from step 5), we apply the Jaccard coefficient on the resulting terms to specify the similarity between the $URI\text{-}R@t$ and $URI\text{-}R@t_0$.

Intersection of the most frequent terms

Term frequency (TF) refers to how often a term appears in a document. The aboutness of a document can be represented using the top-k most frequent terms. After text extraction, we calculated the TF of the text $URI\text{-}R@t$, and then compared the top 20 most frequent terms of the $URI\text{-}R@t$ with the top 20 most frequent terms of the $URI\text{-}R@t_0$. The size


2011/02/13
wayback.archive-it.org
Home US & Canada Latin America UK Africa Asia-Pac Euro
Health Sci/Environment Tech Entertainment Video
2013/07/10
wayback.archive-it.org
NEWS MIDDLE EAST
Home US & Canada Latin America UK Africa Asia Europe Mid-East Business Health Sci/Environ
No term-wise overlap
Mubarak, Tahrir, Square, violence, army
Egypt, protests, Morsi, Cairo, president
bing
Square, violence, Mubarak, Egypt, Tahrir, president, protests, army, Cairo
Egypt, protests, Morsi, Cairo, president

13 February 2011 Last updated at 08:19 ET

Egypt army tries to clear square

There is a stand-off in Cairo's Tahrir Square as protesters who have camped there for 20 days thwart army efforts to clear the area.

1 July 2013 Last updated at 08:56 ET

Brotherhood HQ in Cairo stormed

The Muslim Brotherhood's headquarters in Egypt's capital Cairo is ransacked, as the number of killed at protests against President Morsi rises to 16.

Egyptian democracy 'is blocked'
Anti-Morsi protests: Egyptian voices
'People not expecting this'
BBC reporters sum up mood
Army looms large in Egyptian press
Message to Morsi
In pictures: Egypt protests
Polarisation marks Morsi's first year

Egyptian army retakes Tahrir Square | World news | The ...
www.theguardian.com › World › Egypt
Egypt's army has violently retaken Cairo's Tahrir Square from protesters, less than 48 hours before the former president Hosni Mubarak is to stand trial in the capital.

FIG. 72: An example for increasing the semantic context by the Web based kernel function using a search engine (SE).

of the intersection between the top 20 terms of $URI\text{-}R@t$ and $URI\text{-}R@t_0$ represents the similarity between the mementos. We name this method TF-Intersection.

Web-based kernel function

The previous methods are term-wise similarity measures, i.e., they use lexicographic term matching. But these methods may not suitable for archived collections with a large time span or pages that contain a small amount of text. For example, the Egyptian Revolution collection is from February 2011 until April 2013. Suppose a page in February 2011 has terms like “Mubarak, Tahrir, Square, violence, army” and a page in April 2013 has terms like “Egypt, protests, Morsi, Cairo, president”. The two pages are semantically relevant to each other, but term-wise the previous methods might not detect them as relevant. With a large evolution of pages through a long period of time, we need a method that focuses on the semantic context of the documents.

The work by Sahami and Heilman [274] inspired us to augment the text of $URI\text{-}R@t_0$ with additional terms from the Web using a search engine to increase its semantic context. This approach is based on query expansion techniques [60], which have been well-studied in the Information Retrieval field. We used the contextually descriptive snippet text that was returned with search engine results, which we call "SEKernel". Snippet text has been shown to be a good source for query expansion terms [343]. Snippet text has shown effectiveness in representing the documents. We used the returned results from the Bing Search API.

We augment the terms of $URI\text{-}R@t_0$ with semantic context from the search engine as follows:

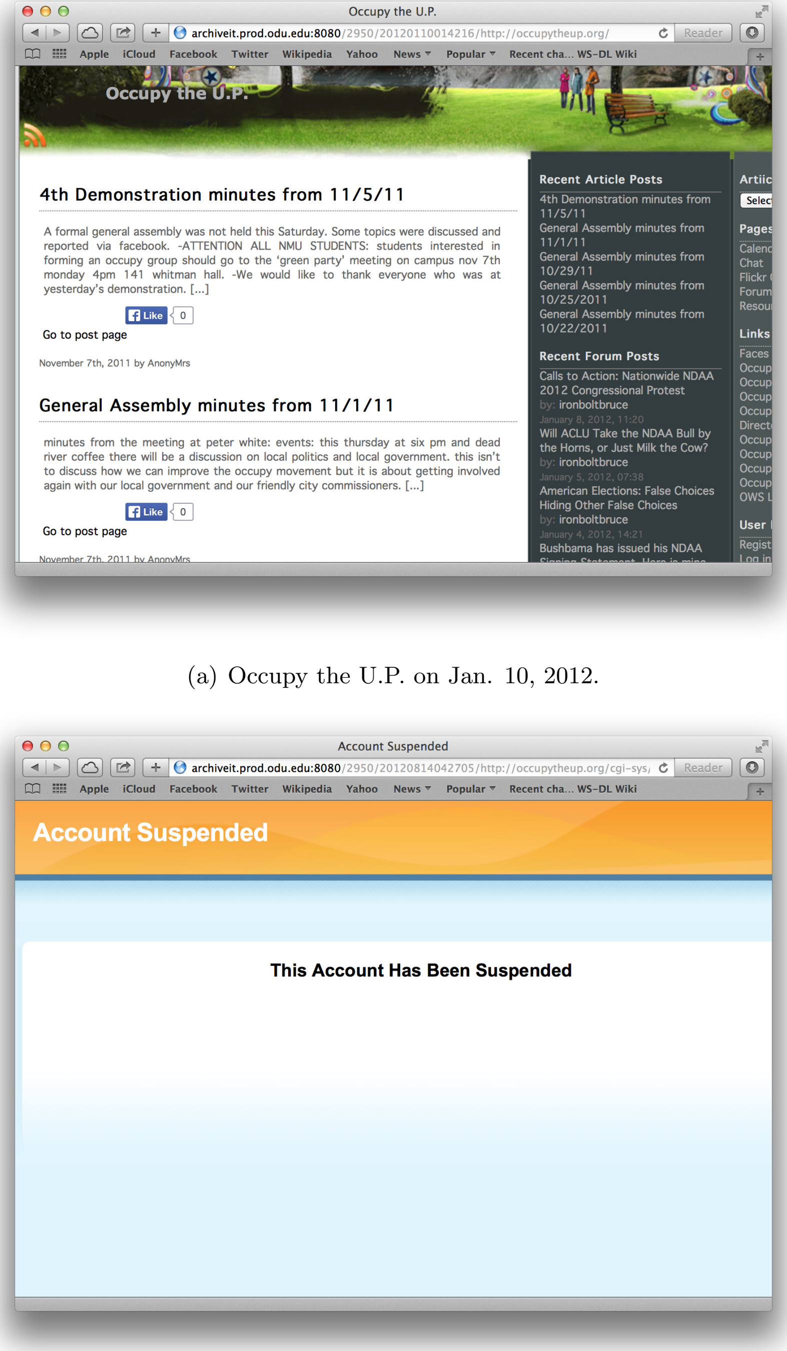
1. Format a query q from the top five words x of the first memento ($URI\text{-}R@t_0$).
2. Issue $q(x)$ to the search engine SE.
3. Extract the terms p from the top 10 snippets returned for $q(x)$.
4. Add the terms of the snippets p to the terms of the original text of the first memento d to have a new list of terms, $ST = p \cup d$.
5. $\forall t$, calculate the Jaccard coefficient between ST (the expanded aboutness of the $URI\text{-}R@t_0$) and the terms of $URI\text{-}R@t$, where $t \geq 1$.

Figure 72 shows an example of how we apply the Web-based kernel function on a memento from the Egyptian Revolution collection. As the figure illustrate, we use terms "Mubarak, Tahrir, Square, violence, army" of the first memento as search keywords to generate semantic context. The resulting snippet will have new terms like "Egypt, President, Cairo, protests", which term-wise overlaps with the page that contains "Egypt, protests, Morsi, Cairo, president". The resulting similarity between the two mementos in Figure 72 after extending the terms of the first memento is 0.4.

Change in size

We noticed that the sizes of off-topic mementos are often much smaller in size than the on-topic mementos. We used the relative change in size to detect when the page goes off-topic. The relative change of the page size can be represented by the content length or the total number of words (e.g., egypt, egypt, tahrir, the, square) in the page. For example, assume $URI\text{-}R@t_0$ contains 100 words and $URI\text{-}R@t$ contains 5 words. This represents a 95% decrease in the number of words between $URI\text{-}R@t_0$ and $URI\text{-}R@t$. The change in size, denoted $d(A, B)$, can be defined formally as following:

$$d(A,B) = 1 - \frac{s(A)}{s(B)}, \qquad \text{where s is the size of document.} \quad (12)$$



4th Demonstration minutes from 11/5/11

A formal general assembly was not held this Saturday. Some topics were discussed and reported via facebook. -ATTENTION ALL NMU STUDENTS: students interested in forming an occupy group should go to the 'green party' meeting on campus nov 7th monday 4pm 141 whitman hall. -We would like to thank everyone who was at yesterday's demonstration. [...]

Like 0

Go to post page

November 7th, 2011 by AnonyMrs

General Assembly minutes from 11/1/11

minutes from the meeting at peter white: events: this thursday at six pm and dead river coffee there will be a discussion on local politics and local government. this isn't to discuss how we can improve the occupy movement but it is about getting involved again with our local government and our friendly city commissioners. [...]

Like 0

Go to post page

November 7th, 2011 by AnonyMrs

Recent Article Posts

4th Demonstration minutes from 11/5/11
General Assembly minutes from 11/1/11
General Assembly minutes from 10/29/11
General Assembly minutes from 10/25/2011
General Assembly minutes from 10/22/2011

Recent Forum Posts

Calls to Action: Nationwide NDAA 2012 Congressional Protest
by: ironboltbruce
January 8, 2012, 11:20

Will ACLU Take the NDAA Bull by the Horns, or Just Milk the Cow?
by: ironboltbruce
January 5, 2012, 07:38

American Elections: False Choices Hiding Other False Choices
by: ironboltbruce
January 4, 2012, 14:21

Bushbama has issued his NDAA

(a) Occupy the U.P. on Jan. 10, 2012.

Account Suspended

This Account Has Been Suspended

(b) Expired on August 14, 2012, but no textual content.

FIG. 73: Later versions of occupytheup.org are off-topic.

TABLE 22: The results of evaluating the similarity approaches averaged on three collections.

Similarity Measure	**Threshold**	**FP**	**FN**	**FP+FN**	**ACC**	F_1	**AUC**
Cosine	0.15	31	22	53	**0.983**	**0.881**	**0.961**
WordCount	−0.85	6	44	50	0.982	0.806	0.870
SEKernel	0.05	64	83	147	0.965	0.683	0.865
Bytes	−0.65	28	133	161	0.962	0.584	0.746
Jaccard	0.05	74	86	159	0.962	0.538	0.809
TF-Intersection	0.00	49	104	153	0.967	0.537	0.740

TABLE 23: The results of the best three combined methods approaches averaged on three collections.

Similarity Measure	**Threshold**	**FP**	**FN**	**FP+FN**	**ACC**	F_1	**AUC**
(Cosine, WordCount)	(0.10, −0.85)	24	10	34	**0.987**	**0.906**	**0.968**
(Cosine, SEKernel)	(0.10, 0.00)	6	35	40	0.990	0.901	0.934
(WordCount, SEKernel)	(−0.80, 0.00)	14	27	42	0.985	0.818	0.885

We tried two methods for measuring the change in size: the content length (bytes) and the number of words (WordCount). Although using the content length, which can be extracted directly from the headers of the WARC files, saves the steps of extracting the text and tokenization, it fails to detect when the page goes off-topic in the case when the page has little to no textual content but the page template is still large. For example, the Facebook page in Figure 69 went off-topic in Figure 69(b) and has 62KB, but the on-topic page in Figure 69(a) is nearly similarly sized with 84KB. Using a significant decrease in byte size allows for rapid detection of potential off-topic pages.

There are many cases where the page goes off-topic and the size of the page decreases or in some cases reaches 0 bytes, e.g., the account is suspended, transient errors, or no content in the page. One of the advantages of using the structural-based methods over the textual-based methods is that structural-based methods are language independent. Many of the collections are multi-lingual, and each language needs special processing. The structural methods are suitable for those collections. Figure 73 has an example where the account is suspended and the size of the page is almost 0 bytes.

8.5 EVALUATION

In this section, we define how we evaluate the methods presented in Section 8.4.2 on our gold standard dataset. Based on these results, we define a threshold th for each method for when a memento becomes off-topic.

8.5.1 EVALUATION METRICS

We used multiple metrics to evaluate the performance of the similarity measures:

- False positives (FP), the number of on-topic pages that are predicted as off-topic.
- False negatives (FN), the number of off-topic pages that are predicted as on-topic.
- Accuracy (ACC), the fraction of the classifications that are correct.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{13}$$

  TP is the number of True Positives (off-topic pages that are predicted as off-topic) and TN is the number of True Negatives (on-topic pages that are predicted as on-topic).

- F_1 score (also known as F-measure or the harmonic mean), the weighted average of precision and recall.

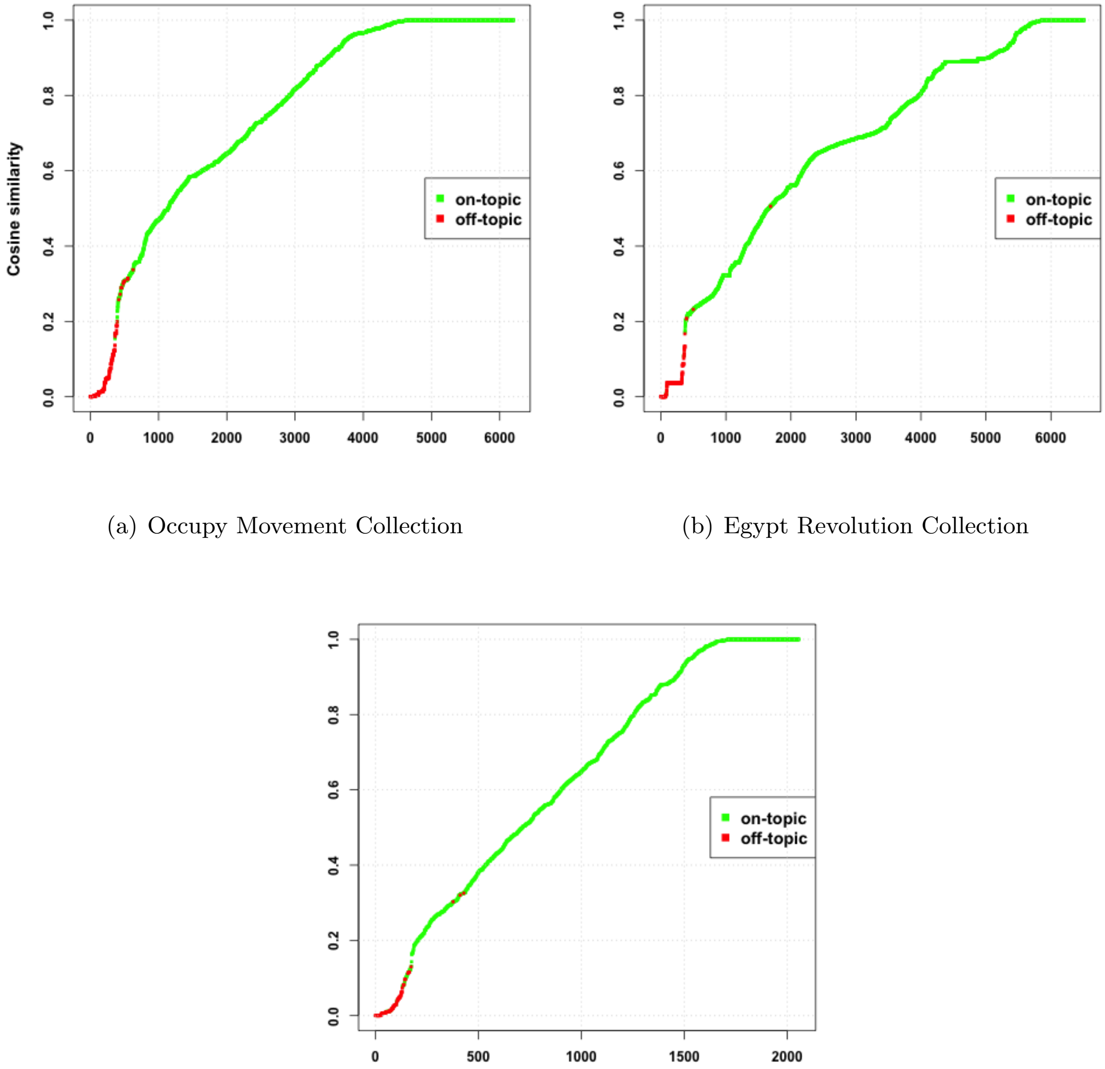
$$F_1 = \frac{2TP}{(2TP + FP + FN)} \tag{14}$$

- The ROC AUC score, a single number that computes the area under the receiver operating characteristic (ROC) [101] curve, which is also denoted as AUC.

8.5.2 RESULTS

We tested each method with 21 thresholds (378 tests for three collections) on our gold standard dataset to estimate which threshold for each method is able to separate the off-topic from the on-topic pages. In order to determine the best threshold, we used the evaluation metrics described in the previous section, and averaged the results based on the F_1 of the three collections at different thresholds. To say that $URI\text{-}R@t$ is off-topic at $th = 0.15$ means that the similarity between $URI\text{-}R@t$ and $URI\text{-}R@t_0$ is < 0.15. On-topic means the similarity between $URI\text{-}R@t$ and $URI\text{-}R@t_0$ is ≥ 0.15.

For each similarity measure, there is an upper bound and lower bound for the value of similarity. For Cosine, TF-Intersection, Jaccard, and SEKernel, the highest value is at 1 and the lowest value is at 0. A similarity of 1 represents a perfect similarity, and 0

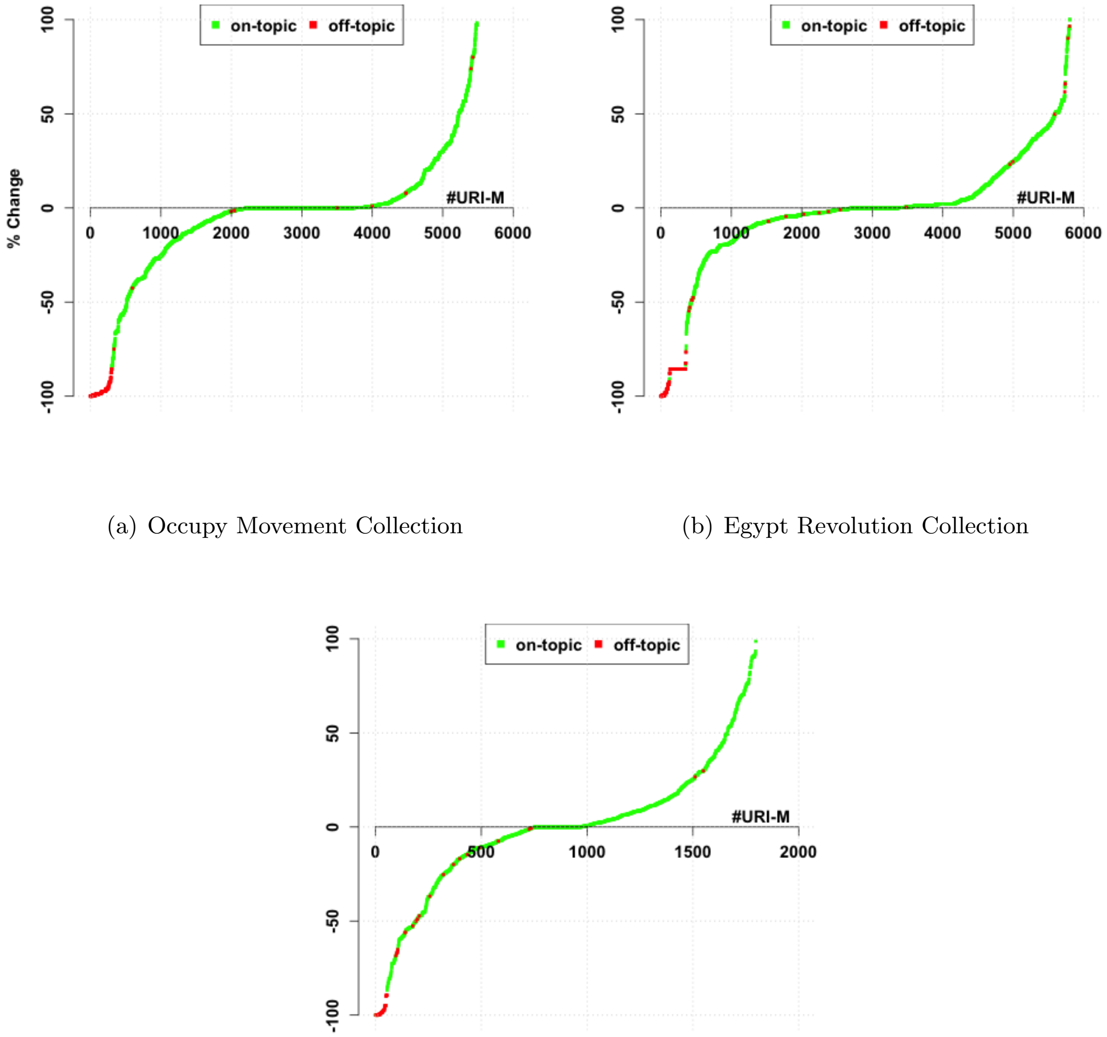


(a) Occupy Movement Collection

(b) Egypt Revolution Collection

(c) Human Rights Collection

FIG. 74: How cosine similarity separates the off-topic from the on-topic pages.



(a) Occupy Movement Collection

(b) Egypt Revolution Collection

(c) Human Rights Collection

FIG. 75: How change of page size (based on word count) separates the off-topic from the on-topic pages.

similarity represents that there is no similarity between the pages. The word count and content length measures can be from -1 to $+1$. The negative values in the change of size measures represent the decrease in size, so -1 means the page has a 100% decrease from $URI\text{-}R@t_0$. When the change in size is 0 that means there is no change in the size of the page. We assume that a large decrease in size between $URI\text{-}R@t$ and $URI\text{-}R@t_0$ indicates that the page might be off-topic. Therefore, if the change in size between $URI\text{-}R@t$ and $URI\text{-}R@t_0$ is a 95% decrease in the size, that means $URI\text{-}R@t$ is off-topic at $th = -0.95$.

Table 22 contains the summary of running the similarity approaches on the three collections. The table shows the best result based on the F_1 score at the underlying threshold measures averaged on all three collections. From the table, the best performing measure is Cosine with average $ACC = 0.983$, $F_1 = 0.881$, and $AUC = 0.961$, followed by WordCount. Using SEKernel performs better than TF-Intersection and Jaccard. Based on the F_1 score, we notice that TF-Intersection and Jaccard similarity are the least effective methods.

Figure 74 shows how Cosine separates the off-topic from the on-topic pages for each collection. It shows that that the off-topic pages are concentrated near 0.0-0.2 similarity and there is no FNs past $th = 0.4$. Figure 75 shows how WordCount identifies on-topic and off-topic mementos at different thresholds. We see from the figure that there are no on-topic pages near 100% decrease (i.e., -100% change), while the majority of the off-topic mementos are concentrated near the 80-100% decrease (i.e., $-(80\text{-}100)\%$ change).

There was consistency among the best-performing values for TF-Intersection, Jaccard, and SEKernel methods over the three collections. For example, for all collections the best performance of the SEKernel method is at $th = 0.05$. However, there was inconsistency among the values of th with the best performance for each collection for Cosine, WordCount, and Bytes measures. For the methods with inconsistent threshold values, we averaged the best thresholds of each collection. For example, the best th values of Cosine for the Occupy Movement collection, Egypt Revolution collection, and Human Rights collection are 0.2, 0.15, 0.1 respectively.

We took the average of the three collections at $th = 0.2$, $th = 0.15$, and $th = 0.1$, then based on the best F_1 score, we specified the threshold that has the best average performance, which is $th = 0.15$. Specifying a threshold for detecting the off-topic pages from archived pages is not easy with the differences in the nature of the collections. For example, long-running collections such as the Human Rights collection (2009-present) have more opportunities for some pages to change dramatically, while staying relevant to the collection. There is more research to be done in exploring the thresholds and methods. We plan to investigate different methods on larger sets of labeled collections, so that we can specify the features that affect choosing the value of the threshold.

TABLE 24: The results of evaluating Archive-It collections through the assessment of the detected off-topic pages using (Cosine, WordCount) methods at th = (0.10, −0.85). Numbers in parenthesis are the total URI-Ms and URI-Rs for the collection.

Collection	ID	Timespan	Off-topic URI-Ms	Affected URI-Rs	TP	FP	P
Global Food Crisis	2893	2011/10/19-2012/10/24	22(3,063)	7(65)	22	0	1.00
Government in Alaska	1084	2006/12/01-2013/04/13	16(506)	4(68)	16	0	1.00
Virginia Tech Shootings	2966	2011/12/08-2012/01/03	24(1,670)	2(239)	24	0	1.00
Wikileaks 2010 Document Release Collection	2017	2010/07/27-2012/08/27	107(2,360)	8(35)	107	0	1.00
DIBAM	1019	2008/02/22-2008/03/24	4(106)	1(25)	4	0	1.00
Global Health Events	4887	2014/10/01-2015/10/21	56(3,518)	8(165)	53	3	0.95
2003 California Recall Election	5947	2003/09/24-2003/12/12	270(2,312)	36(178)	254	16	0.94
Jasmine Revolution - Tunisia 2011	2323	2011/01/19-2012/12/24	114(4,076)	31(231)	107	7	0.94
Academics at Baylor	3497	2013/01/28-2016/04/26	26(414)	13(232)	20	6	0.77
IT Historical Resource Sites	1827	2010/02/23-2012/10/04	59(10,283)	34(1,459)	45	14	0.76
Human Rights Documentation Initiative	1475	2009/04/29-2011/10/31	54(1,530)	20(147)	39	15	0.72
2007 Southern California Wildfires Web Archive	5810	2007/10/23-2007/11/02	335(2,416)	68(156)	215	120	0.64
Maryland State Document Collection	1826	2010/03/04-2012/12/03	0(184)	0(69)	-	-	-
April 16 Archive	694	2007/05/23-2008/04/28	0(118)	0(35)	-	-	-
Brazilian School Shooting	2535	2011/04/09-2011/04/14	0(1,092)	0(476)	-	-	-
Russia Plane Crash Sept 7,2011	2823	2011/09/08-2011/09/15	0(447)	0(65)	-	-	-
Burke Library New York City Religions 340	1945	2011/11/16-2013/02/11	0(208)	0(107)	-	-	-
Hurricane Irene (Aug 2011)	2816	2011/09/02-2011/09/26	0(102)	0(71)	-	-	-

8.5.3 COMBINING THE SIMILARITY MEASURES

We tested 6,615 pairwise combinations (15 method combinations × 21 × 21 threshold values). A page was considered off-topic if either of the two methods declared it off-topic. Performance results of combining the similarity approaches are presented in Table 23. We present the three best average combinations of the similarity measures based on the F_1 score and the AUC. The performance increases with combining Cosine and WordCount (Cosine, WordCount) at th $= (0.1, -0.85)$. There is a 36% decrease in errors (FP+FN) as compared to the best performing single measure, Cosine. Furthermore, (Cosine, WordCount) has a 3% increase in the F_1 score over Cosine. (Cosine, SEKernel) at th $= (0.1, 0.0)$ has 2% increase in F_1 over Cosine. (WordCount, SEKernel) at th $= (-0.80, 0.00)$ has lower performance than Cosine.

In summary, (Cosine, WordCount) gives the best performance at th $= (0.1, -0.85)$ across all the single and combined methods. Moreover, combining WordCount with Cosine does not cause much overhead in processing because WordCount uses tokenized words and needs no extra text processing.

8.6 EVALUATING ARCHIVE-IT COLLECTIONS

We used the best performing method (Cosine, WordCount) on the labeled dataset with the suggested thresholds $(0.10, -0.85)$ and applied them on unlabeled Archive-It collections. We chose different types of collections, e.g., governmental collections (Maryland State Document Collection, Government in Alaska), event-based collections (Jasmine Revolution - Tunisia 2011, Virginia Tech Shootings, Global Health Events (the 2014 Ebola Outbreak)), and theme-based collections (Wikileaks 2010 Document Release Collection, Human Rights Documentation Initiative, Burke Library New York City Religions). Table 24 contains the details of the 18 tested collections, such as the collection's ID, timespan, etc. that comprise 4,019 URI-Rs and 36,785 URI-Ms. We extracted the tested collections from the ODU mirror of Archive-It's collections, except for the Global Health Events Collection[2], the 2007 Southern California Wildfires Web Archive[3], the Academics at Baylor[4], and the 2003 California Recall Election[5], which we recently obtained from Archive-It.

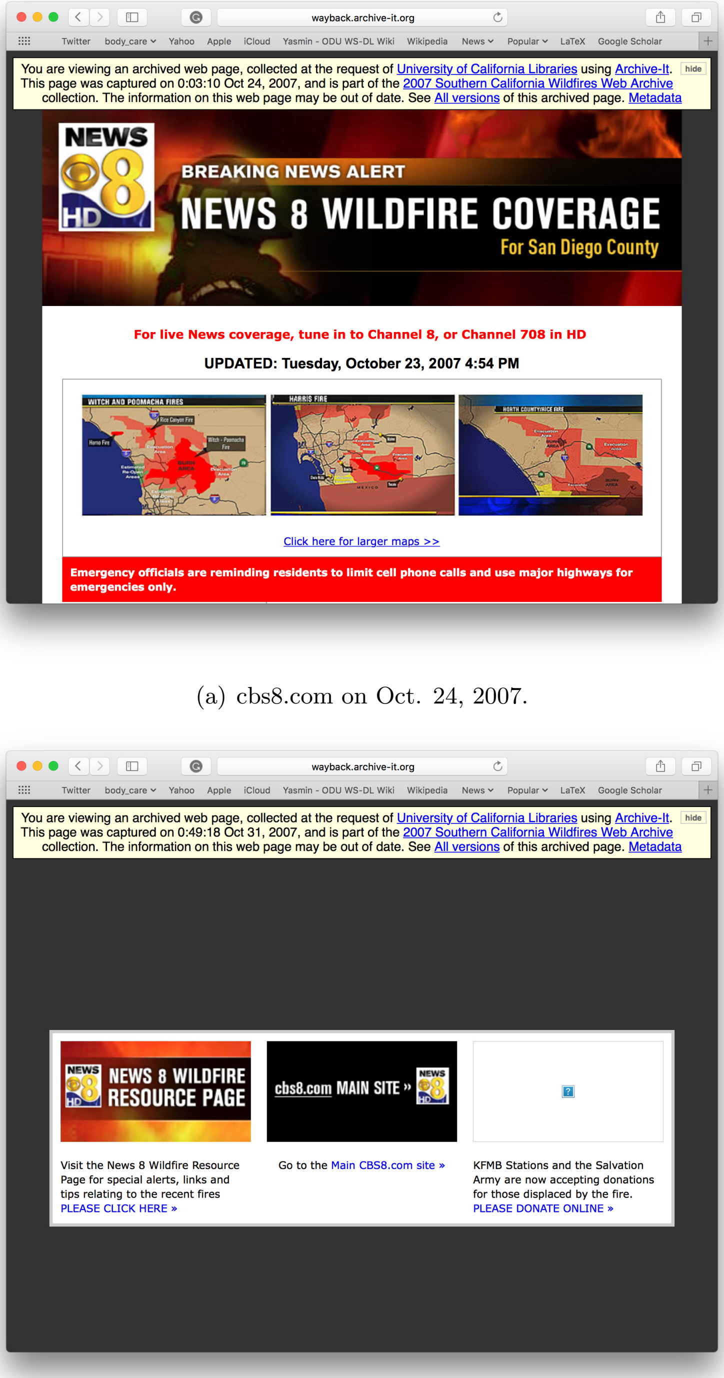
The results of evaluating (Cosine, WordCount) at th $= (0.10, -0.85)$ are shown in Table 24. The table contains the number of affected URI-Rs in each collection. For the reported results, we manually assessed the FP and TP of each TimeMap and then calculated the

[2] `https://archive-it.org/collections/4887/`

[3] `https://archive-it.org/collections/5810/`

[4] `https://archive-it.org/collections/3497/`

[5] `https://archive-it.org/collections/5947/`



You are viewing an archived web page, collected at the request of University of California Libraries using Archive-It. This page was captured on 0:03:10 Oct 24, 2007, and is part of the 2007 Southern California Wildfires Web Archive collection. The information on this web page may be out of date. See All versions of this archived page. Metadata

BREAKING NEWS ALERT

NEWS 8 WILDFIRE COVERAGE

For San Diego County

For live News coverage, tune in to Channel 8, or Channel 708 in HD

UPDATED: Tuesday, October 23, 2007 4:54 PM

Click here for larger maps >>

Emergency officials are reminding residents to limit cell phone calls and use major highways for emergencies only.

(a) cbs8.com on Oct. 24, 2007.

