Detecting off-topic pages within TimeMaps in Web archives

Yasmin AlNoamany[1] · **Michele C. Weigle**[1] · **Michael L. Nelson**[1]

Received: 10 January 2016 / Revised: 24 June 2016 / Accepted: 4 July 2016 / Published online: 18 July 2016

Abstract Web archives have become a significant repository of our recent history and cultural heritage. Archival integrity and accuracy is a precondition for future cultural research. Currently, there are no quantitative or content-based tools that allow archivists to judge the quality of the Web archive captures. In this paper, we address the problems of detecting when a particular page in a Web archive collection has gone off-topic relative to its first archived copy. We do not delete off-topic pages (they remain part of the collection), but they are flagged as off-topic so they can be excluded for consideration for downstream services, such as collection summarization and thumbnail generation. 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 the number of words and content length) to detect when a page has gone off-topic. Those predicted off-topic pages will be presented to the collection's curator for possible elimination from the collection or cessation of crawling. We created a gold standard data set from three Archive-It collections to evaluate the proposed methods at different thresholds. 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. 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 0.89.

✉ Yasmin AlNoamany
yasmin@cs.odu.edu

Michele C. Weigle
mweigle@cs.odu.edu

Michael L. Nelson
mln@cs.odu.edu

[1] Department of Computer Science, Old Dominion University,
Norfolk, VA 23529, USA

Keywords Web archiving · Document filtering · Information retrieval · Document similarity · Archived collections · Web content mining · Internet Archive

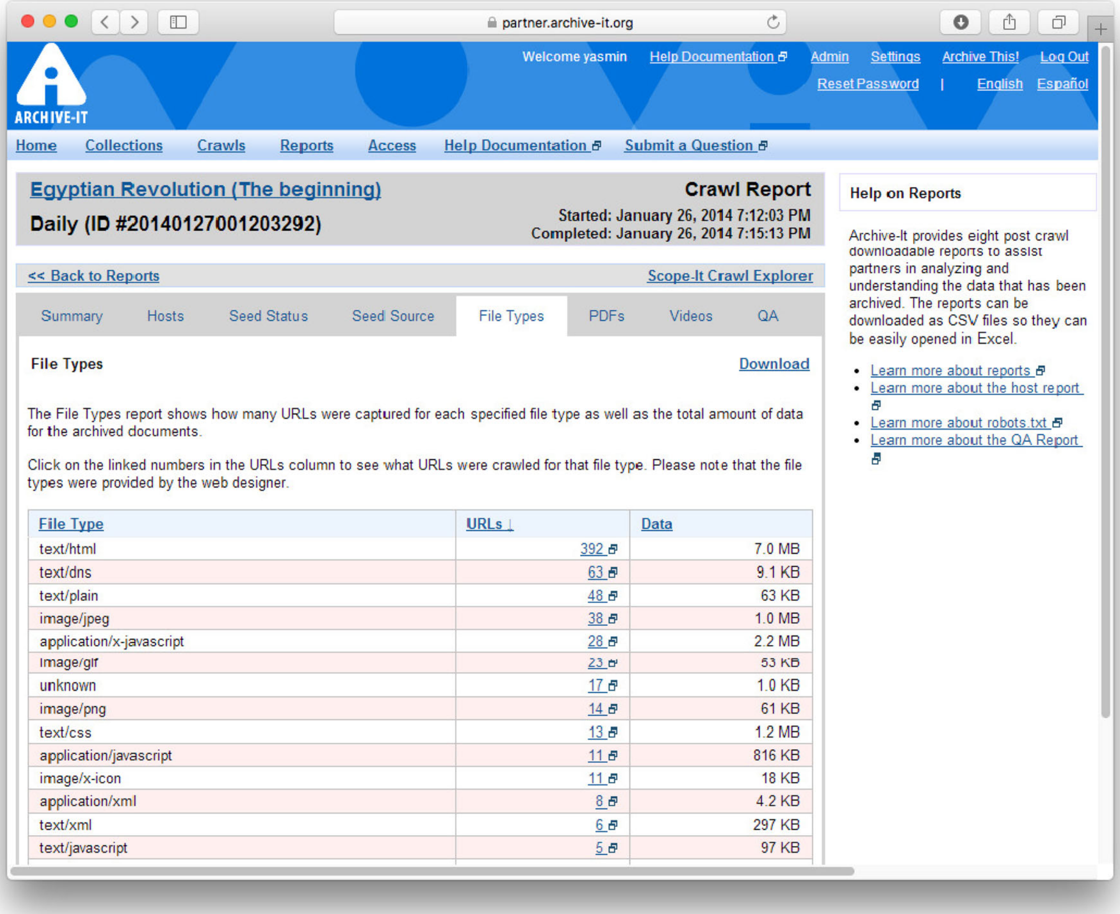
1 Introduction

Much of our cultural discourse occurs primarily on the Web and its preservation is a fundamental pre-condition for research in history, sociology, political science, media, literature, and other related disciplines [41]. We consider archival existence and integrity to be a requirement for any scholarly pursuit that involves Web pages. Since mining the past Web is different from traditional data mining [23], it is necessary to have novel approaches for utilizing the content of page histories for knowledge-discovery purposes and to assist the curators of archived collections to create quality collections that will be ready for researchers and practitioners.