You are viewing an archived web page, collected at the request of University of California Libraries using Archive-It. This page was captured on 0:49:18 Oct 31, 2007, and is part of the 2007 Southern California Wildfires Web Archive collection. The information on this web page may be out of date. See All versions of this archived page. Metadata

NEWS 8 WILDFIRE RESOURCE PAGE

cbs8.com MAIN SITE »

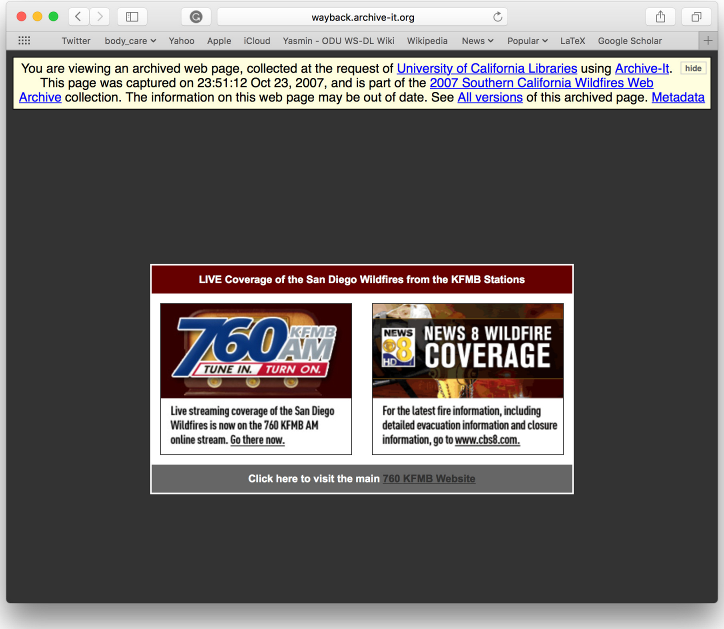
Visit the News 8 Wildfire Resource Page for special alerts, links and tips relating to the recent fires PLEASE CLICK HERE »

Go to the Main CBS8.com site »

KFMB Stations and the Salvation Army are now accepting donations for those displaced by the fire. PLEASE DONATE ONLINE »

(b) cbs8.com on Oct. 31, 2007.

FIG. 76: Example of a significant change in cbs8.com: from Oct. 24, 2007 to Oct. 31, 2007.



You are viewing an archived web page, collected at the request of University of California Libraries using Archive-It. This page was captured on 23:51:12 Oct 23, 2007, and is part of the 2007 Southern California Wildfires Web Archive collection. The information on this web page may be out of date. See All versions of this archived page. Metadata

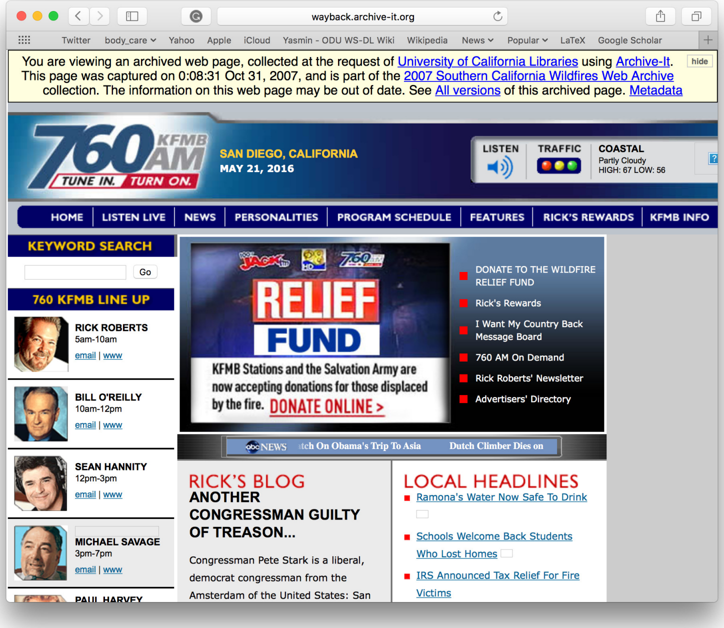
LIVE Coverage of the San Diego Wildfires from the KFMB Stations

Live streaming coverage of the San Diego Wildfires is now on the 760 KFMB AM online stream. Go there now.

For the latest fire information, including detailed evacuation information and closure information, go to www.cbs8.com.

Click here to visit the main 760 KFMB Website

(a) 760kfmb.com on Oct. 23, 2007.



You are viewing an archived web page, collected at the request of University of California Libraries using Archive-It. This page was captured on 0:08:31 Oct 31, 2007, and is part of the 2007 Southern California Wildfires Web Archive collection. The information on this web page may be out of date. See All versions of this archived page. Metadata

SAN DIEGO, CALIFORNIA
MAY 21, 2016

COASTAL
Partly Cloudy
HIGH: 67 LOW: 56

KEYWORD SEARCH

760 KFMB LINE UP

RICK ROBERTS
5am-10am
email | www

BILL O'REILLY
10am-12pm
email | www

SEAN HANNITY
12pm-3pm
email | www

MICHAEL SAVAGE
3pm-7pm
email | www

PAUL HARVEY

KFMB Stations and the Salvation Army are now accepting donations for those displaced by the fire. DONATE ONLINE >

- DONATE TO THE WILDFIRE RELIEF FUND
- Rick's Rewards
- I Want My Country Back Message Board
- 760 AM On Demand
- Rick Roberts' Newsletter
- Advertisers' Directory

tch On Obama's Trip To Asia Dutch Climber Dies on

RICK'S BLOG
ANOTHER CONGRESSMAN GUILTY OF TREASON...

Congressman Pete Stark is a liberal, democrat congressman from the Amsterdam of the United States: San

LOCAL HEADLINES
- Ramona's Water Now Safe To Drink
- Schools Welcome Back Students Who Lost Homes
- IRS Announced Tax Relief For Fire Victims

(b) 760kfmb.com on Oct. 31, 2007.

FIG. 77: Example of a significant change in 760kfmb.com: from Oct. 23, 2007 to Oct. 31, 2007.

precision $P = TP/(TP + FP)$ for each collection. We cannot compute recall since we cannot know how many off-topic mementos were not detected (FN). The precision is near 1.0 for eight collections. $P = 0.72$ for the "Human Rights Documentation" collection, with 15 FPs. Those 15 URI-Ms affected three TimeMaps. An example of an affected TimeMap (`https://wayback.archive-it.org/1475/*/http://www.fafg.org/`) contains 12 FPs. The reason is that the home page of the site changed and the new versions use Adobe Flash. The 14 FPs from the "IT Historical Resource Sites" collection affected 5 URI-Ts. The content of these 5 pages changed dramatically through time, resulting in FPs. The 2007 Southern California Wildfires Web Archive has 44% (68 out of 156) of its TimeMaps affected with off-topic pages. By assessing the detected off-topic pages from this collection, we found that $P = 0.64$, with 120 FPs that affected only 5 URI-Ts because of a significant change in the content of these pages through time. The two pages that dominated the FPs with 88% are shown in Figures 76 and 77.

There are six collections that have no reported off-topic pages. Two of these collections, the Brazilian School Shooting and the Russia Plane Crash, span less than a week, which is typically not enough time for pages to go off-topic. The other collections with no detected off-topic mementos are the Maryland State Document, the April 16 Archive, the Hurricane Irene, and the Burke Library New York City Religions. Perhaps these collections simply had well-behaved URIs.

8.7 SUMMARY

We presented different approaches for assisting curators in identifying off-topic mementos in the archive [20, 22]. We investigated six methods for measuring similarity between pages: cosine similarity, Jaccard similarity, intersection of the most 20 frequent terms, Web-based kernel function, change in number of words, and change in content length. We tested the approaches on three different labeled subsets of collections from Archive-It. We found that of the single methods, the cosine similarity measure is the most effective method for detecting the off-topic pages at $th = 0.15$. The change in size based on the word count comes next at $th = -0.85$. We combined the suggested methods and found that, based on the F_1 score and the AUC, (Cosine, WordCount) at th = (0.1, −0.85) enhances the performance to have the highest F_1 score at 0.9 and the highest AUC at 0.9.

We tested the performance of (Cosine, WordCount) at th = (0.1, −0.85) by applying them on 18 Archive-It collections. We manually assessed the relevancy of the detected off-topic pages. In summary, the suggested approach, (Cosine, WordCount) at th = (0.1, −0.85), has shown good results at detecting the off-topic pages with 0.9 precision. The presented approaches will help curators to judge their crawls and also will prevent users from getting unexpected content when they access archived pages. Besides optimizing the quality of the archived collections, detecting the off-topic pages automatically will help in optimizing storage space and the time required for manual off-topic detection. Furthermore, flagging the off-topic pages will be useful for the quality of the automatically generated stories in the DSA framework and other applications such as thumbnail generation. We generated a gold standard dataset of labeled mementos that is available at `https://github.com/yasmina85/OffTopic-Detection` along with the off-topic detection source code. We are contributing this manually labeled gold standard set to the community for use in future research.

We also identified five different behaviors of changing the aboutness of TimeMaps: Always On, Step Function On, Step Function Off, Oscillating, and Always Off. The ideal behavior for curators is "Always On", in which the pages do not deviate from the theme of the collection. We found that 24% of the TimeMaps in our manually labeled sample had off-topic mementos. The majority of the affected TimeMaps are "Step Function On" and "Oscillating" with 8-13% of the TimeMaps. We found small number of TimeMaps that were "Always Off" or "Step Function Off". These behaviors will inform curators of the different cases of TimeMaps they may have in their collections. Furthermore, they inform us on the challenges of detecting the off-topic pages.

As the results of evaluating the presented approaches in this chapter suggest, (Cosine, WordCount) at th $= (0.1, -0.85)$ are the best performed methods combined. We adopt the two methods for excluding the off-topic pages from the pool of archived pages in Archive-It collections. In the next chapter, we will continue the other steps of selecting the best representative pages for generating a story.

CHAPTER 9

SELECTING REPRESENTATIVE PAGES FOR THE STORIES

The main research question we investigate in this chapter is: How to select k mementos that represent a story? The key element of this task is to evaluate and select the "best" representative k mementos, where k is much smaller than the number of mementos in the collection. Suggested values of k are determined by the results of the study in Chapter 6, and other tunable parameters will include the timeline of the desired story (which may exclude some portions of the collection), the percentage of damage of the memento (incomplete pages are not desirable candidates), the story type (cf. Table 4), etc.

To address our research question, we apply the following steps on an archived collection to reduce the candidate pool of mementos and then select representative mementos for the story (see Figure 59 in Chapter 5):

1. Eliminating the (near-)duplicate mementos: We exclude duplication in TimeMaps based on the duplicate elimination algorithm proposed in Section 9.1.

2. Excluding non-English language pages: We keep only mementos with English language content (Section 9.2).

3. Dynamic time slicing: Based on the dynamic slicing algorithm described in Section 9.3, we divide all the mementos in the collection into a dynamic number of slices that grows slowly based on the collection size.

4. Clustering mementos of each slice: Based on the content, we cluster the mementos of each slice (Section 9.4).

5. Selecting the best representative memento: We evaluate and select a memento from each cluster based on a set of quality metrics we proposed in Section 9.5

6. Chronological ordering: We specify the notions of time for the chosen mementos and extract their metadata to put them in chronological order for visualization (Section 9.6).

7. Visualizing the selected mementos: At the end, we use Storify for visualizing the generated story (Section 9.7).

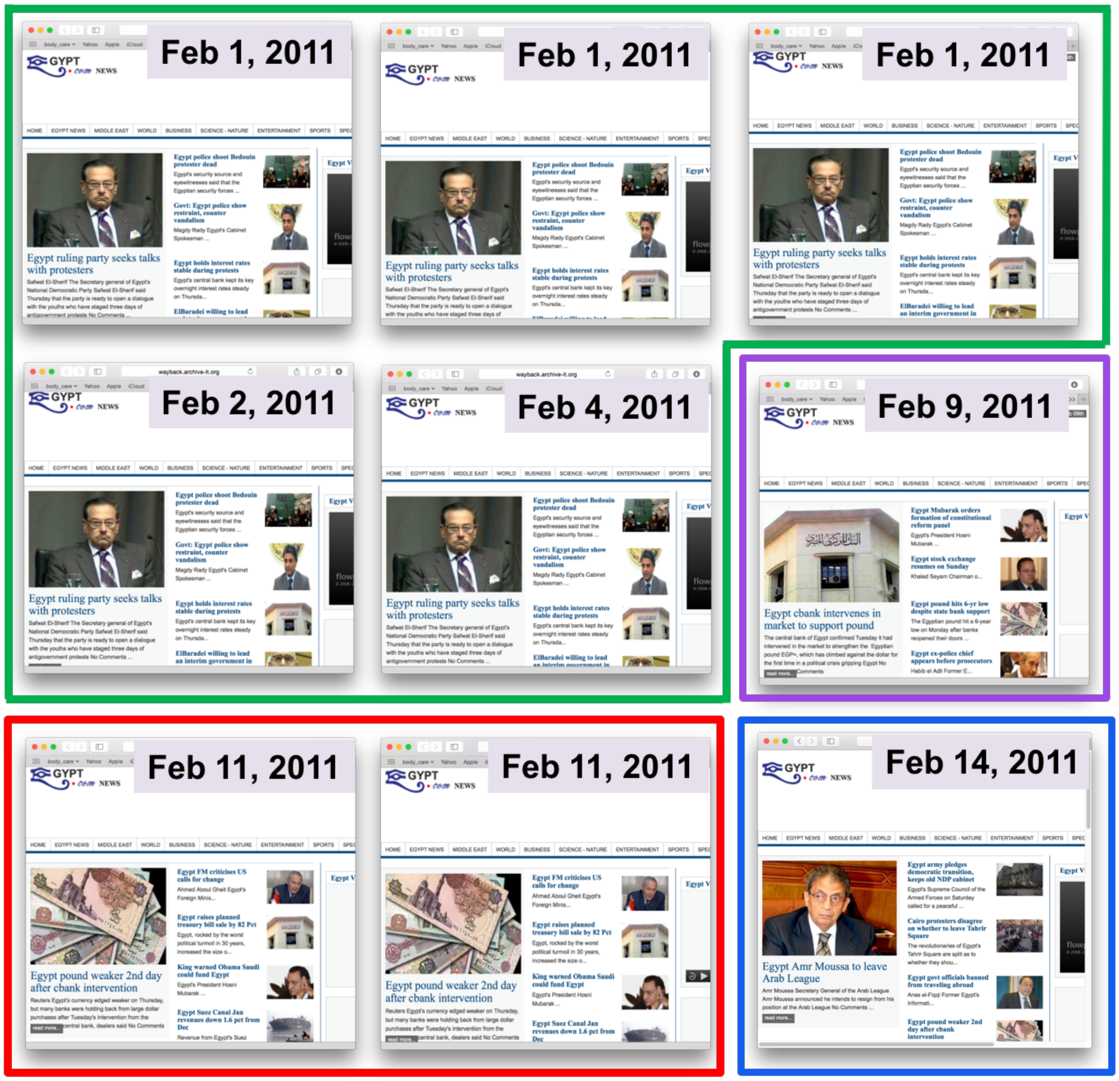


FIG. 78: Snapshots of mementos of news.egypt.com from the Egyptian Revolution collection that have duplication. Each group of similar mementos are grouped and annotated with the same color.

9.1 ELIMINATING (NEAR-)DUPLICATES IN WEB ARCHIVES

Archive-It crawlers grab periodic snapshots of the seed URI based on a predefined frequency set by the collection curator. This frequency may be daily, weekly, or even yearly. Due to the nature of Web evolution, some of these snapshots may change little or not at all.

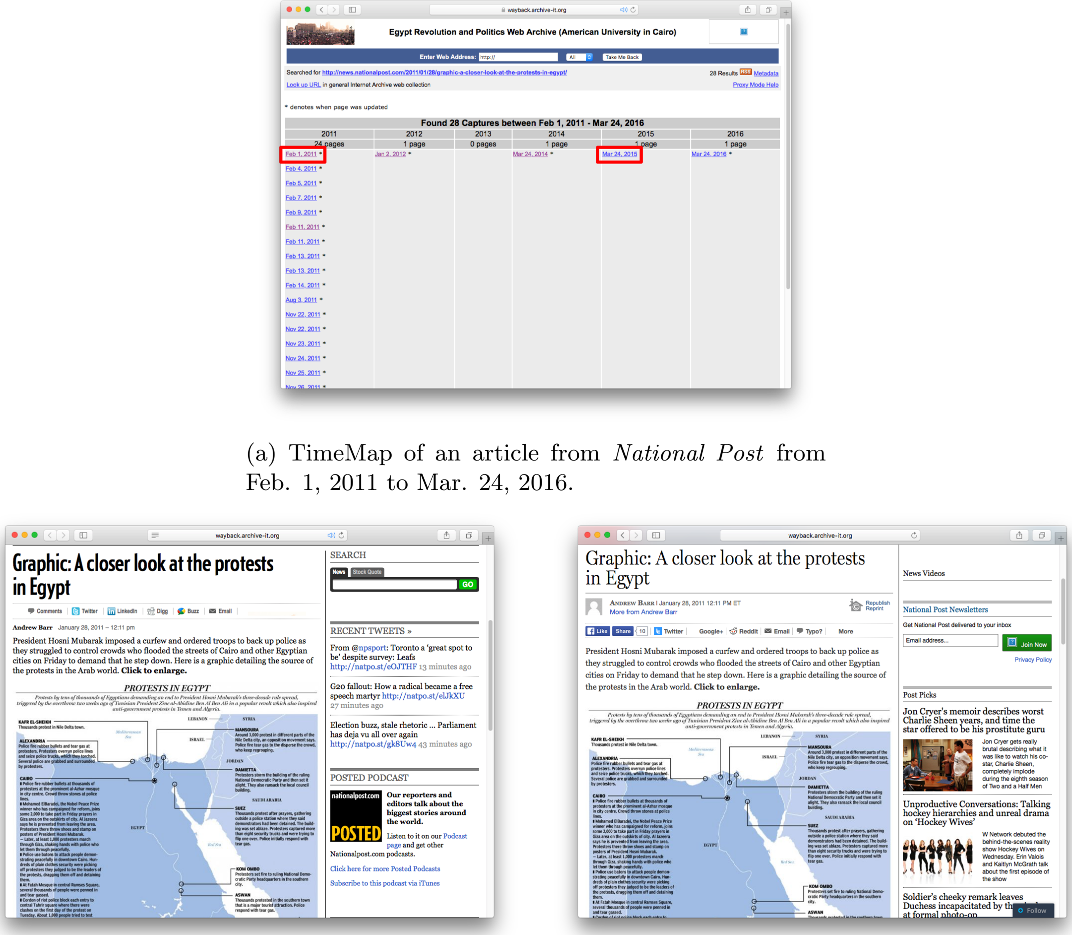
Algorithm 1 Eliminating (near-)duplicates in an individual TimeMap

```
$URI\text{-}T$ has n $URI\text{-}Ms$
$URI\text{-}T_{reduced} = \{\}$
$current = 0$
$next = 1$
Calculate the SimHash $S(URI\text{-}M_{current})$
while $next < n$ do
    Calculate the SimHash $S(URI\text{-}M_{next})$
    Compute Hamming Distance $HD$ between $S(URI\text{-}M_{current}), S(URI\text{-}M_{next})$
    if $HD(S(URI\text{-}M_{current}), S(URI\text{-}M_{next})) > \alpha$ then
        $URI\text{-}T_{reduced} = URI\text{-}T_{reduced} \cup URI - M_{next}$
        $current = next$
    end if
    $next = next + 1$
end while
```

Figure 78 shows part of a TimeMap for an Egyptian news Web site (`http://news.egypt.com/en/`) in the "Egypt Revolution and Politics" collection. The figure illustrates that there are duplicates in this TimeMap. Each group of similar mementos are grouped and annotated with the same color. Different colors reflect different clusters. The first memento, annotated by green, has four duplicates that are exactly the same. The sixth memento, annotated by purple, has no duplicates. The following two mementos are duplicates.

Figure 79(a) shows an example of near-duplication in a TimeMap[1] of a *National Post* article from Feb. 1, 2011 to Mar. 2, 2016 in the Egyptian Revolution collection. The mementos of Feb. 2, 2011 and Mar. 24, 2015 are annotated in red and shown in Figures 79(b) and 79(c). The two copies contain the same content except for the content of the sidebar on the right of the page, which contains recent news on the *National Post*. The article shown in Figure 79 is a good candidate for a story about the Egyptian Revolution, but it should only be considered once. The first memento (shown in Figure 79) will be considered because its Memento-Datetime is the closest to the publication date of the article (Jan. 28, 2011).

[1] `https://wayback.archive-it.org/2358/*/http://news.nationalpost.com/2011/01/28/graphic-a-closer-look-at-the-protests-in-egypt/`



wayback.archive-it.org

Egypt Revolution and Politics Web Archive (American University in Cairo)

Enter Web Address: http:// All Take Me Back

Searched for http://news.nationalpost.com/2011/01/28/graphic-a-closer-look-at-the-protests-in-egypt/ 28 Results RSS Metadata

Look up URL in general Internet Archive web collection Proxy Mode Help

* denotes when page was updated

Found 28 Captures between Feb 1, 2011 - Mar 24, 2016					
2011	2012	2013	2014	2015	2016
24 pages	1 page	0 pages	1 page	1 page	1 page
Feb 1, 2011 *	Jan 2, 2012 *		Mar 24, 2014 *	Mar 24, 2015	Mar 24, 2016 *
Feb 4, 2011 *					
Feb 5, 2011 *					
Feb 7, 2011 *					
Feb 9, 2011 *					
Feb 11, 2011 *					
Feb 11, 2011 *					
Feb 13, 2011 *					
Feb 13, 2011 *					
Feb 14, 2011 *					
Aug 3, 2011 *					
Nov 22, 2011 *					
Nov 22, 2011 *					
Nov 23, 2011 *					
Nov 24, 2011 *					
Nov 25, 2011 *					
Nov 26, 2011 *					

(a) TimeMap of an article from *National Post* from Feb. 1, 2011 to Mar. 24, 2016.

wayback.archive-it.org

Graphic: A closer look at the protests in Egypt

Comments | Twitter | LinkedIn | Digg | Buzz | Email |

Andrew Barr January 28, 2011 – 12:11 pm

President Hosni Mubarak imposed a curfew and ordered troops to back up police as they struggled to control crowds who flooded the streets of Cairo and other Egyptian cities on Friday to demand that he step down. Here is a graphic detailing the source of the protests in the Arab world. **Click to enlarge.**

PROTESTS IN EGYPT

Protests by tens of thousands of Egyptians demanding an end to President Hosni Mubarak's three-decade rule spread, triggered by the overthrow two weeks ago of Tunisian President Zine al-Abidine Ben Al Ben Ali in a popular revolt which also inspired anti-government protests in Yemen and Algeria.

KAFR EL-SHEIKH Thousands protest in Nile Delta town.

ALEXANDRIA Police fire rubber bullets and tear gas at protesters. Protesters overrun police lines and seize police trucks, which they torched. Several police are grabbed and surrounded by protesters.

CAIRO ▮ Police fire rubber bullets at thousands of protesters at the prominent al-Azhar mosque in city centre. Crowd throw stones at police lines. ▮ Mohamed ElBaradei, the Nobel Peace Prize winner who has campaigned for reform, joins some 2,000 to take part in Friday prayers in Giza area on the outskirts of city. Al Jazeera says he is prevented from leaving the area.

MANSOURA Around 3,000 protest in different parts of the Nile Delta city, an opposition movement says. Police fire tear gas to the disperse the crowd, who keep regrouping.

DAMIETTA Protesters storm the building of the ruling National Democratic Party and then set it alight. They also ransack the local council building.

SUEZ Thousands protest after prayers, gathering outside a police station where they said demonstrators had been detained.

SEARCH

News | Stock Quote | GO

RECENT TWEETS »

From @npsport: Toronto a 'great spot to be' despite survey: Leafs http://natpo.st/eOJTHF 13 minutes ago

G20 fallout: How a radical became a free speech martyr http://natpo.st/elJkXU 27 minutes ago

Election buzz, stale rhetoric ... Parliament has deja vu all over again http://natpo.st/gk8Uw4 43 minutes ago

POSTED PODCAST

nationalpost.com POSTED

Our reporters and editors talk about the biggest stories around the world.

Listen to it on our Podcast page and get other Nationalpost.com podcasts.

Click here for more Posted Podcasts

Subscribe to this podcast via iTunes

(b) Feb. 1, 2011 version.

wayback.archive-it.org

Graphic: A closer look at the protests in Egypt

Andrew Barr | January 28, 2011 12:11 PM ET
More from Andrew Barr

Republish Reprint

Like | Share | 10 | Twitter | Google+ | Reddit | Email | Typo? | More

President Hosni Mubarak imposed a curfew and ordered troops to back up police as they struggled to control crowds who flooded the streets of Cairo and other Egyptian cities on Friday to demand that he step down. Here is a graphic detailing the source of the protests in the Arab world. **Click to enlarge.**

PROTESTS IN EGYPT

News Videos

National Post Newsletters

Get National Post delivered to your inbox

Email address... Join Now

Privacy Policy

Post Picks

Jon Cryer's memoir describes worst Charlie Sheen years, and time the star offered to be his prostitute guru

Jon Cryer gets really brutal describing what it was like to watch his co-star, Charlie Sheen, completely implode during the eighth season of Two and a Half Men

Unproductive Conversations: Talking hockey hierarchies and unreal drama on 'Hockey Wives'

W Network debuted the behind-the-scenes reality show Hockey Wives on Wednesday. Erin Valois and Kaitlyn McGrath talk about the first episode of the show

Soldier's cheeky remark leaves Duchess incapacitated by th at formal photo-op

Follow

(c) Mar. 24, 2015 version.

FIG. 79: Example of duplicate in a TimeMap.

There have been several methods for calculating the similarity between Web pages [52, 216, 259, 193, 262]. We used 64-bit SimHash fingerprints with $k = 4$ to calculate the (near-)duplicates between Web pages in individual TimeMaps after excluding the text from the HTML. We propose Algorithm alg:duplicates for eliminating (near-)duplicates of mementos of the same TimeMap $URI\text{-}T$ if the mementos exceed a specific threshold α, which was determined empirically. The goal is to generate a reduced TimeMap $URI\text{-}T_{reduced}$ that contains only unique mementos of the URI. This process is called de-duplication.

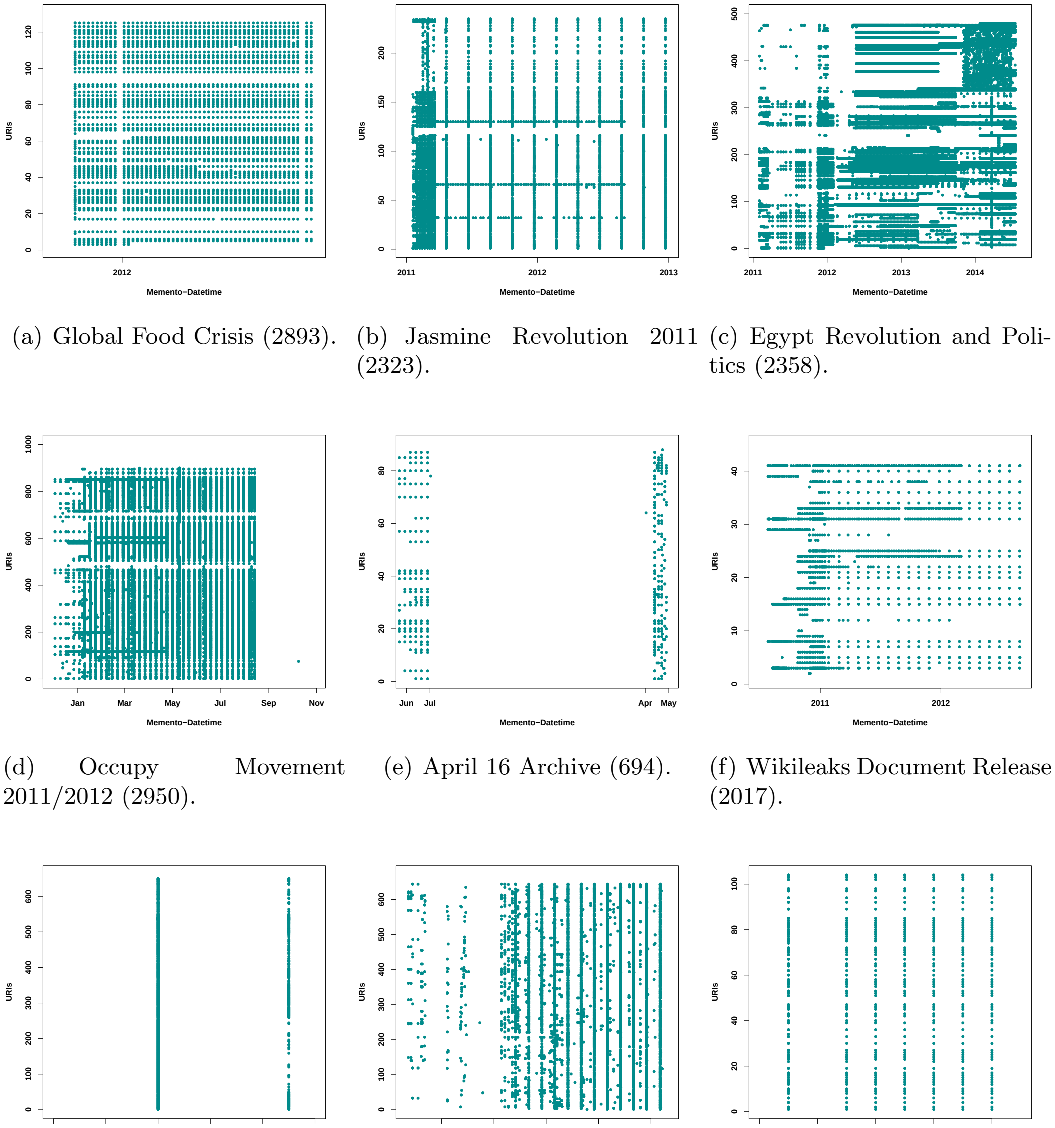
Another example of (near-)duplication that might occur within the collection is when two different sites have the same news story [259, 193, 262]. We select between the repeated news stories based on quality metrics that evaluate the mementos that are close to each others in terms of their content (see Section 9.4).

9.2 EXCLUDING THE NON-ENGLISH LANGUAGE PAGES

We detect the language of the content using the language detection library created by Shuyo [289] with precision $\geq 99\%$ [289, 54]. We select the English mementos and exclude other languages. The DSA framework can be applied on pages with other languages, but currently, we evaluate English language pages only.

9.3 SLICING THE COLLECTIONS

In the story generation process, we divide the collection into slices. The main challenge is how to determine the window size of each slice. For understanding the nature of the archived collections in terms of the crawl frequency of the seed URIs, we picked nine different collections in Archive-It that cover a wide range of topics, such as politics, crisis, health, etc. to tackle the following research questions: Is the crawling frequency similar for all the URIs in the collections?, Do curators take snapshots from the URIs at regular time intervals?, Do the crawls of URIs contains gaps? We extracted the TimeMaps for the seed URIs in each of the nine collections and visualized them. Figure 80 shows the visualizations for the Memento-Datetimes of multiple archived collections. The x-axis represents the Memento-Datetimes and the y-axis represents the seed URI. We expected to see more like Figure 80(a), in which the collection starts with a list of seeds that may increase through time, and the capture of these URIs continues at regular intervals. However, we found that in most cases, the crawl of the pages is not frequent. Furthermore, the crawl of a URI may not start at the same time as other URIs because the collection grows over time as new seed URIs are added, especially for long-running collections. For example, Figures 80(d) and 80(f) show that their seeds have different start dates and end dates. The Egypt Revolution collection (Figure 80(c)) has pages that were crawled starting in 2011 and other



(a) Global Food Crisis (2893).

(b) Jasmine Revolution 2011 (2323).

(c) Egypt Revolution and Politics (2358).

(d) Occupy Movement 2011/2012 (2950).

(e) April 16 Archive (694).

(f) Wikileaks Document Release (2017).

(g) Brazilian School Shooting (2535).

(h) Human Rights (1068).

(i) Russia Plane Crash (2823).

FIG. 80: Visualizations for the Memento-Datetimes of Archive-It collections.

Algorithm 2 Slicing the collection dynamically.

```
Input: $N$ is a reduced set of all mementos after excluding the off-topic (Chapter 8), duplicates (Section 9.1), non-English language mementos (Section 9.2)
Sort all the mementos in $N$ by their Memento-Datetimes
Define $S_r$ as the recommended number of slices
Define $S_a$ as the actual number of slices
if $|N| > 28$ then
    $$S_r = \lceil 28 + log_{10}|N| \rceil \qquad (15)$$
else
    $$S_r = |N| \qquad (16)$$
end if
$Y = \lceil |N|/S_r \rceil$
$S_a = \lceil |N|/Y \rceil$
$i = 1$
while $i < S_a$ do
    Move the next $Y$ mementos from $N$ into slice $i$
    $i = i + 1$
end while
Move the remaining mementos from $N$ into slice $S_a$
```

pages starting in 2013. There are collections with URIs whose crawling date is before the ending crawling date of the collection (for example, Figure 80(f)). Therefore, there are seed URIs that have 1000 mementos, while other seed URIs have just 20 mementos.

One of the reasons for stopping the crawling of a page may be the change of the page's topic through time. For example, the crawl of `hamdeensabahy.com` stopped after it went off-topic for a long time. This can be discovered if the curator checks the relevancy of the page manually. Another reason for ceasing the crawl of a URI is the change of the page's status from HTTP 200 to HTTP 404.

Slicing the collection can be done in two ways: dividing the collection into slices that have equal time intervals or dividing the collection into slices with an equal number of mementos. Slicing by time interval will not be appropriate for collections that have gaps in the middle. For example, the crawl frequency of Figure 80(e), where there is a large gap in crawling, will result in some slice having a large number of mementos and other slices do not have mementos at all. Therefore, we proposed a slicing algorithm that will distribute the mementos equally in a predefined number of slices that will be specified dynamically based on the number of mementos N in the collection after excluding the off-topic pages, non-English language pages, and the (near-)duplicates. The total number of resulted slices should be close to 28, which is suggested based on our study of the popular stories in Storify (Chapter 6).

Algorithm 2 defines the number of slices and the number of mementos per slice ($Y = \lceil N/S \rceil$). This algorithm will secure a uniform representation of the mementos based on the density of the mementos through time. Note that, the actual number of slices S_a will be less than the recommended S_r in some cases. For example, assume $N = 50$. Based on the equation 15, the recommended number of slices S_r will be 29 ($S_r = 29$), so $Y = 2$. Distributing 50 mementos equally by 2 mementos in each slice will result in 25 slices, so $S_a = 25$.

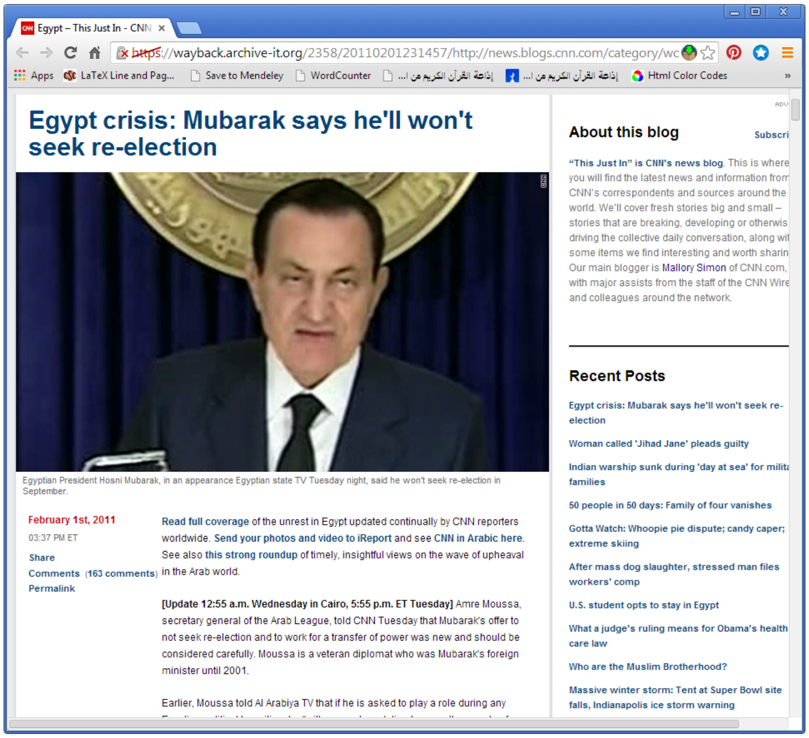
9.4 CLUSTERING THE MEMENTOS OF EACH SLICE

After dividing the collection into S_a slices, we cluster the Y mementos in each slice. The output of this step is a set of k clusters, where $k \geq S_a$. We used the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [97] to cluster the mementos in each slice based on their textual contents. DBSCAN does not require the specification of the number of clusters a priori, as opposed to k-means clustering [125]. DBSCAN needs two parameters: $minPts$, the minimum number of points in each cluster; ϵ, the radius of the cluster. There is no standard similarity cut (ϵ) that represents if two topics are similar. We found empirically that $\epsilon = 0.4$ is a good value for increasing the novelty between clusters and producing stories that have the desired number of resources, which is close to 28.

A resulting cluster from this step contains one or more mementos that are close to one another. For example, the two mementos in Figures 81(a) and 81(b) should be in the same cluster. The choice between them will be based on the quality metrics we specify in the following section.

9.5 SELECT THE BEST REPRESENTATIVE MEMENTOS

The previous steps produce a set of k clusters, in which each cluster contains the mementos that are close to each others in terms of content. In this step, we will select only one memento from each cluster. So the output of this step is k mementos. Choosing the best candidates for each event of the story will tremendously affect the quality of the created story. We specify the memento quality based on the amount of damage for the memento [56] and if the memento generates a visually attractive link preview when inserting to a tool like Storify.


Egypt – This Just In - CNN
https://wayback.archive-it.org/2358/20110201231457/http://news.blogs.cnn.com/category/wc
Apps
LaTeX Line and Pag...
Save to Mendeley
WordCounter
إذاعة القرآن الكريم من ا...
إذاعة القرآن الكريم من ا...
Html Color Codes
ADV
Subscri

Egypt crisis: Mubarak says he'll won't seek re-election

Egyptian President Hosni Mubarak, in an appearance Egyptian state TV Tuesday night, said he won't seek re-election in September.

February 1st, 2011
03:37 PM ET
Share
Comments (163 comments)
Permalink

Read full coverage of the unrest in Egypt updated continually by CNN reporters worldwide. Send your photos and video to iReport and see CNN in Arabic here. See also this strong roundup of timely, insightful views on the wave of upheaval in the Arab world.

[Update 12:55 a.m. Wednesday in Cairo, 5:55 p.m. ET Tuesday] Amre Moussa, secretary general of the Arab League, told CNN Tuesday that Mubarak's offer to not seek re-election and to work for a transfer of power was new and should be considered carefully. Moussa is a veteran diplomat who was Mubarak's foreign minister until 2001.

Earlier, Moussa told Al Arabiya TV that if he is asked to play a role during any

About this blog

"This Just In" is CNN's news blog. This is where you will find the latest news and information from CNN's correspondents and sources around the world. We'll cover fresh stories big and small – stories that are breaking, developing or otherwis driving the collective daily conversation, along wit some items we find interesting and worth sharin Our main blogger is Mallory Simon of CNN.com, with major assists from the staff of the CNN Wire and colleagues around the network.

Recent Posts

Egypt crisis: Mubarak says he'll won't seek re-election

Woman called 'Jihad Jane' pleads guilty

Indian warship sunk during 'day at sea' for milita families

50 people in 50 days: Family of four vanishes

Gotta Watch: Whoopie pie dispute; candy caper; extreme skiing

After mass dog slaughter, stressed man files workers' comp

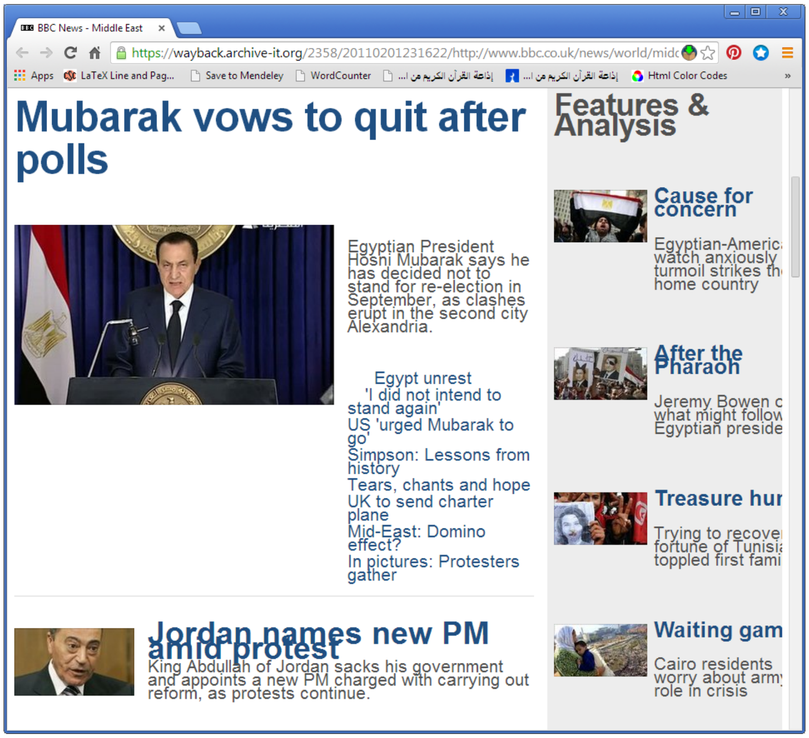
U.S. student opts to stay in Egypt

What a judge's ruling means for Obama's health care law

Who are the Muslim Brotherhood?

Massive winter storm: Tent at Super Bowl site falls, Indianapolis ice storm warning

(a) Feb 01, 2011: CNN covering Mubarak's speech.


BBC News - Middle East
https://wayback.archive-it.org/2358/20110201231622/http://www.bbc.co.uk/news/world/midc
Apps
LaTeX Line and Pag...
Save to Mendeley
WordCounter
إذاعة القرآن الكريم من ا...
إذاعة القرآن الكريم من ا...
Html Color Codes

Mubarak vows to quit after polls

Egyptian President Hosni Mubarak says he has decided not to stand for re-election in September, as clashes erupt in the second city Alexandria.

Egypt unrest
'I did not intend to stand again'
US 'urged Mubarak to go'
Simpson: Lessons from history
Tears, chants and hope
UK to send charter plane
Mid-East: Domino effect?
In pictures: Protesters gather

Jordan names new PM amid protest

King Abdullah of Jordan sacks his government and appoints a new PM charged with carrying out reform, as protests continue.

Features & Analysis

Cause for concern

Egyptian-Americ watch anxiously turmoil strikes th home country

After the Pharaoh

Jeremy Bowen c what might follow Egyptian preside

Treasure hur

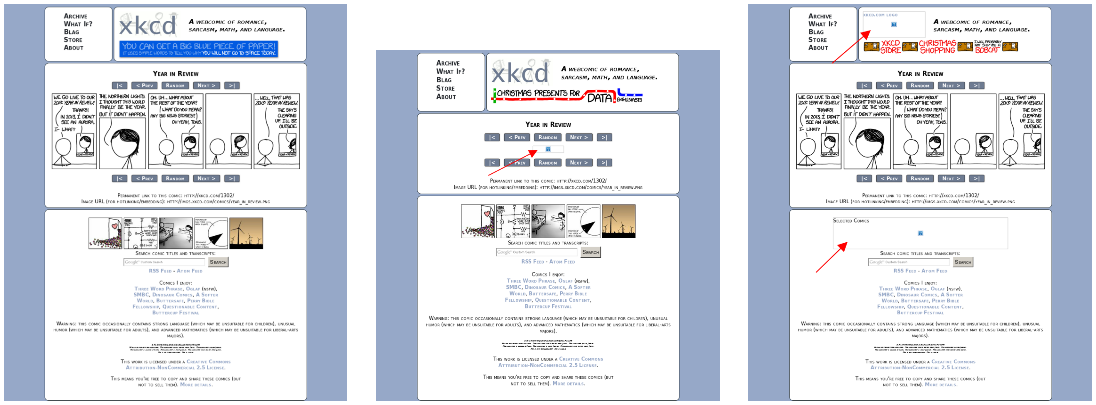
Trying to recove fortune of Tunisi toppled first fami

Waiting gam

Cairo residents worry about army role in crisis

(b) Feb 01, 2011: BBC covering Mubarak Speech.

FIG. 81: The coverage of the same news from two popular Web sites, but the archived version of the BBC page is missing style sheets.



(a) All three of the embedded images are included and identified by the red arrows. Missing resources represent 17%.

(b) The large, central image (that is the main content of the page) was removed, identified by the red arrow. Missing resources represent 24%.

(c) The XKCD logo was removed and banner of comics, identified by the red arrows. Missing resources represent 29%.

FIG. 82: The XKCD example demonstrates that embedded resources have varying human-perceived importance to their page [55].

We weight each memento with quality measure M_q which represents the total quality of each memento. M_q is calculated as follows:

$$M_q = (1 - w_d * D_m) + w_l * M_l + w_c * M_c \tag{17}$$

where D_m is the value of memento damage (Section 9.5.1), M_l is URI level (Section 9.5.2), and M_c is the URI category (Section 9.5.2).

We explain each metric in Equation 17 in detail in the following sections. We tune the system using different weights for each of the quality metrics as shown in Equation 17. We set level weight ($w_l = 0.45$), memento damage weight ($w_d = 0.40$), and category weight ($w_c = 0.15$). Setting these weights needs further testing with multiple collections. In the following subsections, we will explain how we calculate each metric.

9.5.1 MEMENTO DAMAGE

When Web crawlers attempt to capture Web pages, they may not capture every resource on every page, which can result in missing a portion of the embedded resources of the pages (e.g., images and style sheets) [39]. Some of the embedded resources are more important to the user than others [56]. Brunelle et al. [56, 57] used the example in Figure 82 to demonstrate that the proportion of the missing resources of the page is not an accurate representation of the memento damage. Figure 82(a) shows the live Web version of the XKCD[2] page, which is missing two embedded stylesheets that represent 17% of all the embedded resources. They manually removed the main image of the page (the central image is most important to the utility of the page) that resulted in 24% of the embedded resources being missing (Figure 82(b)). Figure 82(c) shows the same page after they manually removed the logo and banner, which are not essential to the user's understanding of the XKCD content. With missing the logo and the banner, the missing resources represent 29% of the total resources. Therefore, the importance of the missing resources is an essential factor in assessing the damage of mementos.

Many approaches have been proposed for measuring memento damage [56, 55, 169]. In the DSA framework, we adopt Brunelle's algorithm for assessing memento damage [56, 57]. The main idea of Brunelle's approach is generating a damage metric that is close to the perception of Web users. They first measure the importance of the embedded resources to rate the damage of the memento. Their proposed algorithm is based on the MIME type, size, and location of the embedded resource to calculate the importance of the embedded resources. They defined D_m as the damage rating, or cumulative damage, which is a normalized value ranging from [0, 1].

D_m is calculated as follows [57]:

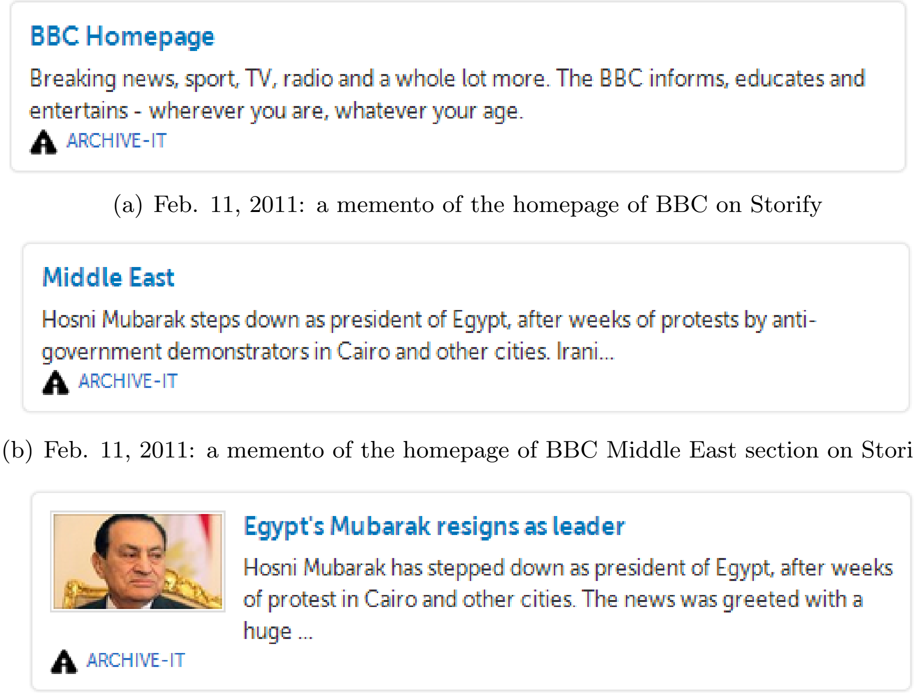
$$D_m = \frac{D_{m_{actual}}}{D_{m_{potential}}} \tag{18}$$

They define the set of all embedded resources R and the set of all missing resources R_r in Equation 19 to determine potential $D_{m_{potential}}$ and actual damage $D_{m_{actual}}$.

$$\begin{aligned} R &= \{\text{All embedded resources requested}\} \\ R_r &= \{\text{All missing embedded resources}\} \\ R_r &\subseteq R \end{aligned} \tag{19}$$

Calculating D_m starts with loading the URI-M with PhantomJS, then finding the potential damage $D_{m_{potential}}$ by determining the importance of CSS, multimedia, and images, and then determining proportion of unsuccessfully dereferenced embedded resources and finding the actual damage $D_{m_{actual}}$ (same as the last step but with only those URI-Ms unsuccessfully dereferenced). The last step is determining the total damage D_m, which we use in the DSA framework (Equation 17) as our indicator for the memento damage.

[2] http://www.xkcd.com/



BBC Homepage

Breaking news, sport, TV, radio and a whole lot more. The BBC informs, educates and entertains - wherever you are, whatever your age.

(a) Feb. 11, 2011: a memento of the homepage of BBC on Storify

Middle East

Hosni Mubarak steps down as president of Egypt, after weeks of protests by anti-government demonstrators in Cairo and other cities. Irani...

(b) Feb. 11, 2011: a memento of the homepage of BBC Middle East section on Storify

Egypt's Mubarak resigns as leader

Hosni Mubarak has stepped down as president of Egypt, after weeks of protest in Cairo and other cities. The news was greeted with a huge ...

(c) Feb. 11, 2011: a memento of the BBC article page on Storify

FIG. 83: Storify creates better snippets from a specific article (i.e., deep links) than a homepage about the same event.

9.5.2 SNIPPET QUALITY

As we declared earlier, we use social media to visualize the generated stories. When a user posts a link on social media networks, e.g., Facebook and Storify, a visual *snippet* with a title, a summary of the content, and an image is extracted from that link. These visual snippets are created from the HTML tags of the Web page. The type and the level of the URI affect the quality of the snippet. In the following subsections, we will illustrate how the level and category of a Web page affect the quality of the snippet and our weighting algorithms for the pages.

URI level-based quality

We experimented with the generation of visual snippets for many different kinds of URIs. We discovered that social media can generate better snippets from articles that focus on only one topic (these articles also often have a long URI path length, e.g., `cnn.com/a/b/c/2011/4/2`), while it does not extract nice snippets from homepages that have an overview of multiple topics (these pages often have a short URI path length, e.g., cnn.com). For instance, Figure 83 shows three different snippets on Storify for three different URIs with the same domain (bbc.com) in the Egypt Revolution collection. Each of the three URIs covers the same event, Mubarak's stepping down. The snippet that is created from a URI[3] of a "deep link"[4] (Figure 83(c)) is better in terms of the title, image, and the summary text than the snippet that is created from a high-level URI (Figure 83(a) and 83(b)). The second best snippet is the one that is generated from the homepage of the Middle East section[5] (Figure 83(b)) and the one with the least quality is the one that is generated from the BBC homepage[6] (Figure 83(a)).

Therefore, if $aboutness(URI\text{-}R_i@t_x) \approx aboutness(URI\text{-}R_j@t_x)$, where $URI\text{-}R_i$ is a deep URI and $URI\text{-}R_j$ is high level URI ($M_l(URI\text{-}R_i) > M_l(URI\text{-}R_j)$), then $URI\text{-}R_i@t_x$ is preferred over $URI\text{-}R_j@t_x$. In the DSA framework, the deeper the URI-R, the higher weight we assign to this URI-R based on its level. The value of M_l is normalized in the range of [0, 1]. For example, the M_l of cnn.com/a/b/c/2011/4/2 will be assigned 0.6 and $M_l = 0.1$ for cnn.com/.

URI category-based quality

By testing Storify, we found that the page category may affect the quality of the extracted snippets. Moreover, there are different kinds of URIs in which the extraction fails to capture information related to the topic of the collection such as URIs for pages on Facebook, Facebook accounts, Twitter accounts, Google groups, etc. When these pages are posted on Storify, the text of the snippet is extracted from the description of the profiles or pages. For example, Figure 84 shows the snippet representation of the memento of the @Haitifeed Twitter account[7] in the "Haiti Earthquake"[8] collection. The text of the snippet in Figure 84(b) shows the description of the Twitter page which does not represent the topic of the collection.

Figure 85 contains a memento of the "We are all Khaled Said" page on Facebook, which started the Egyptian Revolution events on Facebook in January 2011. As we mentioned in Chapter 1, the page was created in June 2010 for bringing the attention to a young man named Khaled Said who was beaten to death by Egyptian security forces in Alexandria, Egypt. Although the page is important to the Egyptian Revolution events, the page is

[3] https://wayback.archive-it.org/2358/20110211192204/http://www.bbc.co.uk/news/world-middle-east-12433045

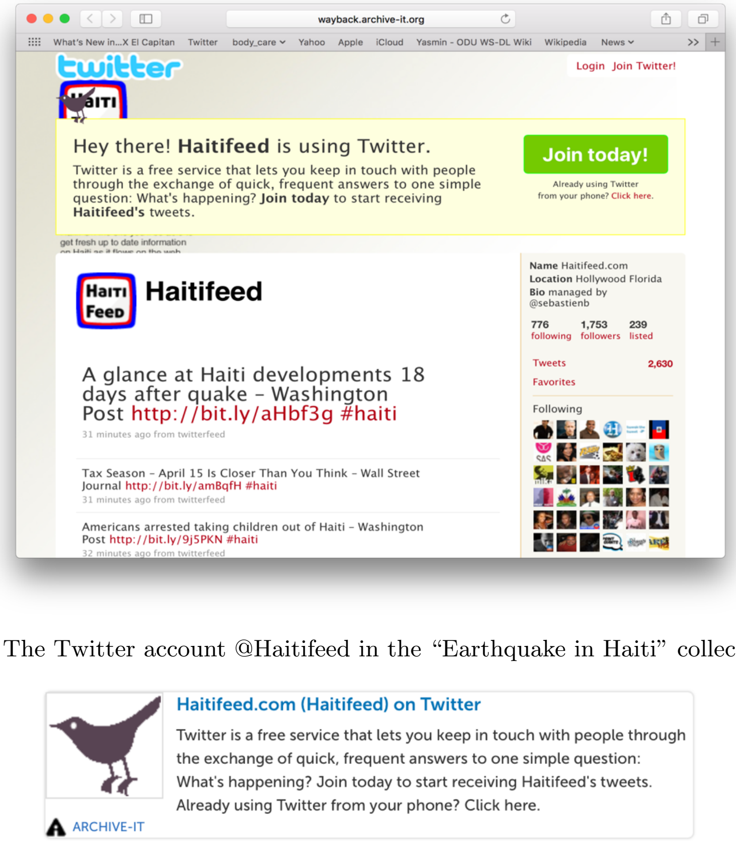
[4] https://en.wikipedia.org/wiki/Deep_linking

[5] https://wayback.archive-it.org/2358/20110211191942/http://www.bbc.co.uk/news/world/middle_east/

[6] https://wayback.archive-it.org/2358/20110211191429/http://www.bbc.co.uk/

[7] http://twitter.com/Haitifeed/

[8] https://archive-it.org/collections/1784/



Hey there! **Haitifeed** is using Twitter.

Twitter is a free service that lets you keep in touch with people through the exchange of quick, frequent answers to one simple question: What's happening? **Join today** to start receiving **Haitifeed's** tweets.

Already using Twitter from your phone? Click here.

get fresh up to date information on Haiti as it flows on the web

Haitifeed

Name Haitifeed.com
Location Hollywood Florida
Bio managed by @sebastienb

776 following 1,753 followers 239 listed

Tweets 2,630

A glance at Haiti developments 18 days after quake - Washington Post http://bit.ly/aHbf3g #haiti
31 minutes ago from twitterfeed

Tax Season - April 15 Is Closer Than You Think - Wall Street Journal http://bit.ly/amBqfH #haiti
31 minutes ago from twitterfeed

Americans arrested taking children out of Haiti - Washington Post http://bit.ly/9j5PKN #haiti
32 minutes ago from twitterfeed

(a) The Twitter account @Haitifeed in the "Earthquake in Haiti" collection

Haitifeed.com (Haitifeed) on Twitter

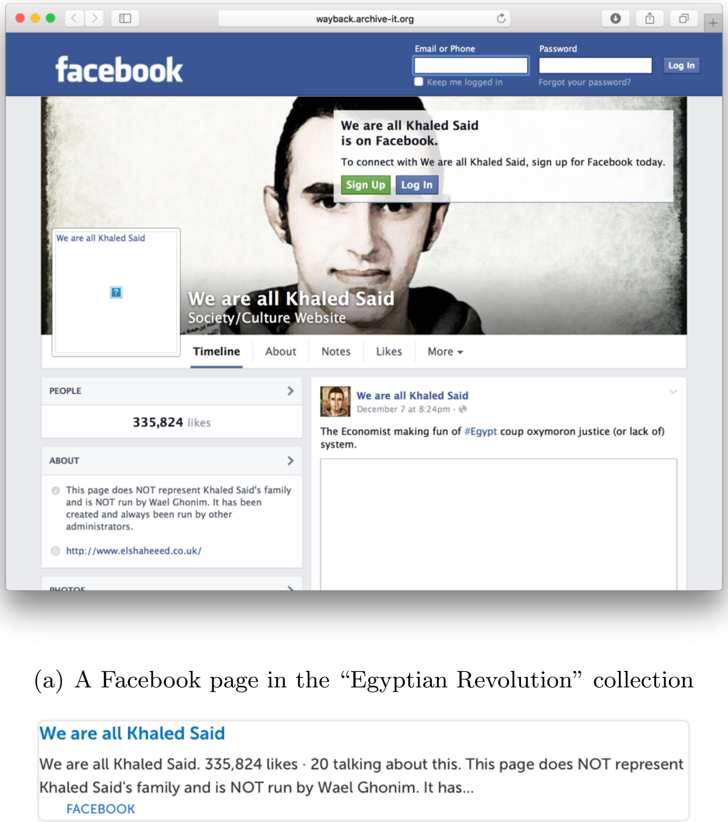
Twitter is a free service that lets you keep in touch with people through the exchange of quick, frequent answers to one simple question: What's happening? Join today to start receiving Haitifeed's tweets. Already using Twitter from your phone? Click here.

(b) The snippet of Twitter account @Haitifeed on Storify

FIG. 84: Frequently a memento of a Twitter account does not produce good representative snippet. Link: `http://wayback.archive-it.org/1784/20100131023240/http://twitter.com/Haitifeed/`

not a good candidate to be included in the story. That is because the extracted snippet contains the description of the page, which does not have anything relevant to the Egyptian Revolution. Including such pages may not be a good selection for generating an attractive story on Storify.

Another example that shows how the type of the Web site affects the quality of the snippet is illustrated in Figure 86. The figure shows Web pages from two different domains (cnn.com and news.blogs.cnn.com) describing the same event for one of the Egyptian Revolution figures who was arrested then released in the 14 days of the Egyptian Revolution. The snippet produced from a news article (Figure 86(c)) is better than the snippet generated from a blog post (Figure 86(d).



We are all Khaled Said is on Facebook.

To connect with We are all Khaled Said, sign up for Facebook today.

We are all Khaled Said

We are all Khaled Said
Society/Culture Website

335,824 likes

This page does NOT represent Khaled Said's family and is NOT run by Wael Ghonim. It has been created and always been run by other administrators.

http://www.elshaheeed.co.uk/

We are all Khaled Said
December 7 at 8:24pm

The Economist making fun of #Egypt coup oxymoron justice (or lack of) system.

(a) A Facebook page in the "Egyptian Revolution" collection

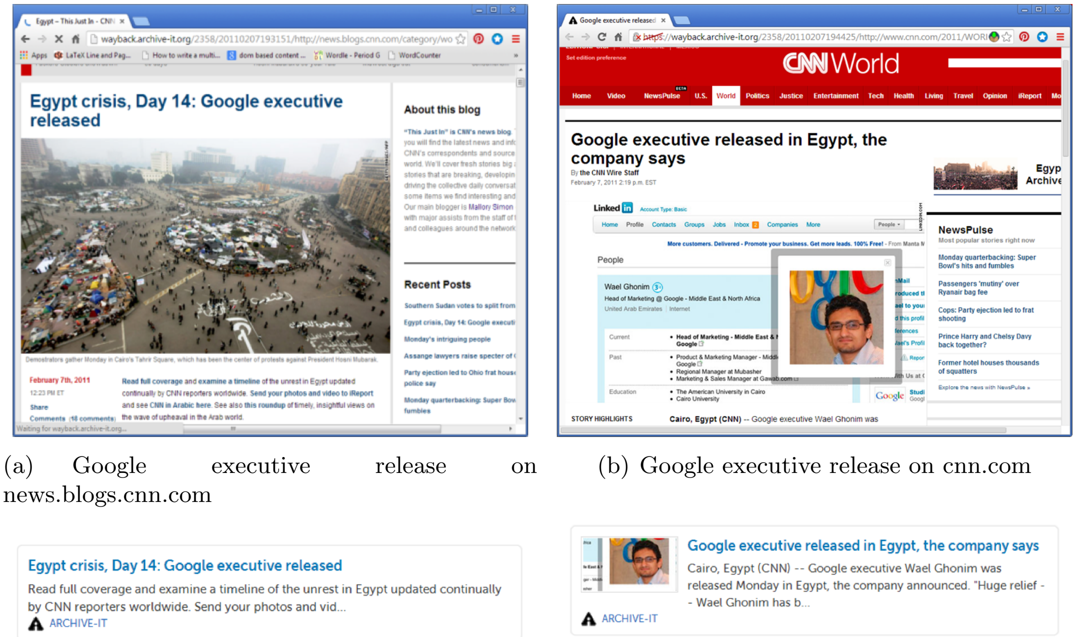
We are all Khaled Said

We are all Khaled Said. 335,824 likes · 20 talking about this. This page does NOT represent Khaled Said's family and is NOT run by Wael Ghonim. It has...

FACEBOOK

(b) The snippet of the Facebook page on Storify

FIG. 85: Frequently a memento of a Facebook page does not produce a good representative snippet. Link: `http://wayback.archive-it.org/2358/20141225080305/https:/www.facebook.com/elshaheeed.co.uk`



Egypt – This Just In - CNN

wayback.archive-it.org/2358/20110207193151/http://news.blogs.cnn.com/category/wo

Apps | LaTeX Line and Pag... | How to write a multi... | dom based content ... | Wordle - Period G | WordCounter

Egypt crisis, Day 14: Google executive released

Demostrators gather Monday in Cairo's Tahrir Square, which has been the center of protests against President Hosni Mubarak.

February 7th, 2011
12:23 PM ET
Share
Comments (18 comments)

Read full coverage and examine a timeline of the unrest in Egypt updated continually by CNN reporters worldwide. Send your photos and video to iReport and see CNN in Arabic here. See also this roundup of timely, insightful views on the wave of upheaval in the Arab world.

About this blog

"This Just In" is CNN's news blog. you will find the latest news and inf CNN's correspondents and source world. We'll cover fresh stories big stories that are breaking, developin driving the collective daily conversat some items we find interesting and Our main blogger is Mallory Simon with major assists from the staff of and colleagues around the network

Recent Posts

- Southern Sudan votes to split from
- Egypt crisis, Day 14: Google executi
- Monday's intriguing people
- Assange lawyers raise specter of
- Party ejection led to Ohio frat house police say
- Monday quarterbacking: Super Bow fumbles

Waiting for wayback.archive-it.org...

(a) Google executive release on news.blogs.cnn.com

Google executive released

https://wayback.archive-it.org/2358/20110207194425/http://www.cnn.com/2011/WORI

Set edition preference

CNN World

Home | Video | NewsPulse | U.S. | World | Politics | Justice | Entertainment | Tech | Health | Living | Travel | Opinion | iReport | Mo

Google executive released in Egypt, the company says

By the CNN Wire Staff
February 7, 2011 2:19 p.m. EST

Linked in Account Type: Basic

Home | Profile | Contacts | Groups | Jobs | Inbox 2 | Companies | More

More customers. Delivered - Promote your business. Get more leads. 100% Free! - From Manta M

People

Wael Ghonim
Head of Marketing @ Google - Middle East & North Africa
United Arab Emirates | Internet

Current	• Head of Marketing - Middle East & N Google
Past	• Product & Marketing Manager - Middl Google • Regional Manager at Mubasher • Marketing & Sales Manager at Gawab.com
Education	• The American University in Cairo • Cairo University

Egyp Archive

NewsPulse
Most popular stories right now

- Monday quarterbacking: Super Bowl's hits and fumbles
- Passengers 'mutiny' over Ryanair bag fee
- Cops: Party ejection led to frat shooting
- Prince Harry and Chelsy Davy back together?
- Former hotel houses thousands of squatters

Explore the news with NewsPulse »

STORY HIGHLIGHTS

Cairo, Egypt (CNN) -- Google executive Wael Ghonim was

(b) Google executive release on cnn.com

Egypt crisis, Day 14: Google executive released

Read full coverage and examine a timeline of the unrest in Egypt updated continually by CNN reporters worldwide. Send your photos and vid...

ARCHIVE-IT

Google executive released in Egypt, the company says

Cairo, Egypt (CNN) -- Google executive Wael Ghonim was released Monday in Egypt, the company announced. "Huge relief - - Wael Ghonim has b...

ARCHIVE-IT

(c) The snippet of news.blogs.cnn.com on Storify

(d) The snippet of cnn.com on Storify

FIG. 86: The snippet of cnn.com on Storify.

When creating a collection at Archive-It, curators may group Web pages into categories to allow easier filtering and browsing. However, many collections lack such grouping, making it cumbersome to find related Web pages based on the categories. Furthermore, several Archive-It collections do not have the sites of the collection organized into groups (e.g., Pakistan floods collection[9]). Thus, we used our previously proposed heuristic-based categorization [252], which classifies the URI based on its domain component, then assigns each category a weight based on how the category affects the snippet quality. Examples for how we categorized the URIs include the following:

- Social Media: Facebook, Twitter, Google Plus, or Reddit.
- News: News sites, such as BBC, CNN, NYTimes.
- Blogs: Blogs or WordPress sites.
- Videos: YouTube or Dailymotion.
- Others: all sites that do not match the previous rules.

[9] https://archive-it.org/collections/2836/

```
{
   "title": "Egypt Revolution - different URIs through time"
  ,"slug": "egypt-revolution-story"
  ,"description": "This is an automatically generated story from
      the Egypt Revolution and Politics collection in Archive-It."
,"thumbnail": "https://storify.com/public/img/default-thumb.gif"
  ,"elements": [
       {"type":"link",
            "permalink":"http://wayback.archive-it.org
               /2358/20110211072257/http://news.blogs.cnn.com/
               category/world/egypt-world-latest-news/",
            "data":{
               "link":{
                  "title":"Egypt     This Just In",
                  "description":"Egyptian opposition leader
                     Mohamed ElBaradei said that...",
                  "thumbnail":"http://wayback.archive-it.org
                     /2358/20110211072257im_/http://i2.cdn.turner
                     .com/cnn/2011/images/02/08/
                     t1larg_assange_gi_afp.jpg"}},
            "source":{
               "name":"news.blogs.cnn.com",
               "href":"http://news.blogs.cnn.com"},
            "attribution":{
               "name":"news.blogs.cnn.com",
               "href":"http://news.blogs.cnn.com"}},
        {"type":"link",
            "permalink":"http://wayback.archive-it.org
               /2358/20110814100103/http://news.egypt.com/en/",
            "data":{
               "link":{
                  "title":"Egypt News",
                  "description":"Telecom Egypt upgrades network
                     cable system...",
                  "thumbnail":"http://wayback.archive-it.org
                     /2358/20110814100103im_/http://news.egypt.
                     com/english/thumbnail.php?file=
                     Mohamed_elBaradei_249905761.jpg&size=
                     article_large"}
            },
            "source":{
               "name":"news.egypt.com",
               "href":"http://news.egypt.com"},
            "attribution":{
               "name":"news.egypt.com",
               "href":"http://news.egypt.com"}},
      ...
    ]
}
```

FIG. 87: The JSON object of a generated story from the Egypt Revolution collection in Archive-It by our implementation of the DSA framework.

We assigned each page a weight $0 \leq M_c \leq 1$ based on its category. We give higher weights to news Web sites, video, social media posts then blogs come next and the lowest weight goes to Facebook pages, Twitter accounts, Google groups, etc.

9.6 ORDER THE SELECTED MEMENTOS CHRONOLOGICALLY

The previous step results in k mementos. In this step, we order the k mementos chronologically. As we discussed in Section 3.5, there are multiple notions of time for an archived Web page: Creation-Datetime (CD), Last-Modified (LM), Memento-Datetime (MD), and Aboutness-Time (AT). The temporal order of the events is important to create a good narrative that the user perceives as it appeared in the past, especially with the broad summary story. We use the "Newspaper: Article scraping and curation" Python library [249] to extract the publishing date of the Web page. It applies multiple strategies such as extracting the date from a URI or from the Web page metadata. If neither of these strategies succeed, we use the Memento-Datetime as the estimated publishing date. Finally, we order the mementos chronologically based on their dates.

9.7 VISUALIZING THE STORIES USING STORIFY

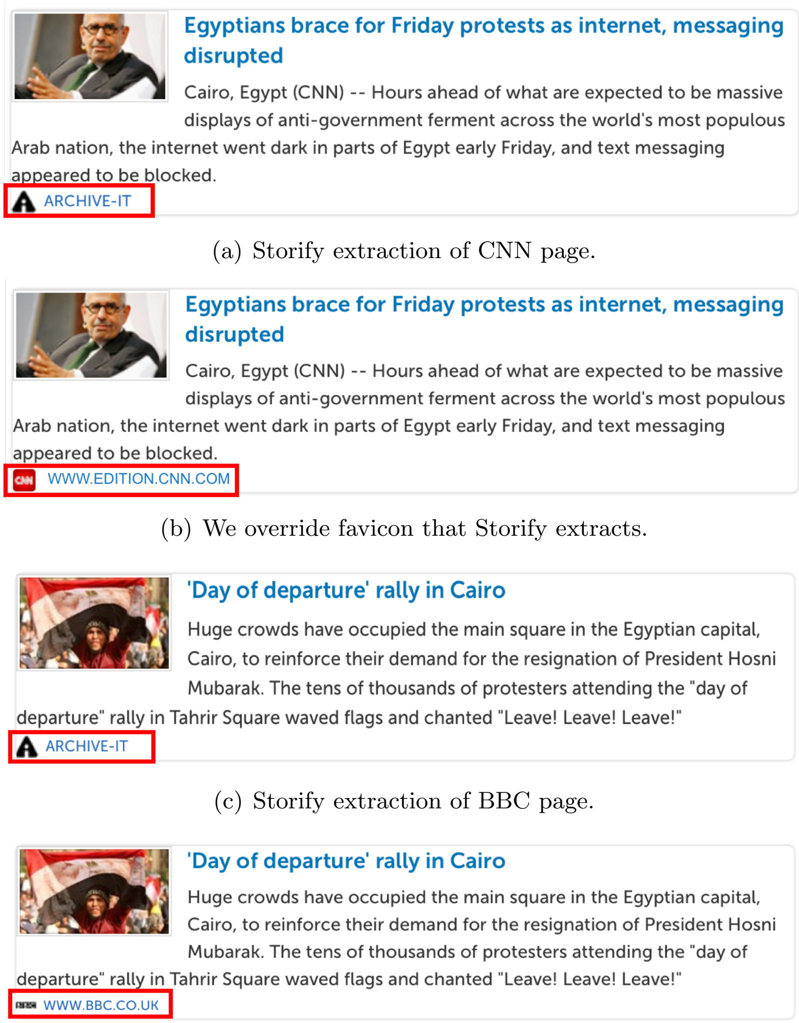
The goal of the DSA framework is to find the $k \approx 28$ samples that best summarize the collection, and then to insert those $k \approx 28$ samples into any visualization tool. In our implementation, we used Storify, a popular platform for storytelling, to visualize the set of $k \approx 28$ mementos that represent the extracted story from the collection. Storify provides an API[10] that allows users to create and publish stories by sending objects of the elements of the stories in JSON format. Once a story is created and pushed to Storify, it can be edited and shared.

Figure 87 shows an example of a Storify story[11] that was generated automatically by our implementation from the Egypt Revolution collection in Archive-It. After we selected $k \approx 28$ mementos that represent the collection, we generated the story elements for the mementos. Each story in a JSON object contains the metadata of the story, such as the story name and description, then the details of each element such as the hyperlink, the extracted title, etc. We use two different methods for creating a snippet on Storify.

1. **Storify extraction:** We send the links of the mementos to Storify and let the snippet extraction be generated by Storify. In this method, we override the favicon of the resource that is created by Storify (see Figure 88), because Storify uses the Archive-It

[10] http://dev.storify.com/api/

[11] https://storify.com/yasmina_anwar/egypt-revolution-story



Egyptians brace for Friday protests as internet, messaging disrupted

Cairo, Egypt (CNN) -- Hours ahead of what are expected to be massive displays of anti-government ferment across the world's most populous Arab nation, the internet went dark in parts of Egypt early Friday, and text messaging appeared to be blocked.

ARCHIVE-IT

(a) Storify extraction of CNN page.

Egyptians brace for Friday protests as internet, messaging disrupted

Cairo, Egypt (CNN) -- Hours ahead of what are expected to be massive displays of anti-government ferment across the world's most populous Arab nation, the internet went dark in parts of Egypt early Friday, and text messaging appeared to be blocked.

WWW.EDITION.CNN.COM

(b) We override favicon that Storify extracts.

'Day of departure' rally in Cairo

Huge crowds have occupied the main square in the Egyptian capital, Cairo, to reinforce their demand for the resignation of President Hosni Mubarak. The tens of thousands of protesters attending the "day of departure" rally in Tahrir Square waved flags and chanted "Leave! Leave! Leave!"

ARCHIVE-IT

(c) Storify extraction of BBC page.

'Day of departure' rally in Cairo

Huge crowds have occupied the main square in the Egyptian capital, Cairo, to reinforce their demand for the resignation of President Hosni Mubarak. The tens of thousands of protesters attending the "day of departure" rally in Tahrir Square waved flags and chanted "Leave! Leave! Leave!"

WWW.BBC.CO.UK

(d) We override favicon that Storify extracts.

FIG. 88: Example for how we override Storify's extracted favicon to generate more visually attractive snippets.

favicon for all the pages regardless of the original source of the page (Figures 88(a) and 88(c)).

2. **Newspaper extraction:** We use the "Newspaper" Python library for extracting and curating articles to extract the metadata of the mementos (the page title, descriptive text, the main image, and the Creation-Datetime or Last-Modified). We put the metadata of the elements in the story's structure in JSON format, then post the story on Storify.

Each extraction method has its advantages and disadvantages. The Newspaper extraction allows us to attach the date to the title and to control the length of the snippet. This gives the story a more organized look (Figure 89). On the other hand, Storify detects the image and the title in most of the pages better than Newspaper extractor (Figure 90). Furthermore, the Storify extractor is much faster because it runs on the server side. Storify takes a fraction of second for publishing $k \approx 28$ mementos, while Newspaper takes ≈ 2 minutes to extract the metadata from the same number of mementos.

9.8 REVISITING THE EGYPTIAN REVOLUTION EXAMPLE IN CHAPTER 1

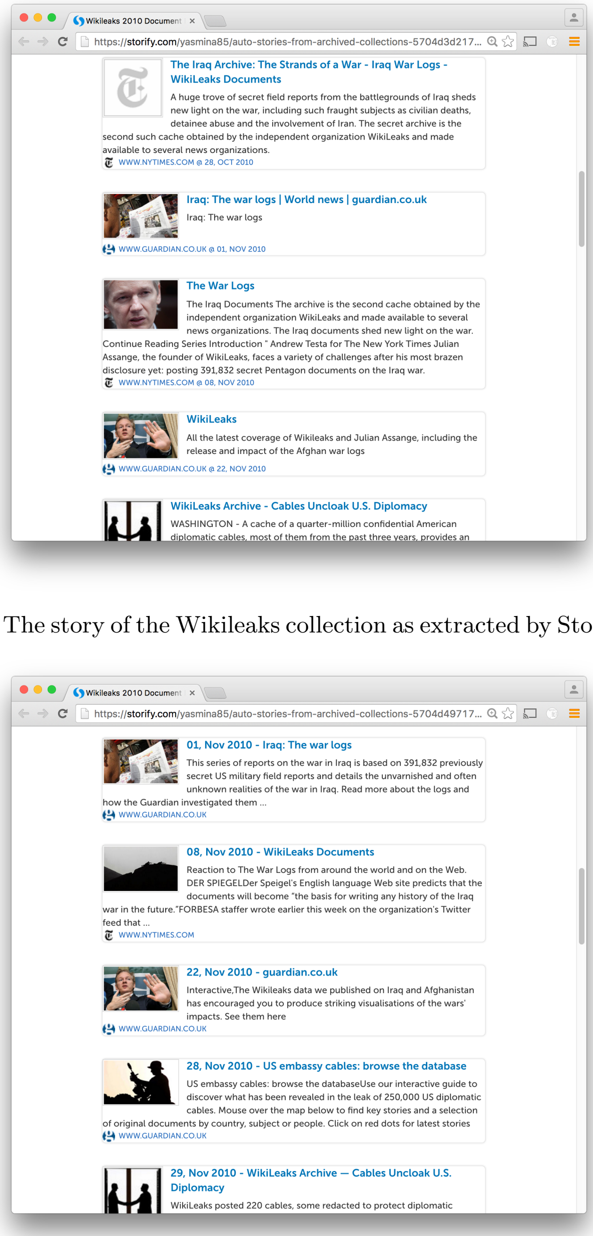
As illustrated in Figure 5, there are three collections that include the Egyptian Revolution and the user may not immediately understand the subtle differences between them. In Section 3.1, we described the problem of understanding each of the three collections and illustrated how it is difficult to browse all the URIs and the mementos in each collection.

We revisit the example of the Egyptian Revolution introduced in Section 3.1 to show the effectiveness of the DSA framework in helping the user to understand each of the three collections. We extracted three broad summaries[12,13,14] from each of the three collections. The resulting generated stories from the DSA framework are shown in Figures 91, 92, and 93. The user can gain an understanding about the holdings of each collection from the snippets of the $k \approx 28$ pages chosen from each collection. We notice from the figure that the resources in the "2011 North Africa and Middle East" collection are about different countries in the Middle East, not only Egypt, while the holdings of the "Egypt Revolution and Politics" collection are only about Egypt. The "2011 Arab Spring" story contains only one element because the "2011 Arab Spring" collection has only five seed URIs. Of those, one seed URI had only two mementos, which were similar to each other, and the remaining four seed URIs were off-topic of the collection.

[12] https://storify.com/yasmina85/auto-stories-from-archived-collections-56fbc3d1b8d27c6f6571c647

[13] https://storify.com/yasmina85/auto-stories-from-archived-collections-5702ff8f228eede273d49c21

[14] https://storify.com/yasmina85/auto-stories-from-archived-collections-5702c7f1228eede273d48ddf



Wikileaks 2010 Document
https://storify.com/yasmina85/auto-stories-from-archived-collections-5704d3d217...

The Iraq Archive: The Strands of a War - Iraq War Logs - WikiLeaks Documents

A huge trove of secret field reports from the battlegrounds of Iraq sheds new light on the war, including such fraught subjects as civilian deaths, detainee abuse and the involvement of Iran. The secret archive is the second such cache obtained by the independent organization WikiLeaks and made available to several news organizations.

WWW.NYTIMES.COM @ 28, OCT 2010

Iraq: The war logs | World news | guardian.co.uk

Iraq: The war logs

WWW.GUARDIAN.CO.UK @ 01, NOV 2010

The War Logs

The Iraq Documents The archive is the second cache obtained by the independent organization WikiLeaks and made available to several news organizations. The Iraq documents shed new light on the war. Continue Reading Series Introduction " Andrew Testa for The New York Times Julian Assange, the founder of WikiLeaks, faces a variety of challenges after his most brazen disclosure yet: posting 391,832 secret Pentagon documents on the Iraq war.

WWW.NYTIMES.COM @ 08, NOV 2010

WikiLeaks

All the latest coverage of Wikileaks and Julian Assange, including the release and impact of the Afghan war logs

WWW.GUARDIAN.CO.UK @ 22, NOV 2010

WikiLeaks Archive - Cables Uncloak U.S. Diplomacy

WASHINGTON - A cache of a quarter-million confidential American diplomatic cables, most of them from the past three years, provides an

(a) The story of the Wikileaks collection as extracted by Storify.

Wikileaks 2010 Document
https://storify.com/yasmina85/auto-stories-from-archived-collections-5704d49717...

01, Nov 2010 - Iraq: The war logs

This series of reports on the war in Iraq is based on 391,832 previously secret US military field reports and details the unvarnished and often unknown realities of the war in Iraq. Read more about the logs and how the Guardian investigated them ...

WWW.GUARDIAN.CO.UK

08, Nov 2010 - WikiLeaks Documents

Reaction to The War Logs from around the world and on the Web. DER SPIEGELDer Speigel's English language Web site predicts that the documents will become "the basis for writing any history of the Iraq war in the future."FORBESA staffer wrote earlier this week on the organization's Twitter feed that ...

WWW.NYTIMES.COM

22, Nov 2010 - guardian.co.uk

Interactive,The Wikileaks data we published on Iraq and Afghanistan has encouraged you to produce striking visualisations of the wars' impacts. See them here

WWW.GUARDIAN.CO.UK

28, Nov 2010 - US embassy cables: browse the database

US embassy cables: browse the databaseUse our interactive guide to discover what has been revealed in the leak of 250,000 US diplomatic cables. Mouse over the map below to find key stories and a selection of original documents by country, subject or people. Click on red dots for latest stories

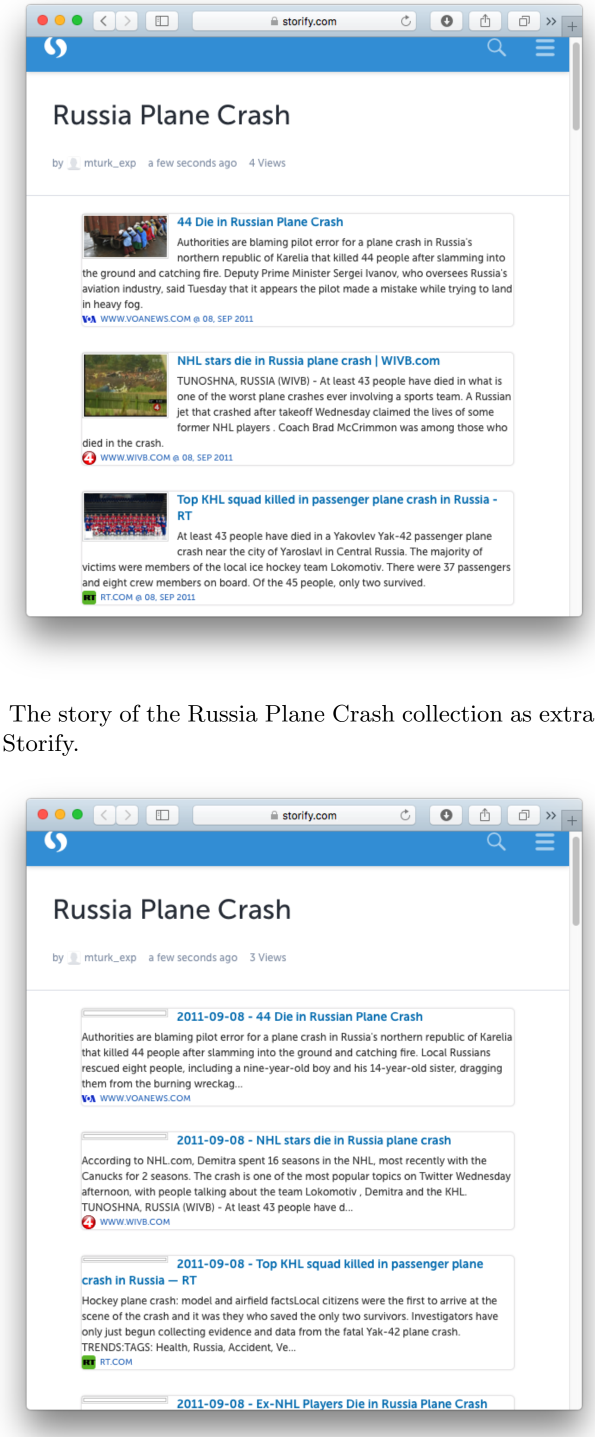
WWW.GUARDIAN.CO.UK

29, Nov 2010 - WikiLeaks Archive — Cables Uncloak U.S. Diplomacy

WikiLeaks posted 220 cables, some redacted to protect diplomatic

(b) The story of the Wikileaks collection as extracted by Newspaper.

FIG. 89: Example for Storify and and Newspaper extraction.



Russia Plane Crash

by mturk_exp a few seconds ago 4 Views

44 Die in Russian Plane Crash

Authorities are blaming pilot error for a plane crash in Russia's northern republic of Karelia that killed 44 people after slamming into the ground and catching fire. Deputy Prime Minister Sergei Ivanov, who oversees Russia's aviation industry, said Tuesday that it appears the pilot made a mistake while trying to land in heavy fog.

WWW.VOANEWS.COM @ 08, SEP 2011

NHL stars die in Russia plane crash | WIVB.com

TUNOSHNA, RUSSIA (WIVB) - At least 43 people have died in what is one of the worst plane crashes ever involving a sports team. A Russian jet that crashed after takeoff Wednesday claimed the lives of some former NHL players . Coach Brad McCrimmon was among those who died in the crash.

WWW.WIVB.COM @ 08, SEP 2011

Top KHL squad killed in passenger plane crash in Russia - RT

At least 43 people have died in a Yakovlev Yak-42 passenger plane crash near the city of Yaroslavl in Central Russia. The majority of victims were members of the local ice hockey team Lokomotiv. There were 37 passengers and eight crew members on board. Of the 45 people, only two survived.

RT.COM @ 08, SEP 2011

(a) The story of the Russia Plane Crash collection as extracted by Storify.

Russia Plane Crash

by mturk_exp a few seconds ago 3 Views

2011-09-08 - 44 Die in Russian Plane Crash

Authorities are blaming pilot error for a plane crash in Russia's northern republic of Karelia that killed 44 people after slamming into the ground and catching fire. Local Russians rescued eight people, including a nine-year-old boy and his 14-year-old sister, dragging them from the burning wreckag...

WWW.VOANEWS.COM

2011-09-08 - NHL stars die in Russia plane crash

According to NHL.com, Demitra spent 16 seasons in the NHL, most recently with the Canucks for 2 seasons. The crash is one of the most popular topics on Twitter Wednesday afternoon, with people talking about the team Lokomotiv , Demitra and the KHL. TUNOSHNA, RUSSIA (WIVB) - At least 43 people have d...

WWW.WIVB.COM

2011-09-08 - Top KHL squad killed in passenger plane crash in Russia — RT

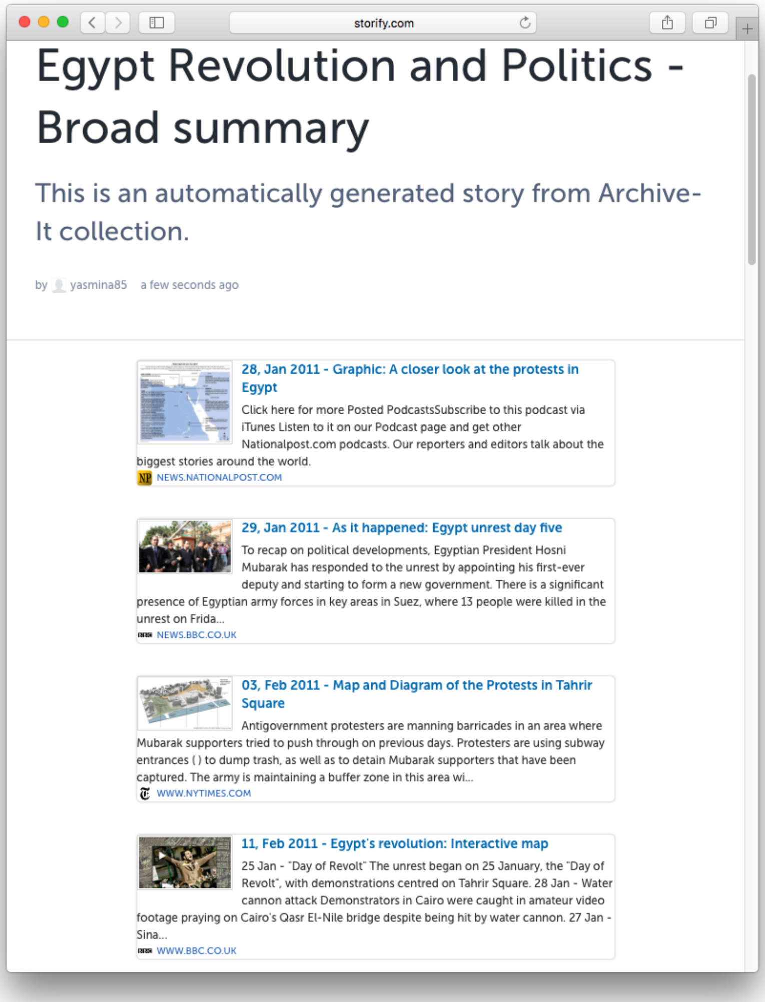
Hockey plane crash: model and airfield factsLocal citizens were the first to arrive at the scene of the crash and it was they who saved the only two survivors. Investigators have only just begun collecting evidence and data from the fatal Yak-42 plane crash. TRENDS:TAGS: Health, Russia, Accident, Ve...

RT.COM

2011-09-08 - Ex-NHL Players Die in Russia Plane Crash

(b) The story of the Russia Plane Crash collection as extracted by Newspaper.

FIG. 90: Storify extracts images better than the Newspaper library.



Egypt Revolution and Politics - Broad summary

This is an automatically generated story from Archive-It collection.

by yasmina85 a few seconds ago

28, Jan 2011 - Graphic: A closer look at the protests in Egypt

Click here for more Posted PodcastsSubscribe to this podcast via iTunes Listen to it on our Podcast page and get other Nationalpost.com podcasts. Our reporters and editors talk about the biggest stories around the world.

NEWS.NATIONALPOST.COM

29, Jan 2011 - As it happened: Egypt unrest day five

To recap on political developments, Egyptian President Hosni Mubarak has responded to the unrest by appointing his first-ever deputy and starting to form a new government. There is a significant presence of Egyptian army forces in key areas in Suez, where 13 people were killed in the unrest on Frida...

NEWS.BBC.CO.UK

03, Feb 2011 - Map and Diagram of the Protests in Tahrir Square

Antigovernment protesters are manning barricades in an area where Mubarak supporters tried to push through on previous days. Protesters are using subway entrances ( ) to dump trash, as well as to detain Mubarak supporters that have been captured. The army is maintaining a buffer zone in this area wi...

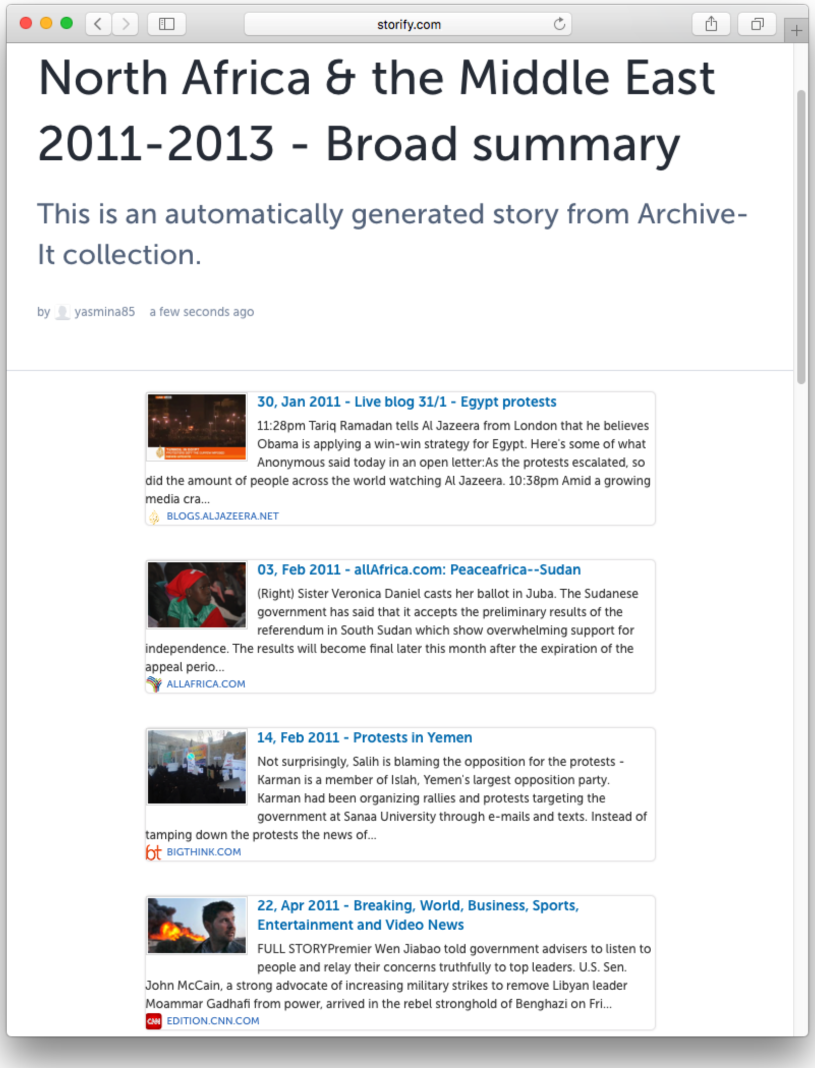
WWW.NYTIMES.COM

11, Feb 2011 - Egypt's revolution: Interactive map

25 Jan - "Day of Revolt" The unrest began on 25 January, the "Day of Revolt", with demonstrations centred on Tahrir Square. 28 Jan - Water cannon attack Demonstrators in Cairo were caught in amateur video footage praying on Cairo's Qasr El-Nile bridge despite being hit by water cannon. 27 Jan - Sina...

WWW.BBC.CO.UK

FIG. 91: The story of the Egyptian Revolution and politics collection.



North Africa & the Middle East 2011-2013 - Broad summary

This is an automatically generated story from Archive-It collection.

by yasmina85 a few seconds ago

30, Jan 2011 - Live blog 31/1 - Egypt protests

11:28pm Tariq Ramadan tells Al Jazeera from London that he believes Obama is applying a win-win strategy for Egypt. Here's some of what Anonymous said today in an open letter:As the protests escalated, so did the amount of people across the world watching Al Jazeera. 10:38pm Amid a growing media cra...

BLOGS.ALJAZEERA.NET

03, Feb 2011 - allAfrica.com: Peaceafrica--Sudan

(Right) Sister Veronica Daniel casts her ballot in Juba. The Sudanese government has said that it accepts the preliminary results of the referendum in South Sudan which show overwhelming support for independence. The results will become final later this month after the expiration of the appeal perio...

ALLAFRICA.COM

14, Feb 2011 - Protests in Yemen

Not surprisingly, Salih is blaming the opposition for the protests - Karman is a member of Islah, Yemen's largest opposition party. Karman had been organizing rallies and protests targeting the government at Sanaa University through e-mails and texts. Instead of tamping down the protests the news of...

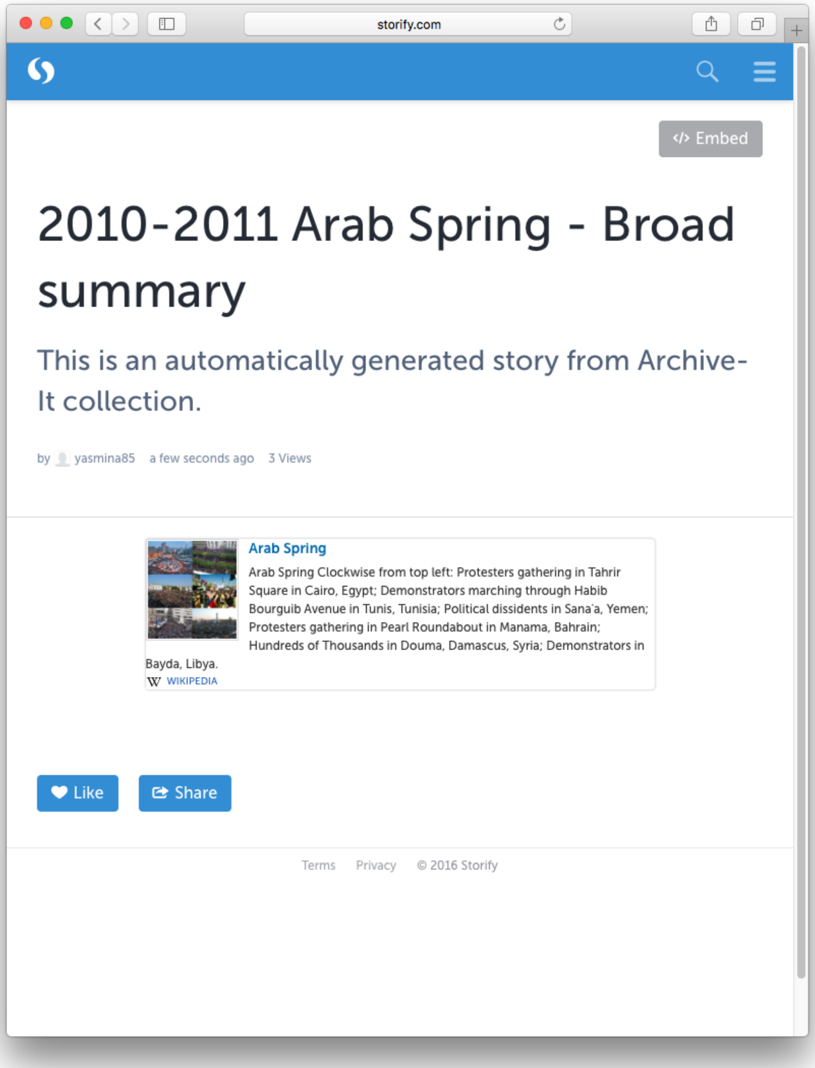
BIGTHINK.COM

22, Apr 2011 - Breaking, World, Business, Sports, Entertainment and Video News

FULL STORYPremier Wen Jiabao told government advisers to listen to people and relay their concerns truthfully to top leaders. U.S. Sen. John McCain, a strong advocate of increasing military strikes to remove Libyan leader Moammar Gadhafi from power, arrived in the rebel stronghold of Benghazi on Fri...

EDITION.CNN.COM

FIG. 92: The story of the North Africa & the Middle East 2011-2013.



2010-2011 Arab Spring - Broad summary

This is an automatically generated story from Archive-It collection.

by yasmina85 a few seconds ago 3 Views

Arab Spring

Arab Spring Clockwise from top left: Protesters gathering in Tahrir Square in Cairo, Egypt; Demonstrators marching through Habib Bourguib Avenue in Tunis, Tunisia; Political dissidents in Sana'a, Yemen; Protesters gathering in Pearl Roundabout in Manama, Bahrain; Hundreds of Thousands in Douma, Damascus, Syria; Demonstrators in Bayda, Libya.

W WIKIPEDIA

FIG. 93: The story of the 2010-2011 Arab Spring collection.

TABLE 25: The characteristics of the collections used for the evaluation.

Collection	ID	Timespan	URI-Rs	URI-Ms
2013 Boston Marathon Bombing	3649	2013/04/19 - 2015/03/03	318	1,907
Occupy Movement 2011/2012	2950	2011/12/03 - 2012/10/09	955	30,581
Egypt Revolution and Politics	2358	2011/02/01 - 2013/04/18	1,112	42,740
April 16 Archive	694	2007/05/23 - 2008/04/28	88	362
2013 Government Shutdown	3936	2013/10/22 - 2013/10/22	186	246
Russia Plane Crash Sept 2011	2823	2011/09/08 - 2011/09/15	104	558
Wikileaks 2010 Document Release Collection	2017	2010/07/27 - 2013/08/26	41	1,126
Earthquake in Haiti	1784	2010/01/20 - 2011/02/27	132	967
Brazilian School Shooting	2535	2011/04/09 - 2011/04/14	650	1,492
Global Health Events	4887	2014/10/01 - 2015/12/21	169	3,026

9.9 EVALUATING THE DARK AND STORMY ARCHIVE FRAMEWORK

In this section, we evaluate the automatically generated stories from archived collections. What makes a good story is a matter of human judgment and is difficult to evaluate. Inspired by the Turing Test [324], we use ground truth dataset of hand-crafted stories from Archive-It collections and let humans select between the human-generated stories and the automatically generated stories. We consider our method to be a success if humans are as likely to choose the automatically generated story as they do the human-generated story.

We asked expert archivists to generate hand-crafted stories from Archive-It collections, then used Amazon's Mechanical Turk[15] (MT) to evaluate the automatically generated stories against the stories that were created by experts. In the following sections, we will present the methodology and the results of evaluating the automatically generated stories from archived collections.

[15] https://www.mturk.com/

9.9.1 HAND-CRAFTED STORIES FROM ARCHIVED COLLECTIONS

We group Archive-It's collections into three main categories [22]. First, there are collections that are devoted to archiving governmental pages (e.g., all Web pages published by the State of South Dakota[16]). Second, there are collections that are event-based (e.g., Occupy Movement collection[17] and SOPA Blackout collection[18]). Third, there are theme-based collections (e.g., the Columbia Human Rights collection[19]).

We tested the DSA framework against event-based collections, which represent a significant portion of Archive-It collections. We asked expert archivists, with the help of the Archive-It team and Archive-It partners, to generate hand-crafted stories from Archive-It collections. We provided them with guideline documents that contained instructions for generating stories from Archive-It collections by selecting 28 representative mementos (more or less based on the collection size) that best represent each collection. We showed them the type of stories that can be generated. We also provided them the criteria for selecting the mementos. They suggested 10 different collections to generate stories from (see Table 25).

Criteria of the generated stories

The following is the list of the guidelines that we provided to the expert archivists for generating the stories:

- The representative mementos should be selected from within the collection. There should not be any memento from outside the collection.
- The default value for the number of selected mementos is $k \approx 28$. This value can be more or less based on the nature and size of each collection.
- We expect to have three generated stories out of each collection. Depending on the nature of the collection, some kind of stories may not be applicable. For those collections, please specify if any of the previous kinds of stories cannot be created.
- You can choose a specific time period for generating the story. If the collection spans many years, you can choose a subset of the timespan of the collection. For example, if you want to know the key events of the 25 Jan Egyptian Revolution during the 18 days of the protests in Egypt until Mubarak stepped down, you can choose pages from within the time range 2011/02/01-2011/02/14.

[16] https://archive-it.org/collections/192/

[17] https://archive-it.org/collections/2950/

[18] https://archive-it.org/collections/3010/

[19] https://archive-it.org/collections/1068/

We also put criteria for selecting the mementos:

- The language of the memento should be in English.
- The memento should be on-topic (the content is related to the topic of the collection).
- The memento should produce a visually attractive snippet on Storify, an article (`cnn.com/a/b/12/2015`) is more preferred than a homepage (`cnn.com`).
- The memento should not be a (near-)duplicate of another memento in the list.
- A better quality memento in terms of the missing resources is a better choice than a memento that is missing resources.

Methodology for Manually Generating a Story from Archived Collections

Along with the criteria of the stories and the selected mementos within each story, we illustrated to the Archive-It team the suggested possible types of stories that can be generated from each collection:

1. Sliding Pages, Sliding Time (SPST): broad summary of different URIs through time that provides an overview of the collection from different Web sites.
2. Sliding Page, Fixed Time (SPFT): different URIs at nearly the same time (for example, how the news covered Feb 11, 2011, when Mubarak stepped down) that provide different perspectives at a point in time.
3. Fixed Page, Sliding Time (FPST): same Web site at different times that provides an evolution of a single page (or domain) through time.

Note that in Chapter 5, we introduced four possible types of stories. However, with the current capabilities of Web archives, the Fixed Page, Fixed Time (FPFT) story cannot be supported because archives currently do not provide users with the ability to navigate representations by their environmental influences [168]. The domain experts provided us with lists of mementos for 23 different stories from the 10 different collections (see Table 26).

An example of a manually generated story by archivists from the Boston Marathon Bombing collection is shown in Figure 94. There were some collections that spanned a short period of time, so the archivists did not provide the FPST stories for these collections (for example, the “Brazilian School Shooting”, which spans over three days only). Another reason for not generating the FPST story is that none of the seeds of the collection change

TABLE 26: The breakdown of the stories that we received from domain experts.

Collection Name	**SPST**	**SPFT**	**FPST**	**No. of stories**
2013 Boston Marathon Bombing	✓	✓	✓	3
Occupy Movement 2011/2012	✓	✓	✓	3
Egypt Revolution and Politics	✓	✓	✓	3
April 16 Archive	✓	✓	✓	3
2013 Government Shutdown	✓	✓	-	2
Russia Plane Crash Sept 2011	✓	✓	-	2
Wikileaks 2010 Document Release Collection	✓	-	✓	2
Earthquake in Haiti	✓	-	✓	2
Brazilian School Shooting	✓	-	✓	2
Global Health Events	✓	-	-	1
Total no. of stories	10	6	7	23

over time (e.g., news articles). For example, the seed URIs of "Russia Plane Crash Sept 2011" collection are all news articles which do not evolve over time.

Table 27 shows the number of resources per story that were generated by experts and by the DSA framework (see Section 9.9.2).

9.9.2 AUTOMATICALLY GENERATED STORIES FROM ARCHIVED COLLECTIONS

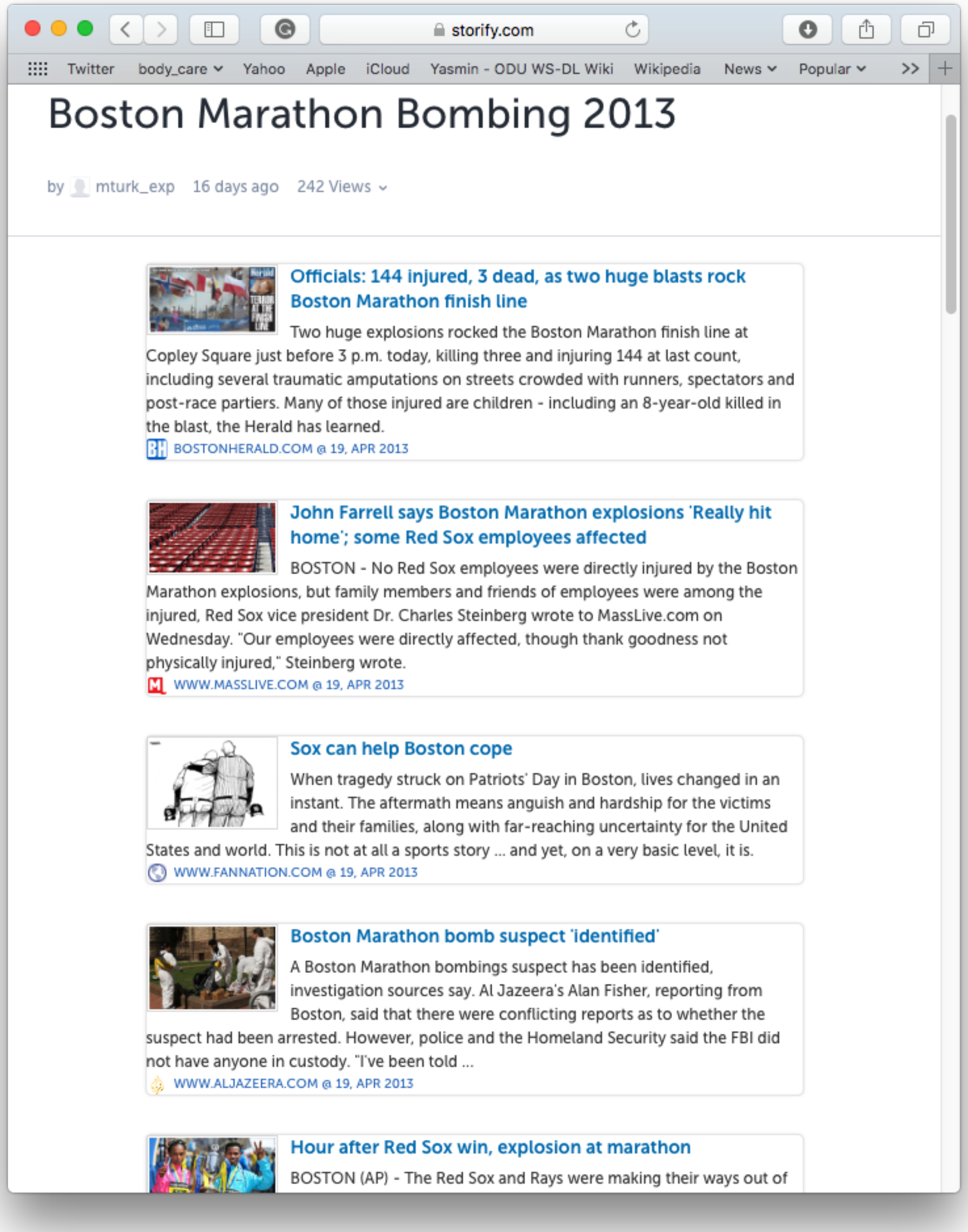
We applied the steps of the DSA framework that were introduced in Chapter 5 on the set of suggested collections in Table 25. We generated 23 stories from the collections. The SPST stories do not require any parameters because they represent a broad summary for the whole collection from all the seed URIs at different times. The FPST story and the SPFT story require input parameters such as URI-T for FPST stories and time frame for SPFT stories. In these stories, we feed DSA with the same parameters that were used in the human-generated stories (Table 26).

Dataset Preprocessing

We applied the following steps to generate stories from the Archive-It collections:

1. Obtain the seed list and the TimeMap of URIs from the front-end interface of Archive-It.

2. Extract the HTML of the mementos from the WARC files (locally hosted at ODU) and download the collections that we do not have in the ODU mirror from Archive-It.



Boston Marathon Bombing 2013

by mturk_exp 16 days ago 242 Views

Officials: 144 injured, 3 dead, as two huge blasts rock Boston Marathon finish line

Two huge explosions rocked the Boston Marathon finish line at Copley Square just before 3 p.m. today, killing three and injuring 144 at last count, including several traumatic amputations on streets crowded with runners, spectators and post-race partiers. Many of those injured are children - including an 8-year-old killed in the blast, the Herald has learned.

BOSTONHERALD.COM @ 19, APR 2013

John Farrell says Boston Marathon explosions 'Really hit home'; some Red Sox employees affected

BOSTON - No Red Sox employees were directly injured by the Boston Marathon explosions, but family members and friends of employees were among the injured, Red Sox vice president Dr. Charles Steinberg wrote to MassLive.com on Wednesday. "Our employees were directly affected, though thank goodness not physically injured," Steinberg wrote.

WWW.MASSLIVE.COM @ 19, APR 2013

Sox can help Boston cope

When tragedy struck on Patriots' Day in Boston, lives changed in an instant. The aftermath means anguish and hardship for the victims and their families, along with far-reaching uncertainty for the United States and world. This is not at all a sports story ... and yet, on a very basic level, it is.

WWW.FANNATION.COM @ 19, APR 2013

Boston Marathon bomb suspect 'identified'

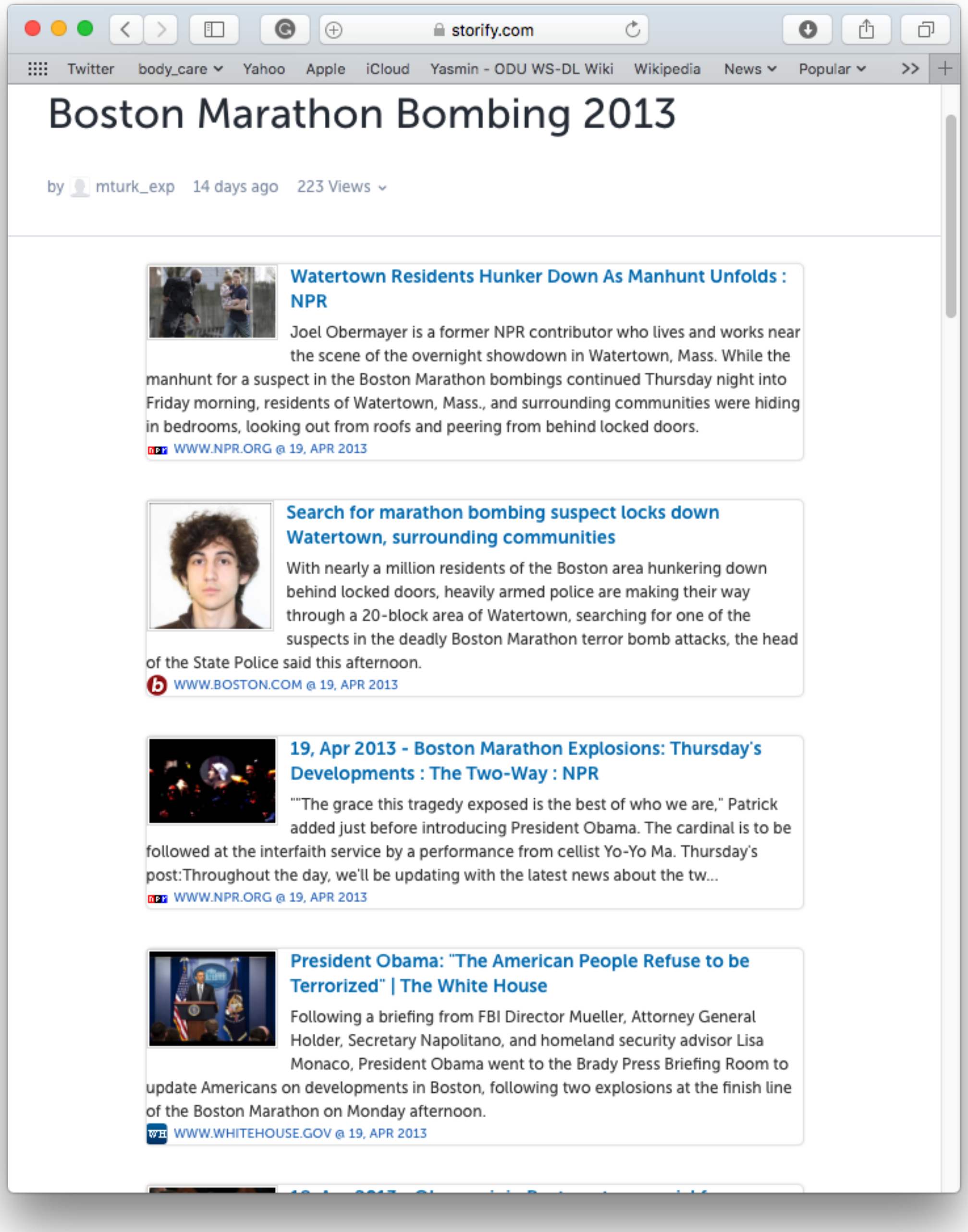
A Boston Marathon bombings suspect has been identified, investigation sources say. Al Jazeera's Alan Fisher, reporting from Boston, said that there were conflicting reports as to whether the suspect had been arrested. However, police and the Homeland Security said the FBI did not have anyone in custody. "I've been told ...

WWW.ALJAZEERA.COM @ 19, APR 2013

Hour after Red Sox win, explosion at marathon

BOSTON (AP) - The Red Sox and Rays were making their ways out of

FIG. 94: An example of a Sliding Page, Sliding Time story from the Boston Marathon Bombing collection that was generated by domain experts. Link: `https://storify.com/mturk_exp/3649b1s-57218803f5db94d11030f90b`



Boston Marathon Bombing 2013

by mturk_exp 14 days ago 223 Views

Watertown Residents Hunker Down As Manhunt Unfolds : NPR

Joel Obermayer is a former NPR contributor who lives and works near the scene of the overnight showdown in Watertown, Mass. While the manhunt for a suspect in the Boston Marathon bombings continued Thursday night into Friday morning, residents of Watertown, Mass., and surrounding communities were hiding in bedrooms, looking out from roofs and peering from behind locked doors.

WWW.NPR.ORG @ 19, APR 2013

Search for marathon bombing suspect locks down Watertown, surrounding communities

With nearly a million residents of the Boston area hunkering down behind locked doors, heavily armed police are making their way through a 20-block area of Watertown, searching for one of the suspects in the deadly Boston Marathon terror bomb attacks, the head of the State Police said this afternoon.

WWW.BOSTON.COM @ 19, APR 2013

19, Apr 2013 - Boston Marathon Explosions: Thursday's Developments : The Two-Way : NPR

""The grace this tragedy exposed is the best of who we are," Patrick added just before introducing President Obama. The cardinal is to be followed at the interfaith service by a performance from cellist Yo-Yo Ma. Thursday's post:Throughout the day, we'll be updating with the latest news about the tw...

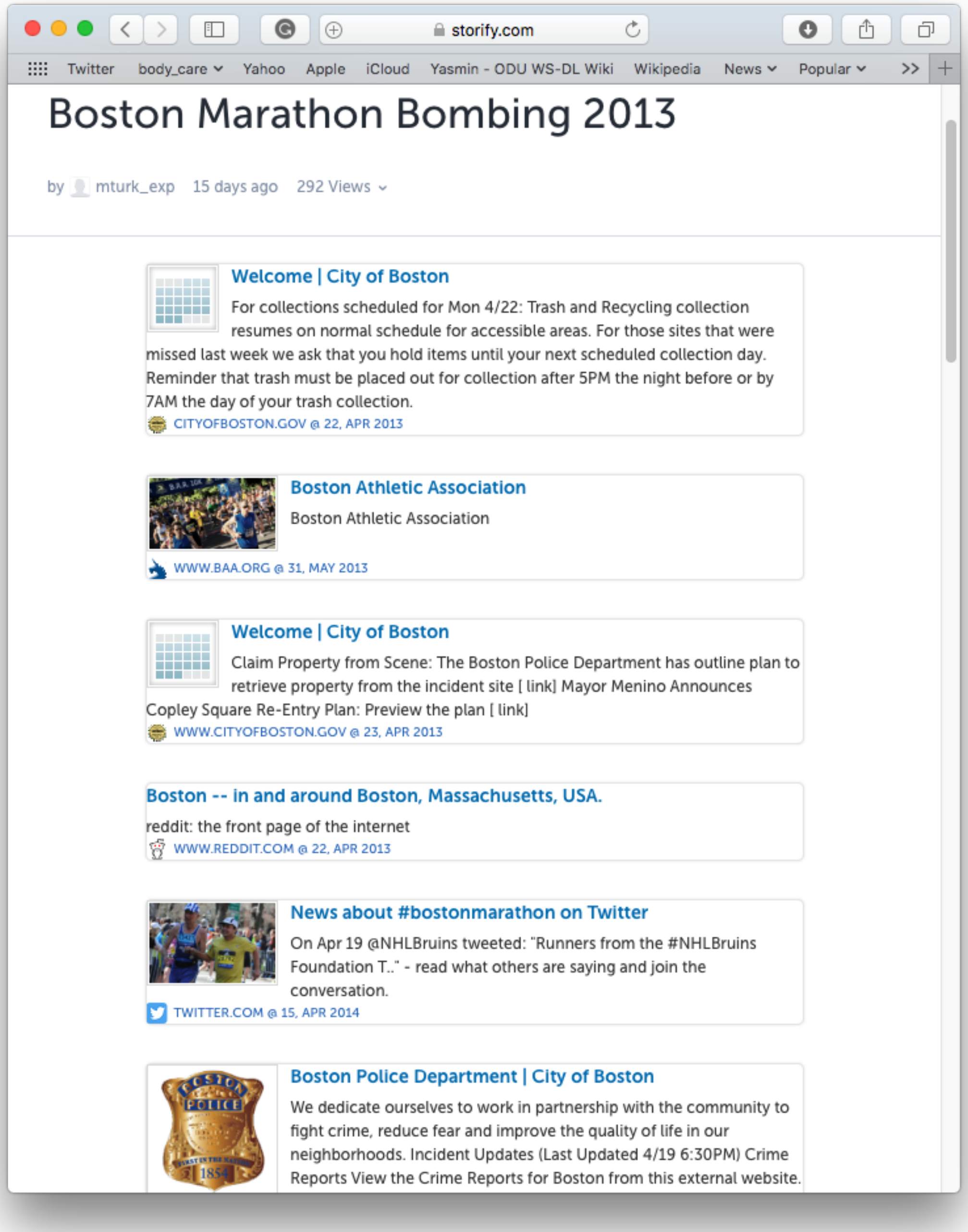
WWW.NPR.ORG @ 19, APR 2013

President Obama: "The American People Refuse to be Terrorized" | The White House

Following a briefing from FBI Director Mueller, Attorney General Holder, Secretary Napolitano, and homeland security advisor Lisa Monaco, President Obama went to the Brady Press Briefing Room to update Americans on developments in Boston, following two explosions at the finish line of the Boston Marathon on Monday afternoon.

WWW.WHITEHOUSE.GOV @ 19, APR 2013

FIG. 95: An example of a Sliding Page, Sliding Time story from the Boston Marathon Bombing collection that was generated automatically. Link: `https://storify.com/mturk_exp/3649b0s`



Boston Marathon Bombing 2013

by mturk_exp 15 days ago 292 Views

Welcome | City of Boston

For collections scheduled for Mon 4/22: Trash and Recycling collection resumes on normal schedule for accessible areas. For those sites that were missed last week we ask that you hold items until your next scheduled collection day. Reminder that trash must be placed out for collection after 5PM the night before or by 7AM the day of your trash collection.

CITYOFBOSTON.GOV @ 22, APR 2013

Boston Athletic Association

Boston Athletic Association

WWW.BAA.ORG @ 31, MAY 2013

Welcome | City of Boston

Claim Property from Scene: The Boston Police Department has outline plan to retrieve property from the incident site [link] Mayor Menino Announces Copley Square Re-Entry Plan: Preview the plan [link]

WWW.CITYOFBOSTON.GOV @ 23, APR 2013

Boston -- in and around Boston, Massachusetts, USA.

reddit: the front page of the internet

WWW.REDDIT.COM @ 22, APR 2013

News about #bostonmarathon on Twitter

On Apr 19 @NHLBruins tweeted: "Runners from the #NHLBruins Foundation T.." - read what others are saying and join the conversation.

TWITTER.COM @ 15, APR 2014

Boston Police Department | City of Boston

We dedicate ourselves to work in partnership with the community to fight crime, reduce fear and improve the quality of life in our neighborhoods. Incident Updates (Last Updated 4/19 6:30PM) Crime Reports View the Crime Reports for Boston from this external website.

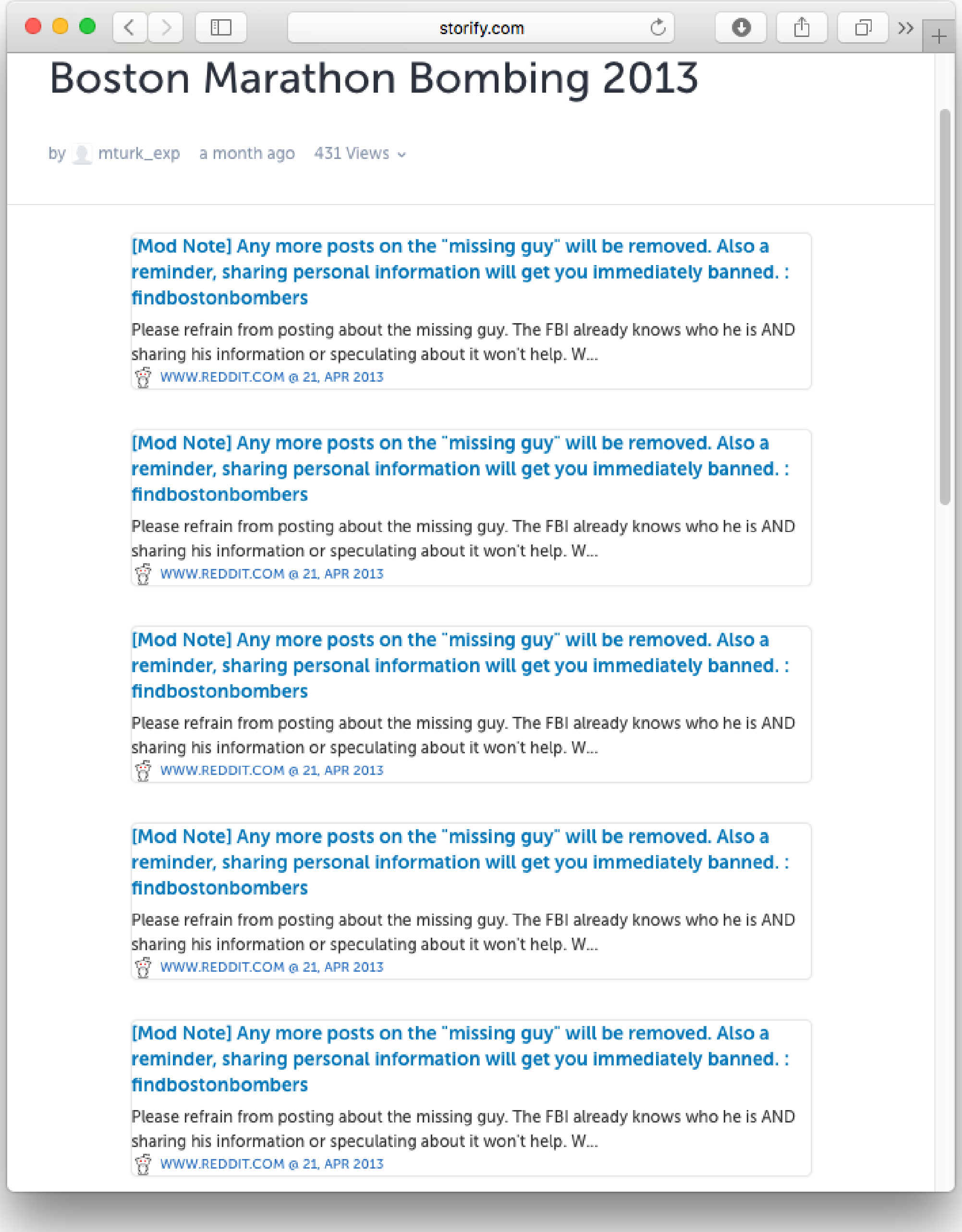
FIG. 96: An example of a Sliding Page, Sliding Time story from the Boston Marathon Bombing collection that was generated randomly. Link: `https://storify.com/mturk_exp/3649b2s-57227227bb79048c2d0388dc`

TABLE 27: The number of resources in the stories generated by domain experts and from the DSA framework.

	SPST		**SPFT**		**FPST**	
Collection	Human	Automatic	Human	Automatic	Human	Automatic
3649	28	29	28	25	7	5
2950	16	45	9	20	9	7
2358	16	20	11	17	12	7
694	17	32	14	19	5	4
3936	17	27	14	15	-	-
2823	28	25	27	23	-	-
2017	25	32	-	-	7	10
1784	28	34	-	-	11	14
2535	26	24	-	-	23	20
4887	36	34	-	-	-	-

3. Extract the text of the page using the Boilerpipe library [184].
4. Eliminate the off-topic pages based on the best-performing method ((Cosine, Word-Count) with the suggested thresholds (0.1, −0.85)), introduced in Chapter 8.
5. Exclude the duplicates of each TimeMap using the algorithm presented in Chapter 9.
6. Detect the language of the content using the language detection library created by Shuyo [289] and then eliminate the non-English language pages.
7. Slice the collection dynamically and then cluster the mementos of each slice using DBSCAN algorithm.
8. Apply the quality metrics introduced in Chapter 9 to select the best representative pages.
9. Sort the selected mementos chronologically then put them and their metadata in a JSON object (see Figure 87).

The number of the resources in the generated stories are presented in Table 27. Note that although the Egypt Revolution and Politics collection is the largest collection in the dataset, the resulting number of the resources for the Sliding Pages, Sliding Time story from this collection is just 20 mementos. That is because we selected the pages from within the same time frame (2011/02/01-2011/02/14) that was used for the human-generated story. An example of an automatically generated story by the DSA framework is illustrated in Figure 95.



Boston Marathon Bombing 2013

by mturk_exp a month ago 431 Views

[Mod Note] Any more posts on the "missing guy" will be removed. Also a reminder, sharing personal information will get you immediately banned. : findbostonbombers

Please refrain from posting about the missing guy. The FBI already knows who he is AND sharing his information or speculating about it won't help. W...

WWW.REDDIT.COM @ 21, APR 2013

[Mod Note] Any more posts on the "missing guy" will be removed. Also a reminder, sharing personal information will get you immediately banned. : findbostonbombers

Please refrain from posting about the missing guy. The FBI already knows who he is AND sharing his information or speculating about it won't help. W...

WWW.REDDIT.COM @ 21, APR 2013

[Mod Note] Any more posts on the "missing guy" will be removed. Also a reminder, sharing personal information will get you immediately banned. : findbostonbombers

Please refrain from posting about the missing guy. The FBI already knows who he is AND sharing his information or speculating about it won't help. W...

WWW.REDDIT.COM @ 21, APR 2013

[Mod Note] Any more posts on the "missing guy" will be removed. Also a reminder, sharing personal information will get you immediately banned. : findbostonbombers

Please refrain from posting about the missing guy. The FBI already knows who he is AND sharing his information or speculating about it won't help. W...

WWW.REDDIT.COM @ 21, APR 2013

[Mod Note] Any more posts on the "missing guy" will be removed. Also a reminder, sharing personal information will get you immediately banned. : findbostonbombers

Please refrain from posting about the missing guy. The FBI already knows who he is AND sharing his information or speculating about it won't help. W...

WWW.REDDIT.COM @ 21, APR 2013

FIG. 97: An example of a poorly generated story from the Boston Marathon Bombing collection to judge the selection of the turkers. Link: `https://storify.com/mturk_exp/3649bads`

9.9.3 RANDOM STORIES AND POOR STORIES

We selected $k \approx 28$ mementos randomly from the TimeMap of each collection as a baseline for evaluating the automatically generated stories (see Figure 96). The selection was done on the mementos in the collection before excluding the off-topic or the duplicates. The selected mementos were not sorted chronologically in the generated stories.

We use randomly generated stories to be compared against the human-generated stories and the automatically generated stories as a baseline for the generated stories by the DSA framework. In other words, we expect that the automatically generated stories will perform better than random stories against human-generated stories.

We generated poor stories by randomly selecting a memento from collection's TimeMap and repeated this memento 28 times. This story represents a control to ensure that the turkers do not choose randomly between the stories.

9.9.4 EXPERIMENT SETUP

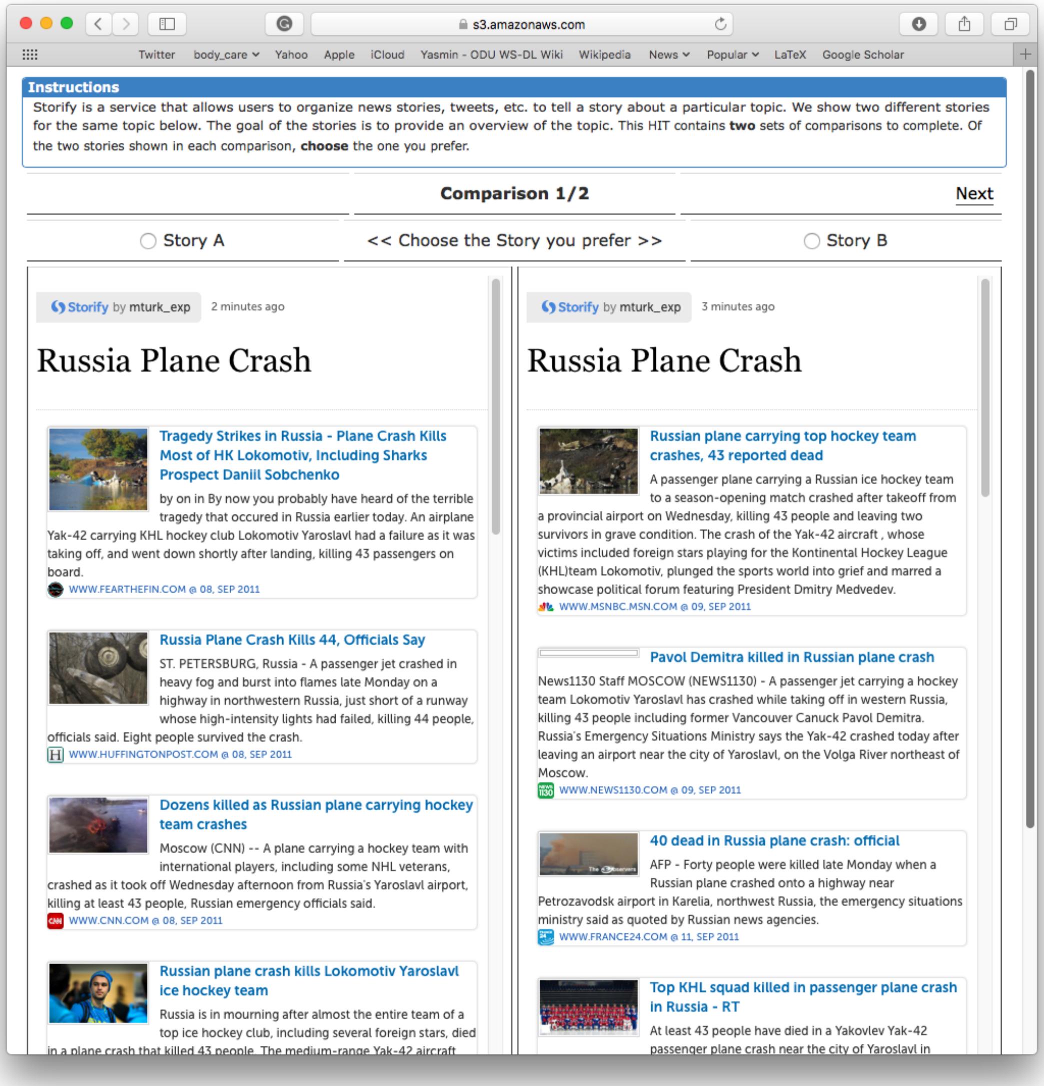
We used the same extraction methods for visualizing the human-generated stories (Figure 94), automatically generated stories (Figure 95), randomly generated stories (Figure 96), and poorly generated stories (Figure 97) on Storify.

Amazon's Mechanical Turk has been widely used for conducting user studies in a cost effective way in the context of time and money [187, 234, 331, 176, 131]. We use Mechanical Turk to compare four types of stories (human-generated, automatically generated, randomly generated, poorly generated), asking Mechanical Turk workers (or turkers) to choose between two stories at a time.

Our goal is to assess if the automatically generated stories by the DSA framework are indistinguishable from the human-generated stories. We provided turkers a description of a simple task to perform (a Human Intelligence Task, or HIT), choosing their preferred story (see Figure 98). We provided a simple generic description for the task as follows:

> *Storify is a service that allows users to organize news stories, tweets, etc. to tell a story about a particular topic. We show two different stories for the same topic below. The goal of the stories is to provide an overview of the topic. This HIT contains two sets of comparisons to complete. Of the two stories shown in each comparison,* ***choose*** *the one you prefer.*

Each HIT consists of two comparisons, in which one of the two comparisons was a control, a comparison between one of the stories and a poorly generated story. We reject the HITs where users selected a poorly generated story (i.e., a false positive selection).



Instructions

Storify is a service that allows users to organize news stories, tweets, etc. to tell a story about a particular topic. We show two different stories for the same topic below. The goal of the stories is to provide an overview of the topic. This HIT contains **two** sets of comparisons to complete. Of the two stories shown in each comparison, **choose** the one you prefer.

Comparison 1/2 Next

○ Story A << Choose the Story you prefer >> ○ Story B

Storify by mturk_exp 2 minutes ago

Russia Plane Crash

Tragedy Strikes in Russia - Plane Crash Kills Most of HK Lokomotiv, Including Sharks Prospect Daniil Sobchenko

by on in By now you probably have heard of the terrible tragedy that occured in Russia earlier today. An airplane Yak-42 carrying KHL hockey club Lokomotiv Yaroslavl had a failure as it was taking off, and went down shortly after landing, killing 43 passengers on board.

WWW.FEARTHEFIN.COM @ 08, SEP 2011

Russia Plane Crash Kills 44, Officials Say

ST. PETERSBURG, Russia - A passenger jet crashed in heavy fog and burst into flames late Monday on a highway in northwestern Russia, just short of a runway whose high-intensity lights had failed, killing 44 people, officials said. Eight people survived the crash.

WWW.HUFFINGTONPOST.COM @ 08, SEP 2011

Dozens killed as Russian plane carrying hockey team crashes

Moscow (CNN) -- A plane carrying a hockey team with international players, including some NHL veterans, crashed as it took off Wednesday afternoon from Russia's Yaroslavl airport, killing at least 43 people, Russian emergency officials said.

WWW.CNN.COM @ 08, SEP 2011

Russian plane crash kills Lokomotiv Yaroslavl ice hockey team

Russia is in mourning after almost the entire team of a top ice hockey club, including several foreign stars, died in a plane crash that killed 43 people. The medium-range Yak-42 aircraft

Storify by mturk_exp 3 minutes ago

Russia Plane Crash

Russian plane carrying top hockey team crashes, 43 reported dead

A passenger plane carrying a Russian ice hockey team to a season-opening match crashed after takeoff from a provincial airport on Wednesday, killing 43 people and leaving two survivors in grave condition. The crash of the Yak-42 aircraft , whose victims included foreign stars playing for the Kontinental Hockey League (KHL)team Lokomotiv, plunged the sports world into grief and marred a showcase political forum featuring President Dmitry Medvedev.

WWW.MSNBC.MSN.COM @ 09, SEP 2011

Pavol Demitra killed in Russian plane crash

News1130 Staff MOSCOW (NEWS1130) - A passenger jet carrying a hockey team Lokomotiv Yaroslavl has crashed while taking off in western Russia, killing 43 people including former Vancouver Canuck Pavol Demitra. Russia's Emergency Situations Ministry says the Yak-42 crashed today after leaving an airport near the city of Yaroslavl, on the Volga River northeast of Moscow.

WWW.NEWS1130.COM @ 09, SEP 2011

40 dead in Russia plane crash: official

AFP - Forty people were killed late Monday when a Russian plane crashed onto a highway near Petrozavodsk airport in Karelia, northwest Russia, the emergency situations ministry said as quoted by Russian news agencies.

WWW.FRANCE24.COM @ 11, SEP 2011

Top KHL squad killed in passenger plane crash in Russia - RT

At least 43 people have died in a Yakovlev Yak-42 passenger plane crash near the city of Yaroslavl in

FIG. 98: A sample HIT that shows two stories that turkers evaluate and select their preferred story. Each HIT contains two comparisons.

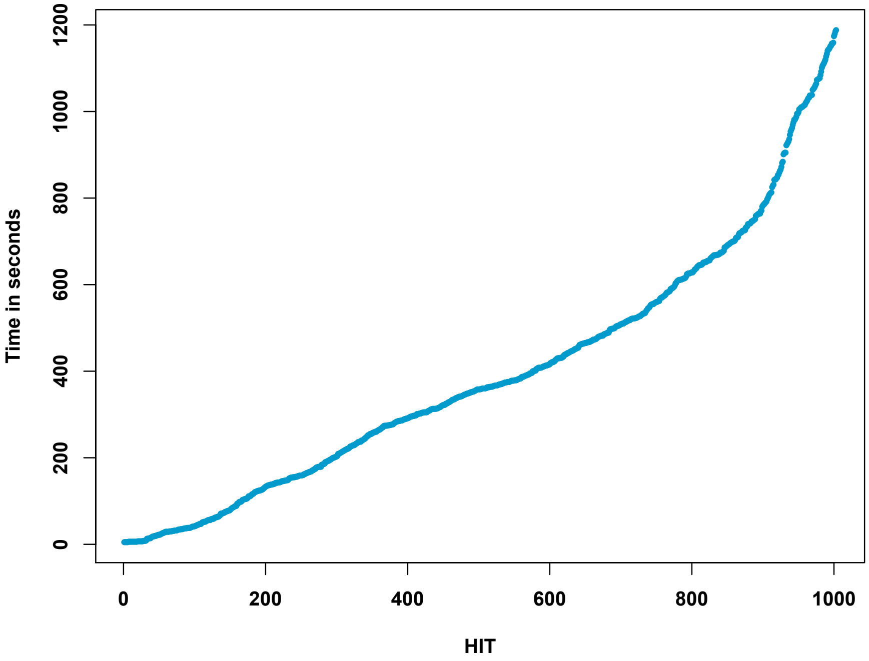


FIG. 99: A plot of the time taken by the turkers for submitting the HITs.

To reduce the cognitive load of the task, we assigned one comparison for each HIT along with the comparison that includes the poor story. Therefore, for evaluating one story, we have three HITs as follows:

HIT_1 : human vs. automatic, human vs. poor

HIT_2 : human vs. random, human vs. poor

HIT_3 : random vs. automatic, automatic vs. poor

We ensured that the position of each pair of composites was reversed among different stories to ensure there was not a bias in the HIT layout. We posted 69 HITs to evaluate 23 different stories. For each HIT, we required 15 turkers with "master" qualification requirements[20]. Based on many studies for deciding the number of participants in user studies, group sizes between eight and 25 are typically good numbers for conducting comparative studies [214, 296]. We chose to use 15 participants for each HIT in our experiment. We

[20] `https://www.mturk.com/mturk/help?helpPage=worker#what_is_master_worker`

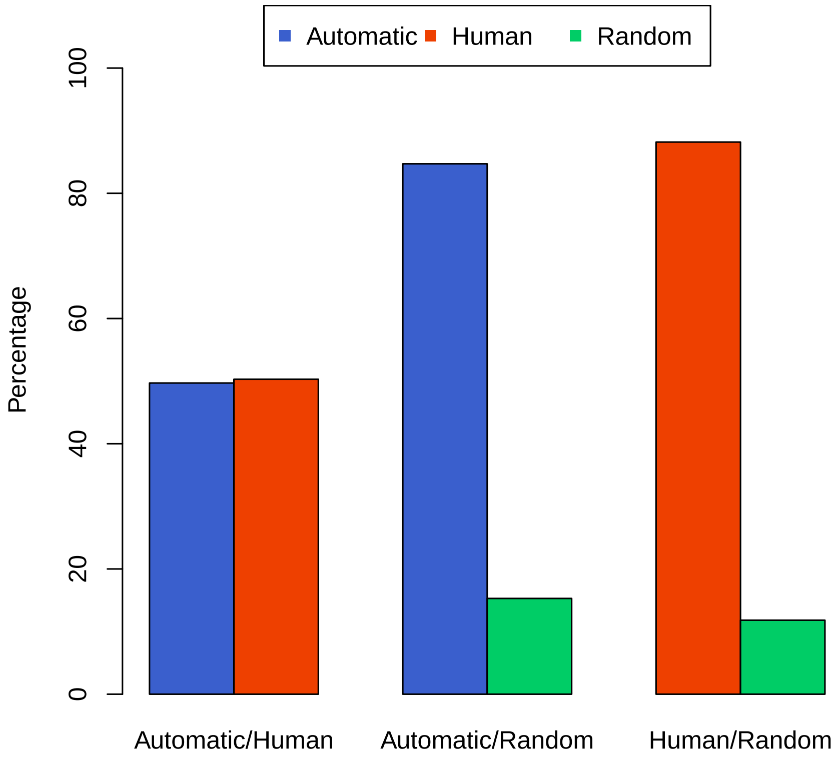


FIG. 100: The summary results of MT evaluation.

TABLE 28: The results of comparing human-generated stories versus automatically generated stories.

	Selections	Human	Automatic
SPST	142	50.7%	49.3%
SPFT	87	46.0%	54.0%
FPST	103	51.5%	48.5%

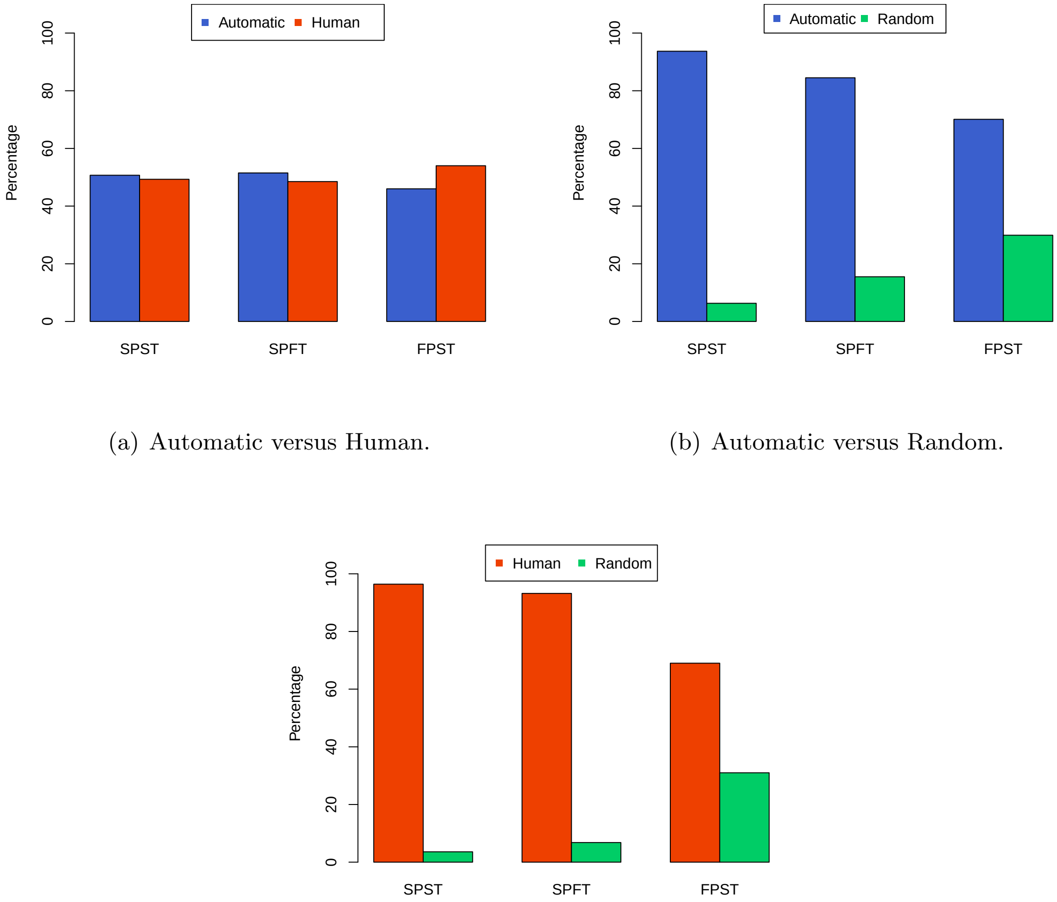
rejected the HITs in which the submissions contained poorly generated stories and the HITs that were completed in less than 10 seconds. TWe rejected a total of 46 HITs. In total, we had 989 out of 1,035 (69×15) valid HITs. These HITs were performed by 30 unique Master level turkers. We awarded the turker $0.50 per HIT. The turkers took seven minutes on average to complete the selections of the two comparisons. Figure 99 shows a plot of the time taken for submitting each HIT.

9.9.5 RESULTS

Figure 100 shows a summary of the results of the turkers selections for the three comparisons: human vs. automatic, random vs. automatic, and human vs. random. The results in Figure 100 shows that both the automatically generated stories and the human-generated were selected $\approx 50\%$ of the time. The figure also shows that the automatic stories are better than the randomly generated stories. Based on the results of the two-tailed t-test, we found that at confidence level 95% the automatically generated stories with $mean = 7.17$ of the votes are indistinguishable from the human-generated stories with $mean = 7.26$ ($p = 0.9134$, $t = 0.1094$, $df = 43.9$). However, at confidence level 95%, the automatically generated stories with $mean = 12.04$ and the human-generated stories with $mean = 12.65$ are significantly different from the random-generated stories with $mean \approx 2$ ($p < 2.2\text{e}{-16}$).

We zoom in on the results of the human-generated stories versus the automatically generated stories to interpret the results based on the different types of stories (SPST, SPFT, FPST). Table 28 shows that for all types of stories, the percentages of the turkers preferences to human and automatic stories are close. We applied a two-sided paired t-test on the samples based on the story type. We found that at confidence level 95% there is no significant difference ($p > 0.5$) between the human-generated stories and the automatically generated stories for all the types of the stories.

Figure 101 shows the breakdown results of the three comparisons based on the story type. Based on the t-test, we found that at confidence level 95%, for each of the three types stories, the results proved that automatically generated stories are significantly similar to the human-generated stories ($p > 0.5$) for all the types of the stories. However, the



(a) Automatic versus Human.

(b) Automatic versus Random.

(c) Random versus Human.

FIG. 101: The results of MT evaluation for each type of story.

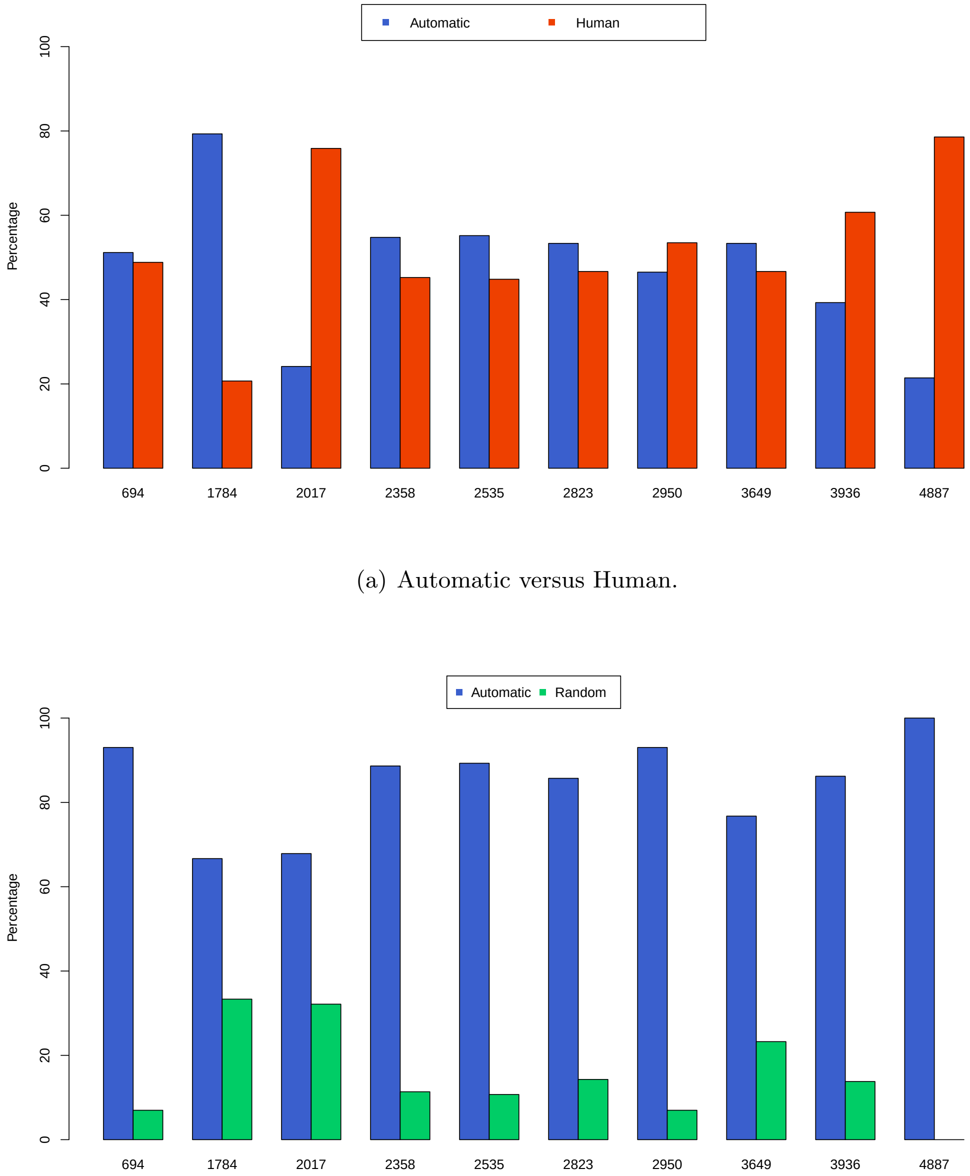
difference between the automatically generated stories and the random-generated story is statistically significant ($p < 0.001$) for all the types of stories at 95% confidence level. There is also a significance difference between the randomly generated stories and the human-generated stories ($p < 0.001$) at 95% confidence level.

We show the results of the turkers' preferences for the three selections for each collection in Figures 102 and 103. Figure 102(a) shows that for most of the collections, the automatically generated stories are indistinguishable from the human-generated stories. There are two collections that human-generated stories were selected more than the automatically generated stories: the "Wikileaks Document Release (2017)" and "Global Health Events". The automatically generated stories for the "Earthquake in Haiti" were preferred by turkers. Further investigation with more collections is required to test if the type of collections affects a human's selection.

9.10 SUMMARY

In this chapter, we described the general methodology for addressing the research question "How to select k mementos that represent a story?". We started with an algorithm for eliminating the (near-)duplicates in Web archives. Then, we provided an algorithm that dynamically slices the collection and divides the pages equally on the number of slices. We introduced a slowly-growing function to specify the number of the pages in the story to be close to 28 mementos (more or less based on the collection size). We also introduced multiple quality metrics for selecting the pages that compose a story, then we put the selected pages into chronological order and generated a JSON object to visualize them using Storify.

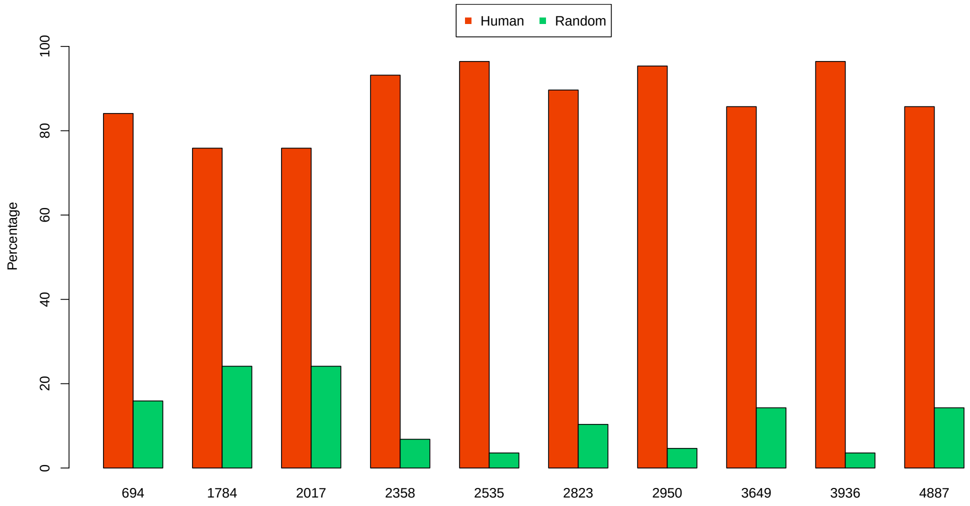
We evaluated the stories generated by the DSA framework in the rest of the chapter. We obtained a ground truth dataset of 23 stories that were generated manually from 10 Archive-It collections by expert archivists. We used Amazon's MT to compare the automatically generated stories with the human-generated stories. Based on 332 comparisons by 30 unique turkers between human-generated stories and automatic stories, the results showed that at confidence level 95%, the automatically generated stories are indistinguishable from the human-generated stories ($p > 0.5$). We also created random stories as a baseline for the automatic stories. The results show that the turkers were able to distinguish the random stories from the automatic and the human stories ($p < 0.001$).



(a) Automatic versus Human.

(b) Automatic versus Random.

FIG. 102: The results of MT evaluation for each collection.



(a) Random versus Human.

FIG. 103: The results of MT evaluation for each collection (continued).

CHAPTER 10

CONTRIBUTIONS, FUTURE WORK, AND CONCLUSIONS

In this chapter, we revisit the research questions with the work that has been done for investigating each question. We will also present the contributions, the future work, and conclusions of this research.

10.1 RESEARCH QUESTIONS REVISITED

RQ1. How do people browse the past Web? One of the concerns in the Web archiving world is how to generate more interest in and use of Web archives. To form our foundation in using the archives, we investigated how users access Web archives based on analyzing the user access logs of the IA's Wayback Machine [18]. We investigated the differences between human and robot accesses of the Wayback Machine, identified four major Web archive access patterns (Dip, Slide, Dive, and Skim), and uncovered the temporal preference for Web archive access (Chapter 4). We found that people come to Web archives because they did not find the pages on the live Web, and likely not because of lengthy browsing sessions of the past Web. Although the IA's Wayback Machine receives a significant amount of traffic, we found that robots outnumber humans 10:1 in terms of sessions.

We checked what users are looking for, why they come to Web archives, where they come from, and how pages link to Web archives [16, 17]. Based on the analysis of referring pages of human users we investigated how humans discover the Wayback Machine, why the referrers link to Web archives, and how they link to Web archives. We found that most human users come to the Wayback Machine via links or direct address presumably because they did not find the requested pages on the live Web. Of the requested archived pages, 65% do not currently exist on the live Web. From analyzing the referrers, we found that more than 82% of human sessions have referrers while only 15% of robot sessions have referrers.

RQ2. Can we automatically generate stories that convey different perspectives of the collection? The culmination of this body of work is the framework for generating stories from the archived collection automatically. It may be possible for a collection to be summarized with more than one kind of story (depending on the nature of the collection as well as the curators' preference), for example, a broadly defined story that samples from different URIs and different times, different URIs at approximately the same time, the same URI at different times, and the same URI at the same time.

We presented the abstract model and the components of the framework along with the definitions in Chapter 5. The framework has three main components: establishing a baseline by quantifying stories in Storify and collections in Archive-It to understand the measurables of both stories and collections, as generated by humans; reducing the candidate pool of archived pages by excluding the irrelevant pages to the topic of the collection, excluding the (near-)duplicates, and excluding the non-English language pages; selecting and evaluating good representative pages by slicing the collection dynamically, clustering the pages of each slice, selecting the best representative page from each cluster based on quality metrics we proposed, and then placing the selected pages in chronological order to be visualized by Storify.

RQ3. How do we build quantitative, descriptive models of human-generated stories and collections in Archive-It? To support automatic story creation, we needed to understand the structural characteristics of popular human-generated stories. We determined the characteristics of the human-generated stories based on a study of stories from Storify. The characteristics we identified included the mean and median length of resources in the stories, the nature of the resources, how quickly do the resources linked to from stories become unavailable (HTTP 404), and the popularity of the resources linked to from stories to (e.g., popular like `cnn.com` or little-known outlets, blogs, and other sites) [19, 21]. We established structural features for what differentiates popular stories from normal stories for building a baseline for the stories we will automatically generate from the archives (Chapter 6). We found that the popular stories have a median value of 28 elements, which will inform our framework for generating stories from archived collections that will be composed of a number of resources that is close to 28.

We also determine the characteristics of Archive-It collections by providing measurements for the statistics of all Archive-It collections such as the number of URIs, the number of mementos, the most used resources in these collections, the average timespan of the collections, etc. (Chapter 7). In summarizing a collection, we can only choose from what is archived. Although some content in Storify stories will not be applicable (e.g., `twitter.com` is popular in Storify, but mostly missing in Archive-It collections), some other characteristics will be applicable, such as the number of resources. Accordingly, our choices of what to select from the collection are informed by what constitutes a popular story.

We contrasted the created descriptive models of the created stories on social media storytelling service and the collections in Archive-It explaining the similarities and the differences between the human-generated stories and the archived collections.

RQ4. How to detect the off-topic Web pages in the archives? Our work toward establishing a framework to create stories from archived collections then combine them with social media begins with filtering the Web archive collections from the off-topic pages. We proposed and evaluated different methods (Cosine similarity, Jaccard similarity, intersection of the 20 most frequent terms, Web-based kernel function, and the change in size using a number of words and content length) at different thresholds to detect when the page has gone off-topic through the subsequent captures [20, 22]. We built a gold standard dataset from three different collections to evaluate the proposed methods. Those predicted off-topic pages will be presented to the collection's curator for possible elimination from the collection or cessation of crawling and not considered for inclusion in stories. We found that combining cosine similarity at threshold 0.10 and change in size using word count at threshold -0.85 performs the best with accuracy $= 0.987$, F_1 score $= 0.906$, and AUC $= 0.968$ (Chapter 8). We evaluated the performance of the proposed method on several Archive-It collections. The average precision of detecting off-topic pages in the collections is 0.92.

We also identified five different behaviors of changing the aboutness of TimeMaps: Always On, Step Function On, Step Function Off, Oscillating, and Always Off. We quantified each behavior based on a gold standard dataset. These behaviors will inform curators of the different cases of TimeMaps they may have in their collections. Furthermore, they inform us on the challenges of detecting the off-topic pages.

RQ5. How do we identify, evaluate, and select candidate (archived) Web pages to support the story? After we built a baseline and decreased the candidate pool of archived pages, we applied several steps on the rest of the mementos to select the best representative set of mementos. We provided a dynamic slicing algorithm to select from all the parts of the collections equally. Then, we clustered the mementos in each slice based on their contents (Chapter 9).

We proposed several metrics to measure memento quality M_q. Based on studying how Storify visualizes different kinds of pages, we defined two metrics that affect the snippet quality: the URI level (deep URI or high-level URI) and the type of the page (e.g., social media, news article, etc). We adopt Brunelle's algorithm [58] for assessing memento damage as another criteria for choosing the page. We also defined different methods for extracting the metadata of the pages to be visualized by Storify.

10.2 CONTRIBUTIONS

We developed techniques to automatically (with optional human review and “steering”) sample pages from a collection that summarize and describe the collection. For example, given a collection of thousands of pages, the DSA framework will automatically select $k \approx 28$ representative pages that will then be linked in storytelling Web services, such as Storify. This dissertation makes ten significant contributions to the field of digital preservation:

1. The basic building blocks (Dip, Slide, Dive, and Skim) of user access patterns in Web archives were introduced through an analysis of the Internet Archive’s Wayback Machine access logs [18]. We also quantified the patterns differentiating robot from human accesses (Chapter 4).

2. We studied of the requests of Web archive users, both humans and robots, to gain insight into what users look for, in the context of the language of the requested pages [16, 17]. We provided an analysis of referring pages of human users to investigate how humans discover the Wayback Machine, why the referrers link to Web archives, and how they link to Web archives (Chapter 4).

3. We proposed different methods to detect when the page has gone off-topic through the subsequent captures [20, 22]. Those predicted off-topic pages will be presented to the collection’s curator for possible elimination from the collection or cessation of crawling (Chapter 8).

4. A gold standard dataset from three different Archive-It collections was created by labeling thousands of archived pages to test different methods for detecting the off-topic pages in Web archives (Chapter 8).

5. We created a command line service that helps curators to detect the off-topic pages then present them to the curator to decide about their relevancy (Chapter 8). The code and gold standard dataset are available at `https://github.com/yasmina85/OffTopic-Detection`.

6. Five different behaviors of changing the aboutness of TimeMaps in Web archives were identified [22]: Always On, Step Function On, Step Function Off, Oscillating, and Always Off (Chapter 8).

7. To support automatic story creation, we built a quantitative, descriptive model of stories that were created manually by Storify users and focusing on the structural characteristics of popular (i.e., receiving the most views) stories [19] (Chapter 6).

8. A baseline for the characteristics of archived collections was presented based on analyzing the whole population of Archive-It collections [21] (Chapter 7).
9. We presented the DSA framework, in which we identify, evaluate, and select candidate mementos to support the events of the stories (Chapter 5).
10. We introduced a ground truth dataset for the human-generated stories, which we evaluate the automatic stories against using human evaluation (Chapter 9).
11. A command line service was created to automatically generate different kinds of stories from archived collections. The code and gold standard dataset are available at `https://github.com/yasmina85/DSA-stories`.

10.3 FUTURE WORK

We believe that Web archives need services to help users and researchers understand the tremendous amount of cultural heritage that Web archives hold. Adopting the DSA framework will help users to understand the important resources of the archived collections. Our future work will focus on helping archivists to integrate our DSA framework into Archive-It to help the curators to discover the non-relevant materials in their collections and generate summaries from these collections. These summaries may attract Web users and help them understand the holding of these collections.

Our future work will continue to improve the framework by integrating a component to recommend to collection curators' new seed URIs that are relevant to the aboutness of the collections. Because different sources provide different URIs for the story with different perspectives, we plan to use different sources for detecting new seed URIs, such as Google search, social media, and the list of references on Wikipedia pages.

We also provided a preliminary investigation of automatically detecting off-topic pages in Web archives. The off-topic detection methods presented in the DSA were able to detect off-topic pages within the context of a single TimeMap. We generated our framework with the assumption that the first memento is on-topic. The next step is to compute the aboutness of the whole collection and compare the aboutness of the mementos to the aboutness of the collection, in part to more easily detect the off-topic pages in the "Always Off" and "Step Function Off" TimeMaps.

We provided preliminary evaluation for the stories generated by the DSA framework. Although the humans were not able to distinguish the automatically generated stories from the human-generated stories, future research should investigate the usefulness of the generated stories and evaluate the discovery tasks for people given the summarized stories.

For example, if we generate 17 stories from the 17 human rights collections that exist in Archive-It, we need to conduct user studies to evaluate if a user can tell which collection is about women's human rights, or which collection is about human rights in Africa. Furthermore, we plan also to collaborate with humanities researchers to conduct user studies on important events, e.g., the Arab Spring, and check if a specific kind of story provides the best insight into the events and the corresponding collections. For example, how do the Sliding Page, Fixed Time stories help humanities researchers to get different perspectives about news coverage and how much time is saved from manual search by providing them this kind of story.

10.4 CONCLUSIONS

Many conversations and important events now start on the Web. The growth rate of content creation in the digital world is exploding incredibly. Unfortunately, the nature of the Web is ephemeral, and the expected lifetime of a Web page is short. This can cause access to the 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. Additionally, archiving Web pages into themed collections is a method for ensuring these resources are available for posterity. Many institutions archive the Web, resulting in tremendous amount of archive pages that have thousands of mementos.

Even though the existence of Web archives can fulfill this important function, we saw from our analysis of user access logs of the largest and oldest Web archive, the IA's Wayback Machine, that Web archives are underutilized by human users. We found that although the Internet Archive receives a lot of traffic, robots outnumber humans 10:1 in the Wayback Machine. Furthermore, the humans that visit the Internet Archive's Wayback Machine typically visit a single page and then leave; depending on the source this can be as often as 64% of the time. In short, Web archives are not well-known by the general Web population (and are not indexed by search engines), and those who do know about Web archives consider them difficult to use.

Our objective is to provide creative and easy approaches to the normal users to browse, explore, and understand the born-digital materials. Furthermore, the curator will have assistance for summarizing the holdings of the archived collections automatically by identifying, evaluating, and selecting candidate Web pages from archived collections. The candidate pages then are used for generating stories that summarize the holdings of the archived collections collection, then arrange these pages in a narrative structure ordered by time and visualize these stories using Storify. Curators will have the option to update the generated stories based on their preference. For example, if there are specific URIs the curators prefer to exclude from the generated story, they can do this. These stories will be a bridge between the current and past Web. They will provide people with multiple perspectives about important events using tools they are already familiar with, such as Storify, and the resources of the generated stories will be persistent.

Using the DSA framework, my son can easily find what he needs to know about the Egyptian Revolution as it happened in the past, in addition to being able to define where to start and which collection will give an insight about the Egyptian Revolution. For example, the resulting stories in Figures 91, 92, and 93 will give him an idea about the holdings of each collection in which each story summarizes and he can decide on the collection that is only about the Egyptian Revolution. In addition to that, the generated story will give him insight into the key events of the Egyptian Revolution.

REFERENCES

[1] Archive.is. http://archive.is/

[2] Category:All articles with dead external links. Wikipedia, http://en.wikipedia.org/w/index.php

[3] Internet Archive. https://archive.org

[4] Microsoft Silverlight. http://www.microsoft.com/silverlight/

[5] PivotViewer. http://www.microsoft.com/silverlight/pivotviewer/

[6] UK Web Archive. http://www.webarchive.org.uk/ukwa/

[7] Web Archiving Services. http://webarchives.cdlib.org/

[8] Wikipedia:Using the Wayback Machine. Wikipedia, http://en.wikipedia.org/wiki/Wikipedia:Using_the_Wayback_Machine

[9] Frequently Asked Questions. https://archive.org/about/faqs.php#21 (2011)

[10] Wikimedia Report Card. http://reportcard.wmflabs.org/ (2014)

[11] Adams, I., Miller, E.L., Storer, M.W.: Analysis of Workload Behavior in Scientific and Historical Long-Term Data Repositories. Tech. Rep. UCSC-SSRC-11-01, University of California, Santa Cruz (2011)

[12] Ahlberg, C., Shneiderman, B., Shneidennan, B.: Visual information seeking: tight coupling of dynamic query filters with starfield displays. In: Proceedings of ACM SIGCHI, pp. 313–317 (1994). DOI 10.1145/191666.191775

[13] Ainsworth, S.G., AlSum, A., SalahEldeen, H., Weigle, M.C., Nelson, M.L.: How Much of the Web Is Archived? In: Proceeding of the 11th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '11, pp. 133–136. ACM Press (2011). DOI 10.1145/1998076.1998100

[14] Ainsworth, S.G., Nelson, M.L.: Evaluating sliding and sticky target policies by measuring temporal drift in acyclic walks through a web archive. International Journal on Digital Libraries **16**(2), 129–144 (2014). DOI 10.1007/s00799-014-0120-4

[15] Allan, J.: Introduction to topic detection and tracking. In: J. Allan (ed.) Topic Detection and Tracking, *The Information Retrieval Series*, vol. 12, pp. 1–16. Springer US (2002). DOI 10.1007/978-1-4615-0933-2_1

[16] AlNoamany, Y., AlSum, A., Weigle, M.C., Nelson, M.L.: Who and What Links to the Internet Archive. In: Proceedings of the 17th International Conference on Theory and Practice of Digital Libraries, *TPDL '13*, vol. 8092, pp. 346–357. Springer International Publishing (2013). DOI 10.1007/978-3-642-40501-3_35

[17] AlNoamany, Y., AlSum, A., Weigle, M.C., Nelson, M.L.: Who and What Links to the Internet Archive. International Journal on Digital Libraries **14**(3), 101–115 (2014). DOI 10.1007/s00799-014-0111-5

[18] AlNoamany, Y., Weigle, M.C., Nelson, M.L.: Access Patterns for Robots and Humans in Web Archives. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '13, pp. 339–348. ACM (2013). DOI 10.1145/2467696.2467722

[19] AlNoamany, Y., Weigle, M.C., Nelson, M.L.: Characteristics of Social Media Stories. In: Proceedings of the 19th International Conference on Theory and Practice of Digital Libraries, TPDL '15, pp. 267–279. Springer International Publishing (2015). DOI 10.1007/978-3-319-24592-8_20

[20] AlNoamany, Y., Weigle, M.C., Nelson, M.L.: Detecting Off-Topic Pages in Web Archives. In: Proceedings of the 19th International Conference on Theory and Practice of Digital Libraries, TPDL '15, pp. 225–237. Springer International Publishing (2015). DOI 10.1007/978-3-319-24592-8_17

[21] AlNoamany, Y., Weigle, M.C., Nelson, M.L.: Characteristics of Social Media Stories. What makes a good story? International Journal on Digital Libraries (2016). DOI 10.1007/s00799-016-0185-3

[22] AlNoamany, Y., Weigle, M.C., Nelson, M.L.: Detecting Off-Topic Pages Within TimeMaps in Web Archives. International Journal on Digital Libraries (2016). DOI 10.1007/s00799-016-0183-5

[23] AlSum, A.: mcurl - Command Line Memento Client. `http://ws-dl.blogspot.com/2013/05/2013-05-29-mcurl-command-line-memento.html` (2013)

[24] AlSum, A.: Web Archive's Services Framework for Tighter Integration between the Past and Present Web. Dissertation, Old Dominion University (2014)

[25] AlSum, A., Nelson, M.L.: ArcLink: Optimization Techniques to Build and Retrieve the Temporal Web Graph. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '13, pp. 377–378. ACM Press (2013). DOI 10.1145/2467696.2467751

[26] AlSum, A., Nelson, M.L.: ArcLink: Optimization Techniques to Build and Retrieve the Temporal Web Graph. Tech. Rep. arXiv:1305.5959 (2013)

[27] AlSum, A., Nelson, M.L.: Thumbnail Summarization Techniques for Web Archives. In: Proceedings of the 36th European Conference on Information Retrieval, ECIR '14, pp. 299–310 (2014). DOI 10.1007/978-3-319-06028-6_25

[28] AlSum, A., Weigle, M., Nelson, M., Sompel, H.: Profiling Web Archive Coverage for Top-Level Domain and Content Language. In: Proceedings of the 17th International Conference on Theory and Practice of Digital Libraries, *Lecture Notes in Computer Science*, vol. 8092, pp. 60–71. Springer Berlin Heidelberg (2013)

[29] Andrew Kehoe, M.G.: Social Tagging: A new perspective on textual 'aboutness'. Studies in Variation, Contacts and Change in English **6** (2011). URL http://www.helsinki.fi/varieng/series/volumes/06/kehoe_gee/

[30] Antonio, A., Tuffley, D., Martin, N.: Creating engagement and cultivating information literacy skills via Scoop.it. In: Proceedings of the 30th Australasian Society for Computers in Learning in Tertiary Education Conference (ASCILITE 2013), pp. 52–62 (2013)

[31] Ardizzone, E., Cascia, M.L.: Automatic video database indexing and retrieval. Multimedia Tools Applications **4**(1), 29–56 (1997). DOI 10.1023/A:1009630331620

[32] Arms, W.Y., Aya, S., Dmitriev, P., Kot, B.J., Mitchell, R., Walle, L.: Building a Research Library for the History of the Web. In: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '06, pp. 95–102 (2006). DOI 10.1145/1141753.1141771

[33] Arthur, C.: Egypt cuts off internet access. The Guardian, https://www.theguardian.com/technology/2011/jan/28/egypt-cuts-off-internet-access (2011)

[34] Ashenfelder, M.: The Average Lifespan of a Webpage. http://blogs.loc.gov/digitalpreservation/2011/11/the-average-lifespan-of-a-webpage/ (2011)

[35] Aye, T.T.: Web log cleaning for mining of web usage patterns. In: Proceedings of the 3rd International Conference on Computer Research and Development, pp. 490–494. IEEE (2011). DOI 10.1109/ICCRD.2011.5764181

[36] Bailey, J.: IMLS National Digital Platform Grant Awarded to Advance Web Archiving. Internet Archive Blogs, https://blog.archive.org/2015/10/08/

imls-national-digital-platform-grant-awarded-to-advance-web-archiving/ (2015)

[37] Bailey, J., Grotke, A., Hanna, K., Hartman, C., Taylor, N.: Web archiving in the united states: A 2013 survey. http://www.digitalpreservation.gov/documents/NDSA_USWebArchivingSurvey_2013.pdf (2004)

[38] Bailey, J., Taylor, N., Rosenthal, D., Cramer, T.: Systems Interoperability and Collaborative Development for Web Archiving. https://www.imls.gov/sites/default/files/proposal_narritive_lg-71-15-0174_internet_archive.pdf (2015)

[39] Banos, V., Kim, Y., Ross, S., Manolopoulos, Y.: CLEAR: a credible method to evaluate website archivability. In: Proceedings of the 9th International Conference on Preservation of Digital Objects, iPRES 2013, pp. 9–18 (2013). URL http://delab.csd.auth.gr/papers/IPRES2013bkrm.pdf

[40] Bar-Yossef, Z., Broder, A.Z., Kumar, R., Tomkins, A.: Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay. In: WWW '04: Proceedings of the 13th international conference on World Wide Web, pp. 328–337 (2004). DOI 10.1145/988672.988716

[41] Beagrie, N.: Digital Curation for Science, Digital Libraries, and Individuals. International Journal of Digital Curation **1**(1), 3–16 (2006). DOI 10.2218/ijdc.v1i1.2

[42] Ben Saad, M., Gançarski, S.: Archiving the Web using Page Changes Patterns: A Case Study. In: Proceedings of the 11th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '11, pp. 113–122 (2012)

[43] Bergmark, D., Lagoze, C., Sbityakov, A.: Focused crawls, tunneling, and digital libraries. In: Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries, ECDL '02, pp. 91–106. Springer-Verlag (2002)

[44] Berners-Lee, T.: Information Management: A Proposal. https://www.w3.org/History/1989/proposal.html (1990)

[45] Berners-Lee, T., Fielding, R., Masinter, L.: RFC 2396 - Uniform Resource Identifiers (URI): Generic Syntax. http://www.ietf.org/rfc/rfc2396.txt (1998)

[46] Bilenko, M., Mooney, R.J.: Adaptive duplicate detection using learnable string similarity measures. In: Proceedings of the ninth ACM SIGKDD international conference

on Knowledge discovery and data mining - KDD '03, p. 39. ACM Press (2003). DOI 10.1145/956750.956759

[47] Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. The Journal of Machine Learning Research **3**, 993–1022 (2003)

[48] Bray, T.: The JavaScript Object Notation (JSON) Data Interchange Format. `https://tools.ietf.org/html/rfc7159` (2014)

[49] Brewington, B., Cybenko, G.: Keeping up with the changing Web. Computer **33**(5), 52–58 (2000). DOI 10.1109/2.841784

[50] Broache, A.: FBI rescinds secret order for Internet Archive records. CNET, `http://news.cnet.com/8301-10784_3-9938603-7.html` (2008)

[51] Broder, A.: On the Resemblance and Containment of Documents. In: Proceedings of Compression and Complexity of Sequences., pp. 21–29. IEEE Computer Society (1997). DOI 10.1109/SEQUEN.1997.666900

[52] Broder, A.Z., Glassman, S.C., Manasse, M.S., Zweig, G.: Syntactic clustering of the Web. Computer Networks and ISDN Systems **29**(8-13), 1157–1166 (1997). DOI 10.1016/S0169-7552(97)00031-7

[53] Brown, A.: Archiving websites: a practical guide for information management professionals. Facet (2006)

[54] Brown, R.: Selecting and Weighting N-Grams to Identify 1100 Languages. In: Text, Speech, and Dialogue, *Lecture Notes in Computer Science*, vol. 8082, pp. 475–483. Springer Berlin Heidelberg (2013)

[55] Brunelle, J.F.: Scripts in a frame: A framework for archiving deferred representations. Dissertation, Old Dominion University (2016)

[56] Brunelle, J.F., Kelly, M., SalahEldeen, H., Weigle, M.C., Nelson, M.L.: Not All Mementos Are Created Equal: Measuring The Impact Of Missing Resources. In: Proceedings of the 14th ACM/IEEE Joint Conference on Digital Libraries, JCDL '14, pp. 321 – 330 (2014). DOI 10.1109/JCDL.2014.6970187

[57] Brunelle, J.F., Kelly, M., SalahEldeen, H., Weigle, M.C., Nelson, M.L.: Not All Mementos Are Created Equal: Measuring The Impact Of Missing Resources. International Journal of Digital Libraries **16**(3), 283–301 (2015). DOI 10.1007/s00799-015-0150-6

[58] Brunelle, J.F., Kelly, M., Weigle, M.C., Nelson, M.L.: The Impact of JavaScript on Archivability. International Journal on Digital Libraries **17**(2), 95–117 (2016). DOI 10.1007/s00799-015-0140-8

[59] Bruza, P., Huibers, T., P. D. Bruza, T.W.C.H.: A Study of Aboutness in Information Retrieva. Artificial Intelligence Review **10**, 1–27 (1996)

[60] Buckley, C., Salton, G., Allan, J., Singhal, A.: Automatic Query Expansion Using SMART: TREC 3. Overview of the Third Text REtrieval Conference (TREC-3) pp. 69–80 (1995)

[61] Campbell, G.: Aboutness and Meaning: How a Paradigm of Subject Analysis Can Illuminate Queer Theory in Literary Studies. In: Canadian Association for Information Science (CAIS) (2000)

[62] Capra, R.G., Lee, C.A., Marchionini, G., Russell, T., Shah, C., Stutzman, F.: Selection and context scoping for digital video collections: an investigation of youtube and blogs. In: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '08, pp. 211–220. ACM (2008). DOI 10.1145/1378889.1378925

[63] Carmel, D., Yom-Tov, E., Roitman, H.: Enhancing Digital Libraries Using Missing Content Analysis. In: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '08, pp. 1–10. ACM (2008)

[64] Castellano, G., Fanelli, A.M., Torsello, M.A.: LODAP: a log data preprocessor for mining web browsing patterns. In: Proceedings of the 6th Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, AIKED '07, pp. 12–17 (2007)

[65] Cathro, W., Webb, C., Whiting, J.: Archiving the Web: The PANDORA Archive at the National Library of Australia (2001). URL `http://www.nla.gov.au/openpublish/index.php/nlasp/article/viewArticle/1314`

[66] Catledge, L.D., Pitkow, J.E.: Characterizing browsing strategies in the World-Wide web. Computer Networks and ISDN Systems **27**(6), 1065–1073 (1995). DOI 10.1016/0169-7552(95)00043-7

[67] Chakrabarti, S., Van den Berg, M., Dom, B.: Focused crawling: a new approach to topic-specific web resource discovery. Computer Networks **31**(11), 1623–1640 (1999). DOI 10.1016/S1389-1286(99)00052-3

[68] Chang, F., Dean, J., Ghemawat, S., Hsieh, W.C., Wallach, D.A., Burrows, M., Chandra, T., Fikes, A., Gruber, R.E.: Bigtable: A distributed storage system for structured data. In: Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation, OSDI '06, pp. 15–15. USENIX Association (2006)

[69] Chang, F., Dean, J., Ghemawat, S., Hsieh, W.C., Wallach, D.A., Burrows, M., Chandra, T., Fikes, A., Gruber, R.E.: Bigtable: A distributed storage system for structured data. ACM Transactions on Computer Systems (TOCS) **26**(2), 4 (2008)

[70] Chang, M., Leggett, J.J., Furuta, R., Kerne, A., Williams, J.P., Burns, S.A., Bias, R.G.: Collection Understanding. In: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '04, p. 334. ACM Press (2004). DOI 10.1145/996350.996426

[71] Charikar, M.S.: Similarity Estimation Techniques from Rounding Algorithms. In: Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing, STOC '02, pp. 380–388. ACM (2002). DOI 10.1145/509907.509965

[72] Cho, J., Garcia-Molina, H.: Estimating frequency of change. ACM Transactions on Internet Technology **3**(3), 256–290 (2003). DOI 10.1145/857166.857170

[73] Cho, J., Garcia-Molina, H., Haveliwala, T., Lam, W., Paepcke, A., Raghavan, S., Wesley, G.: Stanford WebBase components and applications. ACM Transactions on Internet Technology (TOIT) **6**(2), 153–186 (2006)

[74] Cho, J., Garcia-Molina, H., Page, L.: Efficient crawling through URL ordering. Computer Networks and ISDN Systems **30**(1-7), 161–172 (1998). DOI 10.1016/S0169-7552(98)00108-1

[75] Christel, M.G., Smith, M.A., Taylor, C.R., Winkler, D.B.: Evolving video skims into useful multimedia abstractions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '98, pp. 171–178. ACM Press/Addison-Wesley Publishing Co. (1998). DOI 10.1145/274644.274670

[76] Chu, W.T., Lin, C.H.: Automatic Selection of Representative Photo and Smart Thumbnailing Using Near-duplicate Detection. In: Proceedings of the 16th ACM International Conference on Multimedia, MM '08, pp. 829–832. ACM Press (2008). DOI 10.1145/1459359.1459498

[77] Cohen, E.L., Willis, C.: One nation under radio: Digital and public memory after september 11. New Media & Society **6**(5), 591–610 (2004)

[78] Cohen, J., Mihailidis, P.: Storify and News Curation: Teaching and Learning about Digital Storytelling. In: Second Annual Social Media Technology Conference & Workshop, vol. 1, pp. 27–31 (2012)

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

[80] Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995). DOI 10.1007/BF00994018

[81] Costa, M., J. Silva, M.: Characterizing Search Behavior in Web Archives. In: Proceedings of Temporal Web Analytics Workshop, TWAW 2011, pp. 33–40 (2011). URL `http://xldb.fc.ul.pt/xldb/publications/Costa.etal:CharacterizingSearchBehavior:2011_document.pdf`

[82] Costa, M., Silva, M.J.: Understanding the Information Needs of Web Archive Users. In: Proceedings of the 10th International Web Archiving Workshop, pp. 9–16 (2010)

[83] Croft, B., Metzler, D., Strohman, T.: Search Engines: Information Retrieval in Practice. Addison-Wesley Publishing Company (2009)

[84] Cunningham, S.J., Bennett, E.: Understanding collection understanding with collage. In: Proceeding of 11th International Conference on Asian digital Libraries, ICADL 2008, pp. 367–370. Springer (2008). DOI 10.1007/978-3-540-89533-6_46

[85] Cutts, M.: SEO advice: URL canonicalization. `http://www.mattcutts.com/blog/seo-advice-url-canonicalization/` (2006)

[86] Czernicki, B.: Silverlight 4 Business Intelligence Software. Apress (2010). DOI 10.1007/978-1-4302-3061-8

[87] Damnjanovic, U., Izquierdo, E., Grzegorzek, M.: Shot boundary detection using spectral clustering. In: Proceedings of the 15th European Signal Processing Conference, pp. 1779–1783. IEEE (2007)

[88] Deal, L.: Visualizing digital collections. Technical Services Quarterly **32**(1), 14–34 (2015). DOI 10.1080/07317131.2015.972871

[89] Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Communications of the ACM **51**(1), 107–113 (2008). DOI 10.1145/1327452.1327492

[90] Dice, L.R.: Measures of the Amount of Ecologic Association Between Species. Ecology **26**(3), 297–302 (1945). DOI 10.2307/1932409

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

[92] Dirfaux, F.: Key frame selection to represent a video. In: Proceedings of IEEE 2000 International Conference on Image Processing, vol. 2, pp. 275–278. IEEE (2000)

[93] Doran, D., Gokhale, S.S.: Web robot detection techniques: overview and limitations. Data Mining and Knowledge Discovery **22**(1-2), 183–210 (2010). DOI 10.1007/s10618-010-0180-z

[94] Dou, W., Wang, X., Skau, D., Ribarsky, W., Zhou, M.X.: 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, pp. 93–102. IEEE Computer Society (2012)

[95] Dougherty, R.L.: Documenting Revolution in the Middle East. Center for Research Libraries (CRL) **31**, 5–7 (2011). URL `https://www.crl.edu/focus/article/7437`

[96] Duh, K., Hirao, T., Kimura, A., Ishiguro, K., Iwata, T., Yeung, C.M.A.: Creating stories: Social curation of twitter messages. In: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM) (2012)

[97] Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 226–231. AAAI Press (1996)

[98] Evans, A., Martin, K., Poatsy, M.A.: Technology In Action, Complete Version, 7th edn. Prentice Hall Press, Upper Saddle River, NJ, USA (2010)

[99] Eysenbach, G., Trudel, M.: Going, going, still there: using the WebCite service to permanently archive cited web pages. Journal of Medical Internet Research **7**(5), 919 (2005). DOI 10.2196/jmir.7.5.e60

[100] Farag, M.M.G., Fox, E.A.: Intelligent Event Focused Crawling. In: Proceedings of the 11th International ISCRAM Conference, pp. 18–21 (2014)

[101] Fawcett, T.: An Introduction to ROC Analysis. Pattern Recognition Letters **27**(8), 861–874 (2006). DOI 10.1016/j.patrec.2005.10.010

[102] Feinberg, J.: Wordle - Beautiful Word Clouds. `http://www.wordle.net/`

[103] Fielding, R.T., Gettys, J., Mogul, J.C., Frystyk, H., Masinter, L., Leach, P.J., Berners-Lee, T.: RFC 2616 - Hypertext Transfer Protocol. `http://www.ietf.org/rfc/rfc2616.txt` (1999)

[104] Fiscus, J., Doddington, G.: Topic Detection and Tracking Evaluation Overview . Topic detection and tracking **12**, 17–31 (2002). DOI 10.1007/978-1-4615-0933-2_2

[105] Foot, K., Schneider, S.: Web Campaigning. MIT press Cambridge, MA (2006)

[106] Francisco-Revilla, L., Trace, C.B., Li, H., Buchanan, S.A.: Encoded archival description: Data quality and analysis. Proceedings of the American Society for Information Science and Technology **51**(1), 1–10 (2014). DOI 10.1002/meet.2014.14505101043

[107] Fukuda, K., Cho, K., Esaki, H.: The Impact of Residential Broadband Traffic on Japanese ISP Backbones. ACM SIGCOMM Computer Communication Review **35**(1), 15–22 (2005)

[108] Gamon, M., Yano, T., Song, X., Apacible, J., Pantel, P.: Understanding Document Aboutness-Step One: Identifying Salient Entities. Tech. Rep. MSR-TR-201 (2013). URL `http://research.microsoft.com/pubs/198455/msrtr13.pdf`

[109] Gershon, N., Page, W.: What storytelling can do for information visualization. Communications of the ACM **44**(8), 31–37 (2001). DOI 10.1145/381641.381653

[110] Ghobrial, B.G., Wilkins, K.G.: The politics of political communication: Competing news discourses of the 2011 Egyptian protests. International Communication Gazette pp. 1–22 (2014). DOI 10.1177/1748048514564027

[111] Giaretta, D.: DCC approach to digital curation. `http://twiki.dcc.rl.ac.uk/bin/view/OLD/DCCApproachToCuration` (2005)

[112] Gilbert, E., Bakhshi, S., Chang, S., Terveen, L.: "I Need to Try This"?: A Statistical Overview of Pinterest. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13, pp. 2427–2436. ACM (2013). DOI 10.1145/2470654.2481336

[113] Gomes, D., Miranda, J., Costa, M.: A Survey on Web Archiving Initiatives. In: Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries, TPDL '11, pp. 408–420. Springer International Publishing (2011). DOI 10.1007/978-3-642-24469-8_41

[114] Good, R.: The Future Of Content Curation Tools - Part II. `http://www.masternewmedia.org/content-curation-tools-future-part2/` (2013)

[115] Gorg, C., Liu, Z., Parekh, N., Singhal, K., Stasko, J.: Visual Analytics with Jigsaw. In: 2007 IEEE Symposium on Visual Analytics Science and Technology, pp. 201–202. IEEE (2007). DOI 10.1109/VAST.2007.4389017

[116] Graham, A., Garcia-Molina, H., Paepcke, A., Winograd, T.: Time As Essence for Photo Browsing Through Personal Digital Libraries. In: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '02, pp. 326–335. ACM Press (2002). DOI 10.1145/544220.544301

[117] Guo, W., Zhong, Y., Xie, J.: A Web Crawler Detection Algorithm Based on Web Page Member List. In: 2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics, pp. 189–192. IEEE (2012). DOI 10.1109/IHMSC.2012.54

[118] Hall, C., Zarro, M.: Social curation on the website Pinterest.com. Proceedings of the American Society for Information Science and Technology **49**(1), 1–9 (2012)

[119] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software. ACM SIGKDD Explorations Newsletter **11**(1), 10 (2009). DOI 10.1145/1656274.1656278

[120] Hammoud, R., Mohr, R.: A probabilistic framework of selecting effective key frames for video browsing and indexing. In: Proceedings of International Workshop on Real-Time Image Sequence Analysis, RISA'00, pp. 79–88 (2000)

[121] Han, J., Choi, D., Choi, A.Y., Choi, J., Chung, T., Kwon, T.T., Rha, J.Y., Chuah, C.N.: Sharing topics in pinterest: Understanding content creation and diffusion behaviors. In: Proceedings of the 2015 ACM on Conference on Online Social Networks, COSN '15, pp. 245–255. ACM (2015). DOI 10.1145/2817946.2817961

[122] Handley, A.: Content Curation Definitions & Context for Content Marketing. `http://www.toprankblog.com/2010/06/content-marketing-curation-context/` (2010)

[123] Hanjalic, A., Zhang, H.: An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. IEEE Transactions on Circuits and Systems for Video Technology **9**(8), 1280–1289 (1999). DOI 10.1109/76.809162

[124] Harrison, T.L., Nelson, M.L.: Just-In-Time Recovery of Missing Web Pages. In: Proceedings of the 17th Conference on Hypertext and Hypermedia, HT '06, pp. 145–156. ACM (2006)

[125] Hartigan, J.A., Wong, M.A.: Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) **28**(1), 100–108 (1979)

[126] Hatcher, E., Gospodnetic, O.: Lucene in Action (In Action series). Manning Publications Co. (2004)

[127] Hauptmann, A.G., Witbrock, M.J.: Intelligent multimedia information retrieval. chap. Informedia: News-on-demand Multimedia Information Acquisition and Retrieval, pp. 215–239. MIT Press (1997)

[128] Hauslohner, A.: Egyptians, Inspired by Tunisia, Use Facebook to Set Up Protest March. `http://content.time.com/time/world/article/0,8599,2044142,00.html` (2011)

[129] Havre, S., Hetzler, B., Nowell, L.: ThemeRiver: Visualizing Theme Changes over Time. In: Proceedings of the IEEE Symposium on Information Vizualization 2000, INFOVIS '00, pp. 115–123. IEEE Computer Society (2000). DOI 10.1109/INFVIS.2000.885098

[130] Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., Yee, K.P.: Finding the flow in web site search. Communications of the ACM **45**(9), 42–49 (2002)

[131] Heer, J., Bostock, M.: Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In: Proceedings of the 26th SIGCHI Conference on Human Factors in Computing Systems, CHI '10, pp. 203–212. ACM (2010). DOI 10.1145/1753326.1753357

[132] Henzinger, M.: Finding near-duplicate web pages: a large-scale evaluation of algorithms. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, pp. 284–291. ACM Press (2006). DOI 10.1145/1148170.1148222

[133] Hockx-Yu, H.: The Past Issue of the Web. In: Proceedings of the 3rd International Web Science Conference, WebSci'11, pp. 1–8. ACM (2011)

[134] Holton, A.E., Chyi, H.I.: News and the Overloaded Consumer: Factors Influencing Information Overload Among News Consumers. Cyberpsychology, Behavior, and Social Networking **15**(11), 619–624 (2012)

[135] Horrigan, J.: Broadband adoption and use in America. Federal Communications Commission (2010). URL https://apps.fcc.gov/edocs_public/attachmatch/DOC-296442A1.pdf

[136] Howard, P.N., Duffy, A., Freelon, D., Hussain, M.M., Mari, W., Mazaid, M.: Opening closed regimes: what was the role of social media during the Arab Spring? Social Science Research Network (SSRN) (2011). DOI 10.2139/ssrn.2595096

[137] Hu, K.: Visarchive: A time and relevance based visual interface for searching, browsing, and exploring project archives (with timeline and relevance visualization). Dissertation, University of Victoria (2014)

[138] Hu, K., Tory, M., Staub-French, S., Nepal, M.P.: Visarchive: a time and relevance based visual interface for searching, browsing and exploring project archives. Visualization in Engineering (2016). URL http://eprints.qut.edu.au/93193/

[139] Hullman, J., Diakopoulos, N.: Visualization rhetoric: Framing effects in narrative visualization. Visualization and Computer Graphics, IEEE Transactions on **17**(12), 2231–2240 (2011)

[140] Hullman, J., Drucker, S., Henry Riche, N., Lee, B., Fisher, D., Adar, E.: A deeper understanding of sequence in narrative visualization. Visualization and Computer Graphics, IEEE Transactions on **19**(12), 2406–2415 (2013)

[141] Hungerford, K.: Keep the Presses Rolling! http://blog.paper.li/2011/03/keep-presses-rolling.html (2011)

[142] Hwang, E.: 100 million of the most interesting people we know. https://blog.pinterest.com/en/100-million-most-interesting-people-we-know (2015)

[143] Indyk, P., Motwani, R.: Approximate nearest neighbors: Towards removing the curse of dimensionality. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, STOC '98, pp. 604–613. ACM (1998). DOI 10.1145/276698.276876

[144] Irmak, U., von Brzeski, V., Kraft, R.: Contextual Ranking of Keywords Using Click Data. In: Proceedings of the 2009 IEEE International Conference on Data Engineering - ICDE '09, pp. 457–468. IEEE (2009). DOI 10.1109/ICDE.2009.76

[145] Jacobs, I., Walsh, N.: Architecture of the World Wide Web, Volume One. Tech. Rep. W3C Recommendation 15 December 2004, W3C (2004). URL http://www.w3.org/TR/webarch/

[146] Jaffe, A., Naaman, M., Tassa, T., Davis, M.: Generating summaries and visualization for large collections of geo-referenced photographs. In: Proceedings of the 8th ACM international workshop on Multimedia information retrieval, pp. 89–98. ACM (2006)

[147] Jain, A., Murty, M., Flynn, P.: Data clustering: a review. ACM computing surveys (CSUR) **31**(3), 264–323 (1999). DOI 10.1145/331499.331504

[148] Järvelin, K., Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '00, pp. 41–48. ACM Press (2000). DOI 10.1145/345508.345545

[149] Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems **20**(4), 422–446 (2002). DOI 10.1145/582415.582418

[150] Jatowt, A., Ishizuka, M.: Temporal Web Page Summarization. In: Proceedings of the 5th International Conference on Web Information Systems Engineering, vol. 3306, pp. 303–312. Springer Berlin Heidelberg (2004). DOI 10.1007/978-3-540-30480-7_31

[151] Jatowt, A., Kanazawa, K., Oyama, S., Tanaka, K.: 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, pp. 115–124. ACM (2009)

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

[153] Jatowt, A., Kawai, Y., Tanaka, K.: Personalized detection of fresh content and temporal annotation for improved page revisiting. In: Database and Expert Systems Applications, pp. 832–841. Springer (2006)

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

[155] Jatowt, A., Kawai, Y., Tanaka, K.: Visualizing historical content of web pages. In: Proceeding of the 17th International World Wide Web Conference, WWW '08, pp. 1221–1222. ACM Press (2008). DOI 10.1145/1367497.1367736

[156] Jatowt, A., Kawai, Y., Tanaka, K.: Page History Explorer: Visualizing and Comparing Page Histories. IEICE Transactions on Information and Systems **94**(3), 564–577 (2011)

[157] Jatowt, A., Tanaka, K.: Towards mining past content of Web pages. New Review of Hypermedia and Multimedia **13**(1), 77–86 (2007). DOI 10.1080/13614560701478897

[158] Jones, M.L., Jones, E., Zeide, E., Dupre, J., Mai, J.E., Richards, N.: The right to be forgotten. In: Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, ASIST '15, pp. 10:1–10:3. American Society for Information Science, Silver Springs, MD, USA (2015)

[159] Jones, R., Klinkner, K.L.: 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, pp. 699–708. ACM Press (2008). DOI 10.1145/1458082.1458176

[160] Jung, B., Kwak, T., Song, J., Lee, Y.: Narrative abstraction model for story-oriented video. In: Proceedings of the 12th Annual ACM International Conference on Multimedia, MULTIMEDIA '04, pp. 828–835. ACM, New York, NY, USA (2004). DOI 10.1145/1027527.1027720

[161] Jung, B., Song, J., Lee, Y.: A narrative-based abstraction framework for story-oriented video. The ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **3**(2) (2007). DOI 10.1145/1230812.1230817

[162] Kahle, B.: Preserving The Internet. Scientific American **276**(3), 82–83 (1997). URL http://cat.inist.fr/?aModele=afficheN&cpsidt=10731235

[163] Kahle, B.: Reader Privacy at the Internet Archive. Internet Archive Blogs, http://blog.archive.org/2013/10/25/reader-privacy-at-the-internet-archive/ (2013)

[164] Kahle, B.: Wayback Machine Hits 400,000,000,000! Internet Archive Blogs, http://blog.archive.org/2014/05/09/wayback-machine-hits-400000000000/ (2014)

[165] Kahle, B., Burner, M.: Arc File Format. http://archive.org/web/researcher/ArcFileFormat.php (1996)

[166] Kang, H.B.: Video Abstraction Techniques for a Digital Library. IGI Global (2002)

[167] Kehoe, A., Gee, M.: Reader comments as an aboutness indicator in online texts: introducing the Birmingham Blog Corpus. Studies in Variation, Contacts and Change in English **12** (2012). URL `http://www.helsinki.fi/varieng/journal/volumes/12/kehoe_gee`

[168] Kelly, M., Brunelle, J., Weigle, M., Nelson, M.: A Method for Identifying Personalized Representations in Web Archives. D-Lib Magazine **19**, 2 (2013). DOI 10.1045/november2013-kelly

[169] Kelly, M., Brunelle, J.F., Weigle, M.C., Nelson, M.L.: On the Change in Archivability of Websites Over Time. In: Proceedings of the 3rd International Conference on Theory and Practice of Digital Libraries, pp. 35–47 (2013). DOI 10.1007/978-3-642-40501-3_5

[170] Kelly, M., Nelson, M.L., Weigle, M.C.: Mink: Integrating the Live and Archived Web Viewing Experience Using Web Browsers and Memento. In: Proceedings of the 14th ACM/IEEE Joint Conference on Digital Libraries, pp. 469–470 (2014). DOI 10.1109/JCDL.2014.6970229

[171] Kemper, E.A., Stringfield, S., Teddlie, C.: Mixed methods sampling strategies in social science research. Handbook of mixed methods in social and behavioral research pp. 273–296 (2003)

[172] Kessler, S.: Facebook & Twitter Both Blocked in Egypt. Mashable, `http://mashable.com/2011/01/26/facebook-blocked-in-egypt/` (2011)

[173] Kibriya, A., Frank, E., Pfahringer, B., Holmes, G.: Multinomial naive bayes for text categorization revisited. AI 2004: Advances in Artificial Intelligence **3339**, 488–499 (2005). DOI 10.1007/978-3-540-30549-1_43

[174] Kieu, B.T., Ichise, R., Pham, S.B.: Predicting the popularity of social curation. In: Knowledge and Systems Engineering, pp. 413–424. Springer (2015)

[175] Kim, P., Myaeng, S.H.: Usefulness of temporal information automatically extracted from news articles for topic tracking. ACM Transactions on Asian Language Information Processing **3**(4), 227–242 (2004). DOI 10.1145/1039621.1039624

[176] Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with mechanical turk. In: Proceedings of the 24th SIGCHI Conference on Human Factors in Computing Systems, CHI '08, pp. 453–456 (2008)

[177] Klein, M., Nelson, M.: Find, new, copy, web, page-tagging for the (re-) discovery of web pages. In: Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries, TPDL '11, pp. 27–39. Springer Berlin Heidelberg (2011). DOI 10.1007/978-3-642-24469-8_5

[178] Klein, M., Nelson, M.L.: Revisiting Lexical Signatures to (Re-)Discover Web Pages. In: Proceedings of the 12th European conference on Research and Advanced Technology for Digital Libraries, *ECDL '08*, vol. 5173, pp. 371–382. Springer Berlin Heidelberg (2008). DOI 10.1007/978-3-540-87599-4

[179] Klein, M., Shipman, J., Nelson, M.L.: Is this a good title? In: Proceedings of the 21st Conference on Hypertext and Hypermedia, HT '10, pp. 3–12. ACM (2010). DOI 10.1145/1810617.1810621

[180] Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., Tobin, R.: Scholarly context not found: One in five articles suffers from reference rot. PloS ONE **9**(12), e115,253 (2014). DOI 10.1371/journal.pone.0115253

[181] Klein, M., Ware, J., Nelson, M.L.: Rediscovering Missing Web Pages Using Link Neighborhood Lexical Signatures. In: Proceedings of the 11th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '11, pp. 137–140. ACM Press (2011). DOI 10.1145/1998076.1998101

[182] Koehler, W.: Web Page Change and Persistence-A Four-Year Longitudinal Study. Journal of the American Society for Information Science and Technology **53**(2), 162–171 (2002)

[183] Koehler, W.: A longitudinal study of web pages continued: a consideration of document persistence. Information Research **9**(2), 9–2 (2004)

[184] Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate Detection Using Shallow Text Features. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM '10, pp. 441–450. ACM (2010). DOI 10.1145/1718487.1718542

[185] Korany, B., El-Mahdi, R.: Arab Spring in Egypt: Revolution and Beyond. American University in Cairo Press (2012). URL `http://www.jstor.org/stable/j.ctt15m7mbm`

[186] Kosala, R., Blockeel, H.: Web Mining Research: A Survey. SIGKDD Exploration Newsletter. **2**(1), 1–15 (2000). DOI 10.1145/360402.360406

[187] Kosara, R., Ziemkiewicz, C.: Do mechanical turks dream of square pie charts? In: Proceedings of the 3rd BELIV'10 Workshop: BEyond Time and Errors: Novel evaLuation Methods for Information Visualization, BELIV '10, pp. 63–70. ACM (2010). DOI 10.1145/2110192.2110202

[188] Kramer-Smyth, J., Nishigaki, M., Anglade, T.: ArchivesZ: Visualizing Archival Collections. `http://archivesz.com/ArchivesZ.pdf` (2007)

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

[190] Krstajic, M., Najm-Araghi, M., Mansmann, F., Keim, D.A.: Incremental Visual Text Analytics of News Story Development. In: Proceedings of IS&T/SPIE Electronic Imaging, vol. 8294, pp. 829,407–829,412 (2012)

[191] Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association **47**(260), 583–621 (1952). DOI 10.1080/01621459.1952.10483441

[192] Kullback, S., Leibler, R.: On information and sufficiency. The Annals of Mathematical Statistics (1951). URL `http://www.jstor.org/stable/2236703`

[193] Kumar, J.P., Govindarajulu, P.: 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)

[194] Kumar, R., Tomkins, A.: A Characterization of Online Browsing Behavior. In: Proceedings of the 19th International World Wide Web Conference, WWW '10, pp. 561–570. ACM (2010). DOI 10.1145/1772690.1772748

[195] Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a Social Network or a News Media? In: Proceedings of the 19th international World Wide Web Conference, WWW '10, pp. 591–600. ACM (2010)

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

[197] Laire, D., Casteleyn, J., Mottart, A.: Social Media's Learning Outcomes within Writing Instruction in the EFL Classroom: Exploring, Implementing and Analyzing Storify. Procedia-Social and Behavioral Sciences **69**, 442–448 (2012)

[198] Lampos, C., Eirinaki, M.: Archiving the Greek Web. Proceedings of the 4th International Web Archiving Workshop (IWAW'04) (2004). URL `http://sjsulug.engr.sjsu.edu/meirinaki/papers/LEJV04-IWAW.pdf`

[199] Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V., Thomas, S.: Relevance models for topic detection and tracking. In: Proceedings of the Second International Conference on Human Language Technology Research, HLT '02, pp. 115–121. Morgan Kaufmann Publishers Inc. (2002)

[200] Lavrenko, V., Croft, W.B.: Relevance based language models. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01, pp. 120–127. ACM Press (2001). DOI 10.1145/383952.383972

[201] Lawrence, S., Pennock, D.M., Flake, G.W., Krovetz, R., Coetzee, F.M., Glover, E., Nielsen, F.A., Kruger, A., Giles, C.L.: Persistence of Web References in Scientific Research. Computer **34**(2), 26–31 (2001). DOI 10.1109/2.901164

[202] Leek, T., Schwartz, R., Sista, S.: Probabilistic approaches to topic detection and tracking. Topic detection and tracking pp. 67–83 (2002). DOI 10.1007/978-1-4615-0933-2_4

[203] Levenshtein, V.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics-Doklady **10**(8) (1966). URL `http://profs.sci.univr.it/~liptak/ALBioinfo/files/levenshtein66.pdf`

[204] Li, J., Lim, J., Tian, Q.: Automatic Summarization for Personal Digital Photos. In: Proceedings of the 2003 Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing and Fourth Pacific-Rim Conference on Multimedia, vol. 3, pp. 1536–1540. IEEE (2003)

[205] Li, Y., Zhang, T., Tretter, D.: An overview of video abstraction techniques. Tech. Rep. HPL-2001-191, HP Laboratory (2001)

[206] Lienhart, R., Pfeiffer, S., Effelsberg, W.: Video abstracting. Communications of the ACM **40**(12), 54–62 (1997). DOI 10.1145/265563.265572

[207] Lin, J.: Scaling down distributed infrastructure on wimpy machines for personal web archiving. In: Proceedings of the 24th International Conference on World Wide Web, WWW '15 Companion, pp. 1351–1355. ACM (2015). DOI 10.1145/2740908.2741695

[208] Lin, J., Gholami, M., Rao, J.: Infrastructure for supporting exploration and discovery in web archives. In: Proceedings of the 23rd International Conference on World Wide Web, WWW '14 Companion, pp. 851–856. ACM (2014). DOI 10.1145/2567948.2579045

[209] Liptak, A.: In Supreme Court Opinions, Web Links to Nowhere. The New York Times, http://www.nytimes.com/2013/09/24/us/politics/in-supreme-court-opinions-clicks-that-lead-nowhere.html

[210] Liu, S., Wu, Y., Wei, E., Liu, M., Liu, Y.: Storyflow: Tracking the evolution of stories. IEEE Transactions on Visualization and Computer Graphics **19**(12), 2436–2445 (2013). DOI 10.1109/TVCG.2013.196

[211] Liu, S.B.: The Rise of Curated Crisis Content. In: Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM 2010) (2010)

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

[213] Lyman, P.: Archiving the World Wide Web. Tech. rep., Building a national strategy for digital preservation. Council on Library and Information Resources and Library of Congress (2002). URL http://www.clir.org/pubs/reports/pub106/web.html

[214] Macefield, R.: How to specify the participant group size for usability studies: a practitioner's guide. Journal of Usability Studies **5**(1), 34–45 (2009)

[215] MacNeil, H.: Metadata strategies and archival description: Comparing apples to oranges. Archivaria **1**(39), 22–32 (1995)

[216] Manku, G.S., Jain, A., Das Sarma, A.: Detecting near-duplicates for web crawling. In: Proceedings of the 16th international conference on World Wide Web, WWW '07, pp. 141–150. ACM Press (2007). DOI 10.1145/1242572.1242592

[217] Manning, C.D., Raghavan, P., Schütze, H., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press (2008). DOI 10.1017/CBO9780511809071

[218] Maqsoud, S.A.: Egypt's media revolution. The Guardian, http://www.theguardian.com/commentisfree/2011/feb/15/egypt-media-revolution-mubarak-lies (2011)

[219] Marchionini, G., Shah, C., Lee, C.A., Capra, R.: Query parameters for harvesting digital video and associated contextual information. In: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '09, pp. 77–86. ACM (2009). DOI 10.1145/1555400.1555414

[220] Marcu, D.: From discourse structures to text summaries. In: Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, pp. 82–88 (1997)

[221] Markov, Z., Larose, D.T.: Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage. John Wiley & Sons, Inc., Hoboken, New Jersey (2007). DOI 10.1002/0470108096

[222] Marshall, C., McCown, F., Nelson, M.: Evaluating Personal Archiving Strategies for Internet-based Information. In: Proceedings of Archiving 2007, 1, pp. 151–156 (2007). URL http://www.ingentaconnect.com/content/ist/ac/2007/00002007/00000001/art00036

[223] Marshall, C.C., Shipman, F.M.: On the Institutional Archiving of Social Media. In: Proceedings of the 12th ACM/IEEE Joint Conference on Digital Libraries, pp. 1–10 (2012). DOI 10.1145/2232817.2232819

[224] Masanès, J.: Web archiving methods and approaches: A comparative study. Library trends **54**(1), 72–90 (2005)

[225] Masanès, J.: Web Archiving. Springer (2006)

[226] McCormack, C.: Storying stories: a narrative approach to in-depth interview conversations. International Journal of Social Research Methodology **7**(3), 219–236 (2004). DOI 10.1080/13645570210166382

[227] McCown, F., Bollen, J., Chan, S., Nelson, M.L.: The Availability and Persistence of Web References in D-Lib Magazine. In: Proceedings of the 5th International Web Archiving Workshop and Digital Preservation, IWAW '05 (2005)

[228] McCown, F., Nelson, M.L.: Evaluation of crawling policies for a web-repository crawler. In: Proceedings of the 17th Conference on Hypertext and Hypermedia, HT '06, pp. 157–168. ACM Press (2006). DOI 10.1145/1149941.1149972

[229] Mendi, E., Bayrak, C.: Shot boundary detection and key frame extraction using salient region detection and structural similarity. In: Proceedings of the 48th Annual Southeast Regional Conference, ACM SE '10, pp. 66:1–66:4. ACM (2010). DOI 10.1145/1900008.1900096

[230] Metzler, D., Dumais, S., Meek, C.: Similarity Measures for Short Segments of Text. In: Proceedings of the 29th European Conference on IR Research, ECIR '07, pp. 16–27. Springer-Verlag (2007)

[231] Mihailidis, P., Cohen, J.N.: Exploring Curation as a Core Competency in Digital and Media Literacy Education. Journal of Interactive Media in Education **1**, 1–19 (2013). DOI 10.5334/2013-02

[232] Mobasher, B., Dai, H., Luo, T., Sun, Y., Zhu, J.: 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, pp. 165–176. Springer-Verlag (2000)

[233] Mohr, G., Stack, M., Ranitovic, I., Avery, D., Kimpton, M.: An Introduction to Heritrix An open source archival quality web crawler. In: Proceedings of the 4th International Web Archiving Workshop, IWAW '04, pp. 43–49 (2004). URL `http://iwaw.europarchive.org/04/Mohr.pdf`

[234] Molina, A., SanJuan, E., Moreno, J.M.T.: A turing test to evaluate a complex summarization task. In: Proceedings of the 4th International Conference of the CLEF Initiative, *CLEF '13*, vol. 8138, pp. 75–80. Springer (2013)

[235] Mori, M., Miura, T., Shioya, I.: Topic Detection and Tracking for News Web Pages. In: 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06), pp. 338–342. IEEE (2006). DOI 10.1109/WI.2006.171

[236] Naaman, M., Harada, S., Wang, Q., Garcia-Molina, H., Paepcke, A.: Context Data in Geo-Referenced Digital Photo Collections. In: Proceedings of the 12th Annual ACM International Conference on Multimedia, MULTIMEDIA '04, pp. 196–203. ACM Press (2004). DOI 10.1145/1027527.1027573

[237] Naphade, M.R., Yeung, M.M., Yeo, B.L.: Novel scheme for fast and efficent video sequence matching using compact signatures. In: Electronic Imaging, pp. 564–572. International Society for Optics and Photonics (1999)

[238] Negulescu, K.C.: Web Archiving @ the Internet Archive. Presentation at the 2010 Digital Preservation Partners Meeting, `http://www.digitalpreservation.gov/meetings/documents/ndiipp10/NDIIPP072110FinalIA.ppt` (2010)

[239] Nelson, M.L.: 2010-11-05: Memento-Datetime is not Last-Modified. http://ws-dl.blogspot.com/2010/11/2010-11-05-memento-datetime-is-not-last.html (2010)

[240] Nelson, M.L.: A Plan For Curating "Obsolete Data or Resources". Tech. Rep. arXiv:1209.2664 (2012)

[241] Nelson, M.L., Allen, B.D.: Object Persistence and Availability in Digital Libraries. D-Lib Magazine **8**(1) (2002). URL http://www.dlib.org/dlib/january02/nelson/01nelson.html

[242] Ngo, C.W., Pong, T.C., Zhang, H.J.: On clustering and retrieval of video shots. In: Proceedings of the Ninth ACM International Conference on Multimedia, MULTIMEDIA '01, pp. 51–60. ACM (2001). DOI 10.1145/500141.500151

[243] Ngo, C.W., Pong, T.C., Zhang, H.J.: On clustering and retrieval of video shots through temporal slices analysis. Multimedia, IEEE Transactions on **4**(4), 446–458 (2002)

[244] Nguyen, G., Worring, M.: Interactive access to large image collections using similarity-based visualization. Journal of Visual Languages and Computing **19**(2), 203–224 (2008). DOI 10.1.1.175.1179

[245] Nithya, P., Sumathi, P.: Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise, Cookies and Web Robots. International Journal of Computer Applications **53**(17), 1–6 (2012). URL http://www.ijcaonline.org/archives/volume53/number17/8510-1684

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

[247] Oh, J., Wen, Q., Hwang, S., Lee, J.: Video abstraction. Video data management and information retrieval pp. 321–346 (2004)

[248] Olston, C., Pandey, S.: Recrawl scheduling based on information longevity. In: Proceeding of the 17th International World Wide Web Conference, WWW '08, pp. 437–446. ACM Press (2008). DOI 10.1145/1367497.1367557

[249] Ou-Yang, L.: Newspaper: Article scraping & curation. http://newspaper.readthedocs.io/ (2013)

[250] Ovadia, S.: Digital Content Curation and Why It Matters to Librarians. Behavioral & Social Sciences Librarian **32**(1), 58–62 (2013). DOI 10.1080/01639269.2013.750508

[251] Padia, K.: Visualizing Digital Collections at Archive-It. Master's thesis, Old Dominion University (2012)

[252] Padia, K., AlNoamany, Y., Weigle, M.C.: Visualizing Digital Collections at Archive-It. In: Proceeding of the 12th Annual International ACM/IEEE Joint Conference on Digital Libraries, JCDL '12, pp. 437–438 (2012). DOI 10.1145/2232817.2232821

[253] Papadimitriou, C.H., Tamaki, H., Raghavan, P., Vempala, S.: Latent Semantic Indexing: A Probabilistic Analysis. In: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, PODS '98, pp. 159–168. ACM Press (1998). DOI 10.1145/275487.275505

[254] Paranjpe, D.: Learning document aboutness from implicit user feedback and document structure. In: Proceeding of the 18th ACM conference on Information and knowledge management, CIKM '09, pp. 365–374. ACM Press (2009). DOI 10.1145/1645953.1646002

[255] Park, S.T., Pennock, D.M., Giles, C.L., Krovetz, R.: Analysis of lexical signatures for finding lost or related documents. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '02, p. 11. ACM Press (2002). DOI 10.1145/564376.564381

[256] Pasha, T.: Islamists in the headlines: Critical discourse analysis of the representation of the Muslim Brotherhood in Egyptian newspapers. Dissertation, University of Utah (2011)

[257] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

[258] Pitti, D.V.: Encoded archival description: An introduction and overview. New Review of Information Networking **5**(1), 61–69 (1999). DOI 10.1080/13614579909516936

[259] Plegas, Y., Stamou, S.: Reducing Information Redundancy in Search Results. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing, SAC '13, pp. 886–893. ACM (2013)

[260] Pope, A., Kumar, R., Sawhney, H., Wan, C.: Video abstraction: Summarizing video content for retrieval and visualization. In: The 29th Asilomar Conference on Signals, Systems and Computers, vol. 1, pp. 915–919. IEEE (1998). DOI 10.1109/ACSSC.1998.751015

[261] Popescu, A.: LiveFyre Acquires Social Storytelling Tool Storify. Mashable, `http://mashable.com/2013/09/09/livefyre-acquires-storify/` (2013)

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

[263] Reddy, K.S., Varma, G.P.S., Babu, I.R.: Preprocessing the Web Server Logs An illustrative approach for effective usage mining. ACM SIGSOFT Software Engineering Notes **37**(3), 1–5 (2012). DOI 10.1145/180921.2180940

[264] Reilly, B., Palaima, C., Norsworthy, K., Myrick, L., Tuchel, G., Simon, J.: Political Communications Web Archiving: Addressing Typology and Timing for Selection, Preservation and Access. In: Proceedings of the 3rd Workshop on Web Archives (2003)

[265] Reisinger, D.: Netflix gobbles a third of peak Internet traffic in North America. CNET, `http://news.cnet.com/8301-1023_3-57546405-93/netflix-gobbles-a-third-of-peak-internet-traffic-in-north-america/` (2012)

[266] Robertson, N.D.: Content Curation and the School Library. Knowledge Quest **41**(2), E1–E5 (2012)

[267] Robertson, S.: Understanding inverse document frequency: On theoretical arguments for IDF. Journal of Documentation **60**(5), 503–520 (2004)

[268] Rose, S., Butner, S., Cowley, W., Gregory, M., Walker, J.: Describing Story Evolution from Dynamic Information Streams. In: Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on, pp. 99–106 (2009). DOI 10.1109/VAST.2009.5333437

[269] Rosen, J.: The right to be forgotten. Stanford Law Review **64**, 88 (2012). URL `http://www.stanfordlawreview.org/online/privacy-paradox/right-to-be-forgotten`

[270] Rossi, A.: Fixing Broken Links on the Internet. Internet Archive Blogs, `https://blog.archive.org/2013/10/25/fixing-broken-links/` (2013)

[271] Rushdy, H., Soueif, A.: 18 Days in Tahrir: Stories from Egypt's Revolution. Haven Books (2011)

[272] Saaya, Z., Rafter, R., Schaal, M., Smyth, B.: The Curated Web: A Recommendation Challenge. In: Proceedings of the 7th ACM Conference on Recommender Systems, RecSys '13, pp. 101–104. ACM (2013). DOI 10.1145/2507157.2507216

[273] Saaya, Z., Schaal, M., Rafter, R., Smyth, B.: Recommending Topics for Web Curation. In: S. Carberry, S. Weibelzahl, A. Micarelli, G. Semeraro (eds.) User Modeling, Adaptation, and Personalization, *Lecture Notes in Computer Science*, vol. 7899, pp. 242–253. Springer Berlin Heidelberg (2013). DOI 10.1007/978-3-642-38844-6_20

[274] Sahami, M., Heilman, T.D.: A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets. In: Proceedings of the 15th International Conference on World Wide Web, WWW '06, pp. 377–386. ACM (2006). DOI 10.1145/1135777.1135834

[275] SalahEldeen, H.M.: Losing My Revolution: A Year After The Egyptian Revolution. `http://ws-dl.blogspot.com/2010/11/2010-11-05-memento-datetime-is-not-last.html` (2012)

[276] SalahEldeen, H.M., Nelson, M.L.: Losing my revolution: How many resources shared on social media have been lost? In: Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries, TPDL '12, pp. 125–137. Springer-Verlag (2012). DOI 10.1007/978-3-642-33290-6_14

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

[278] SalahEldeen, H.M., Nelson, M.L.: Resurrecting my revolution. In: Proceedings of the 3rd International Conference on Theory and Practice of Digital Libraries, TPDL '13, pp. 333–345. Springer Berlin Heidelberg (2013). DOI 10.1007/978-3-642-40501-3_34

[279] Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of (1989)

[280] Salton, G., Wong, A., Yang, C.S.: A Vector Space Model for Automatic Indexing. Communications of the ACM **18**(11), 613–620 (1975). DOI 10.1145/361219.361220

[281] Sandelowski, M.: Telling Stories: Narrative Approaches in Qualitative Research. Journal of Nursing Scholarship **23**(3), 161–166 (1991)

[282] Sanderson, R., AlSum, A.: MementoFox 0.9.52.1. `https://addons.mozilla.org/En-us/firefox/addon/mementofox/` (2011)

[283] Sastry, N.: Predicting pinterest: Organising the world's images with human-machine collaboration. In: Proceedings of the 24th International Conference on World Wide Web, WWW '15 Companion, pp. 1065–1065. International World Wide Web Conferences Steering Committee (2015). DOI 10.1145/2740908.2744719

[284] Schneider, S.M., Foot, K., Kimpton, M., Jones, G.: Building Thematic Web Collections: Challenges and Experiences from the September 11 Web Archive and the Election 2002 Web Archive. In: Proceedings of the 3rd Workshop on Web Archives (2003)

[285] Segel, E., Heer, J.: Narrative visualization: Telling stories with data. Visualization and Computer Graphics, IEEE Transactions on **16**(6), 1139–1148 (2010)

[286] Shankar, H.: Memento Time Travel. `https://chrome.google.com/webstore/detail/memento-time-travel/jgbfpjledahoajcppakbgilmojkaghgm` (2015)

[287] Shaw, E.J.: Rethinking EAD: Balancing flexibility and interoperability. New Review of Information Networking **7**(1), 117–131 (2001). DOI 10.1080/13614570109516972

[288] Shenker, J.: The struggle to document Egypt's revolution. The Guardian, `http://www.theguardian.com/world/2011/jul/15/struggle-to-document-egypt-revolution` (2011)

[289] Shuyo, N.: Language Detection Library for Java. `http://code.google.com/p/language-detection/` (2012)

[290] Sigursson, K.: Incremental crawling with Heritrix. In: Proceedings of the 5th International Web Archiving Workshop (2005). URL `http://www.iwaw.net/05/papers/iwaw05-sigurdsson.pdf`

[291] Silva, A.J.C., Gonçalves, M.A., Laender, A.H.F., Modesto, M.A.B., Cristo, M., Ziviani, N.: Finding What is Missing from a Digital Library: A Case Study in the Computer Science Field. Information Processing and Management **45**(3), 380–391 (2009)

[292] Singhal, A.: Modern information retrieval: A brief overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering **24**(4) (2001). URL http://act.buaa.edu.cn/hsun/IR2013/ref/mir.pdf

[293] Sinha, P.: Summarization of Archived and Shared Personal Photo Collections. In: Proceedings of the 20th International World Wide Web Conference, WWW '11, pp. 421–425. ACM Press (2011). DOI 10.1145/1963192.1963354

[294] Sinha, P., Mehrotra, S., Jain, R.: Effective Summarization of Large Collections of Personal Photos. In: Proceedings of the 20th International Conference Companion on World Wide Web, WWW '11, pp. 127–127. ACM Press (2011). DOI 10.1145/1963192.1963257

[295] Sisodia, D.S., Verma, S.: Web usage pattern analysis through web logs: A review. In: 2012 Ninth International Conference on Computer Science and Software Engineering (JCSSE), pp. 49–53. IEEE (2012). DOI 10.1109/JCSSE.2012.6261924

[296] Six, J.M., Macefield, R.: How to determine the right number of participants for usability studies. UXmatters, http://www.uxmatters.com/mt/archives/2016/01/how-to-determine-the-right-number-of-participants-for-usability-studies.php (2016)

[297] Smith, A.: Home broadband 2010. Tech. rep., Pew Internet & American Life Project, An initiative of the Pew Research Center (2010). URL http://pewinternet.org/Reports/2010/Home-Broadband-2010.aspx

[298] Smith, M., Kanade, T.: Video skimming for quick browsing based on audio and image characterization. Tech. Rep. CMU-CS-95-186, Computer Science Department, Pittsburgh, PA (1995)

[299] Smith, M.A.: Video Skimming and Characterization Through the Combination of Image and Language Understanding Techniques. In: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition, CVPR '97, pp. 775–781. IEEE Computer Society (1997)

[300] Spaniol, M., Weikum, G.: Tracking Entities in Web Archives: The LAWA Project. In: Proceedings of the 21st International Conference Companion on World Wide Web, WWW '12 Companion, pp. 287–290. ACM (2012). DOI 10.1145/2187980.2188030

[301] Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of documentation **28**(1), 11–21 (1972)

[302] Spinellis, D.: The decay and failures of web references. Communications of the ACM **46**(1), 71–77 (2003). DOI 10.1145/602421.602422

[303] Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. ACM SIGKDD Explorations Newsletter **1**(2), 12–23 (2000). DOI 10.1145/846183.846188

[304] Stanoevska-Slabeva, K., Sacco, V., Giardina, M.: Content Curation : a new form of gatewatching for social media? In: Proceeding of the 12th International Symposium on Online Journalism (2012). URL `http://online.journalism.utexas.edu/2012/papers/Katarina.pdf`

[305] Stasko, J., Gorg, C., Liu, Z., Singhal, K.: Jigsaw: Supporting Investigative Analysis through Interactive Visualization. In: Proceedings of IEEE VAST, pp. 131–138 (2007). DOI 10.1109/VAST.2007.4389006

[306] Stassopoulou, A., Dikaiakos, M.D.: Web robot detection: A probabilistic reasoning approach. Computer Networks **53**(3), 265–278 (2009). DOI 10.1016/j.comnet.2008.09.021

[307] Streitfeld, A.: Internet Archive Will Shield Visitors - NYTimes.com. The New York Times Bits Blog, `http://bits.blogs.nytimes.com/2013/10/24/internet-archive-will-shield-visitors/` (2013)

[308] Sutter, J.D.: The faces of Egypt's 'Revolution 2.0'. `http://edition.cnn.com/2011/TECH/innovation/02/21/egypt.internet.revolution/index.html` (2011)

[309] Tan, P.N., Kumar, V.: Discovery of Web Robot Sessions Based on their Navigational Patterns. Data Mining and Knowledge Discovery **6**(1), 9–35 (2002). DOI 10.1023/A:1013228602957

[310] Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining, 1 edn. Addison-Wesley (2005)

[311] Tanahashi, Y., Ma, K.L.: Design considerations for optimizing storyline visualizations. IEEE Transactions on Visualization and Computer Graphics **18**(12), 2679–2688 (2012)

[312] Tanasa, D., Trousse, B.: Advanced data preprocessing for intersites Web usage mining. IEEE Intelligent Systems **19**(2), 59–65 (2004). DOI 10.1109/MIS.2004.1274912

[313] Taskiran, C.M., Pizlo, Z., Amir, A., Ponceleon, D., Delp, E.J.: Automated video program summarization using speech transcripts. IEEE Transactions on Multimedia **8**(4), 775–791 (2006). DOI 10.1109/TMM.2006.876282

[314] Technical Committee ISO/TC 46: The WARC File Format (ISO 28500). `http://bibnum.bnf.fr/warc/WARC_ISO_28500_version1_latestdraft.pdf` (2008)

[315] Teddlie, C., Yu, F.: Mixed Methods Sampling: A Typology With Examples. Journal of Mixed Methods Research **1**(1), 77–100 (2007). DOI 10.1177/2345678906292430

[316] Teevan, J., Dumais, S.T., Liebling, D.J., Hughes, R.L.: Changing How People View Changes on the Web. In: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, UIST '09, pp. 237–246. ACM (2009). DOI 10.1145/1622176.1622221

[317] Thelwall, M., Vaughan, L.: A fair history of the Web? Examining country balance in the Internet Archive. Library & Information Science Research **26**(2), 162–176 (2004)

[318] Theobald, M., Siddharth, J., Paepcke, A.: Spotsigs: Robust and efficient near duplicate detection in large web collections. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, pp. 563–570. ACM (2008). DOI 10.1145/1390334.1390431

[319] Tofel, B.: Wayback for Accessing Web Archives. In: Proceedings of International Web Archiving Workshop. IWAW (2007). URL `http://iwaw.europarchive.org/07/IWAW2007_tofel.pdf`

[320] Truong, B.T., Dorai, C., Venkatesh, S.: New enhancements to cut, fade, and dissolve detection processes in video segmentation. In: Proceedings of the Eighth ACM International Conference on Multimedia, MULTIMEDIA '00, pp. 219–227. ACM (2000). DOI 10.1145/354384.354481

[321] Truong, B.T., Venkatesh, S.: Video abstraction: A systematic review and classification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **3**(1) (2007). DOI 10.1145/1198302.1198305

[322] Tuffley, D., Antonio, A.: First Year Engagement & Retention: A Goal-Setting Approach. Journal of Information Technology Education: Innovations in Practice **12**, 239–251 (2013)

[323] Tufte, E.R.: The Visual Display of Quantitative Information. Graphics Press (1986)

[324] Turing, A.M.: Computing machinery and intelligence. Mind **59**(236), 433–460 (1950)

[325] Tweedy, H., McCown, F., Nelson, M.L.: A Memento Web Browser for iOS. In: Proceedings of the 13th ACM/IEEE Joint Conference on Digital Libraries, pp. 371–372 (2013)

[326] Uchihashi, S., Foote, J., Girgensohn, A., Boreczky, J.: Video manga: Generating semantically meaningful video summaries. In: Proceedings of the Seventh ACM International Conference on Multimedia (Part 1), MULTIMEDIA '99, pp. 383–392. ACM (1999). DOI 10.1145/319463.319654

[327] Van de Sompel, H., Nelson, M.L., Sanderson, R.: RFC 7089 - HTTP framework for time-based access to resource states – Memento. `http://tools.ietf.org/html/rfc7089` (2013)

[328] Van de Sompel, H., Nelson, M.L., Sanderson, R., Balakireva, L.L., Ainsworth, S., Shankar, H.: Memento: Time Travel for the Web. Tech. Rep. arXiv:0911.1112 (2009)

[329] Van de Sompel, H., Sanderson, R., Nelson, M.L., Balakireva, L.L., Shankar, H., Ainsworth, S.: An HTTP-Based Versioning Mechanism for Linked Data. In: Proceedings of the 3rd Linked Data on the Web Workshop (2010)

[330] Viégas, F.B., Wattenberg, M., Feinberg, J.: Participatory Visualization with Wordle. IEEE transactions on visualization and computer graphics **15**(6), 1137–44 (2009). DOI 10.1109/TVCG.2009.171

[331] Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M.J., Zheng, H., Zhao, B.Y.: Social turing tests: Crowdsourcing sybil detection. In: NDSS. The Internet Society (2013)

[332] Wasserman, T.: Netflix takes up 32.7% of Internet bandwidth. Mashable, `http://mashable.com/2011/10/27/netflix-takes-up-32-7-of-internet-bandwidth-study/` (2011)

[333] Wayne, C.L.: Topic Detection & Tracking (TDT). In: Proceedings of the Broadcast News Transcription and Understanding Workshop (1997). DOI 10.1.1.27.2955

[334] Wei, F., Liu, S., Song, Y., Pan, S., Zhou, M.X., Qian, W., Shi, L., Tan, L., Zhang, Q.: TIARA: A Visual Exploratory Text Analytic System. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, pp. 153–162 (2010)

[335] Weiss, R.: On the Web, Research Work Proves Ephemeral. `http://stevereads.com/cache/ephemeral_web_pages.html` (2003)

[336] Whitelaw, M.: Exploring Archival Collections with Interactive Visualisation. In: Proceedings of E-Research Australasia Conference (2009)

[337] Williams, C.: How Egypt shut down the internet. The Telegraph, `http://www.telegraph.co.uk/news/worldnews/africaandindianocean/egypt/8288163/How-Egypt-shut-down-the-internet.html` (2011)

[338] Yan, L., Zhou, X., Lu, L., Centeno, A.G., Kuan, L., Hawrylycz, M., Larimer, M., Rosen, G.D., Williams, R.W.: Global exploratory analysis of massive neuroimaging collections using Microsoft Silverlight PivotViewer. In: Proceedings of the 2011 Biomedical Sciences and Engineering Conference: Image Informatics and Analytics in Biomedicine, pp. 1–4. IEEE (2011). DOI 10.1109/BSEC.2011.5872323

[339] Yang, Y., Carbonell, J.J.G., Brown, R.R.D., Pierce, T., Archibald, B.T.B., Liu, X.: Learning Approaches for Detecting and Tracking News Events. Intelligent Systems and their Applications, IEEE **14**(4), 32–43 (2000). DOI 10.1109/5254.784083

[340] Yang, Y., Pierce, T., Carbonell, J.: A study of retrospective and on-line event detection. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98, pp. 28–36. ACM Press (1998). DOI 10.1145/290941.290953

[341] Yeung, M.M., Liu, B.: Efficient Matching and Clustering of Video Shots. In: Proceedings of the 1995 International Conference on Image Processing, *ICIP '95*, vol. 1, pp. 338 – 341. IEEE Computer Society, Washington, DC, USA (1995)

[342] Yih, W., Goodman, J., Carvalho, V.R.: Finding advertising keywords on web pages. In: Proceedings of the 15th International World Wide Web Conference, WWW '06, p. 213. ACM Press (2006). DOI 10.1145/1135777.1135813

[343] Yin, Z., Shokouhi, M., Craswell, N.: Query Expansion Using External Evidence. In: Advances in Information Retrieval, pp. 362–374. Springer (2009)

[344] Youssef, A.: A Critical Analysis on Media Coverage of the Egyptian Revolution: The Case of Al-Ahram, Al-Masry Al-Youm, The Telegraph and The Washington Post. Master's thesis, Örebro University (2012)

[345] Zarro, M., Hall, C.: Exploring Social Curation. D-Lib Magazine **18**(11), 6 (2012)

[346] Zhang, H.J., Wu, J., Zhong, D., Smoliar, S.W.: An integrated system for content-based video retrieval and browsing. Pattern recognition **30**(4), 643–658 (1997)

[347] Zhong, C., Shah, S., Sundaravadivelan, K., Sastry, N.: Sharing the Loves: Understanding the How and Why of Online Content Curation. Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM) (2013)

[348] Zhuang, Z., Wagle, R., Giles, C.L.: What's There and What's Not? Focused Crawling for Missing Documents in Digital Libraries. In: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '05, pp. 301–310 (2005)

VITA

Yasmin AlNoamany
Department of Computer Science
Old Dominion University
Norfolk, VA 23529

EDUCATION

Ph.D.	Computer Science, Old Dominion University, 2016
M.S.	Computer Science, Mansoura University, 2009
B.S.	Computer Science, Mansoura University, 2006

EMPLOYMENT

02/2011 - Present	Research Assistant, Old Dominion University
07/2014 - 12/2014	Software Engineer Intern, Internet Archive

PUBLICATIONS

A complete list is available at `https://scholar.google.com/citations?user=bRlYmNcAAAAJ&hl=en`

PROFESSIONAL SOCIETIES

Arab Women In Computing (ArabWIC)

Typeset using LaTeX.