The Internet Archive [40] (IA) is the largest and oldest of the various Web archives, holding over 400 billion Web pages (as of November 2015) with archives as far back as 1996 [25]. Archive-It[1] is a collection development service that has been operated by the Internet Archive since 2006. As of November 2015, Archive-It has been used by over 340 institutions in 48 states, and featured over 9 billion archived Web pages in nearly 3200 separate collections. Archive-It's collections have three main categories. First, there are collections that are devoted to archiving governmental pages (e.g., all Web pages published by the State of South Dakota[2]). Second, there

[1] https://archive-it.org/.

[2] https://archive-it.org/collections/192.



Egyptian Revolution (The beginning)

Daily (ID #20140127001203292)

Crawl Report

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

Help on Reports

Archive-It provides eight post crawl downloadable reports to assist partners in analyzing and understanding the data that has been archived. The reports can be downloaded as CSV files so they can be easily opened in Excel.

- Learn more about reports
- Learn more about the host report
- Learn more about robots.txt
- Learn more about the QA Report

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. 1 Archive-It provides the collection curators with information about their crawls

are collections that are event based (e.g., Occupy Movement collection[3] and SOPA Blackout collection[4]). Third, there are theme-based collections (e.g., the Columbia Human Rights collection[5]).

Archive-It provides its 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 *seeds*, uniform resource identifiers (URIs) that should be crawled periodically (the frequency is tunable by the user), and to what depth (e.g., follow the pages linked to from the seeds two levels out). Archive-It also creates collections of global events under the name of Internet Archive Global Events. The seed URIs by asking the people to nominate URIs that are related to these events. The seed URIs are manually collected by asking people to nominate URIs that are related to these events, or are selected by the collection's curator(s).

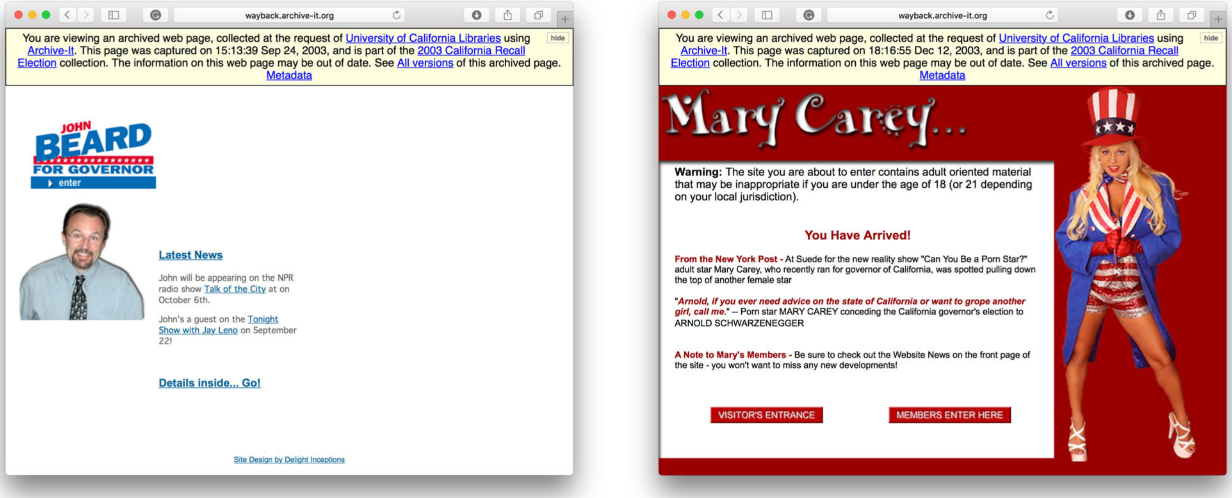
The Heritrix [39] crawler at Archive-It crawls or recrawls these seeds based on the predefined frequency and depth to build a collection of archived Web pages that the curator believes best exemplifies the topic of the collection. Archive-It has deployed tools that allow collection's curators to perform quality control on their crawls, as shown in Fig. 1. 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. We define off-topic archived pages as those that have changed through time to move away from the scope of the initial seed URI, which should be relevant to the topic of the collection.

In previous work [3] and extended in this paper, we introduce different approaches for detecting off-topic pages of the seed URI in the archives. The approaches depend

[3] https://archive-it.org/collections/2950.

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

[5] https://archive-it.org/collections/1068.



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

(a)

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!

(b)

Fig. 2 Example of johnbeard4gov.com in the 2003 California Recall Election collection that went off-topic. **a** September 24, 2003: johnbeard4gov.com was for a California gubernatorial candidate. **b** December 12, 2003: johnbeard4gov.com became spam

on comparing the versions of the pages through time. We assume that the first copies of seed URIs were originally on-topic for the topic of the collection, since they were selected manually and judged by the collection curator. 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 the number of words and the content length). In this paper, we extend our analysis by providing more details of how these methods work. We also investigate how a page's aboutness changes through time, based on a data set from Archive-It. We found five different behaviors for how the aboutness of a page changes between on-topic and off-topic. We also quantified these behaviors based on a data set from archived collections.

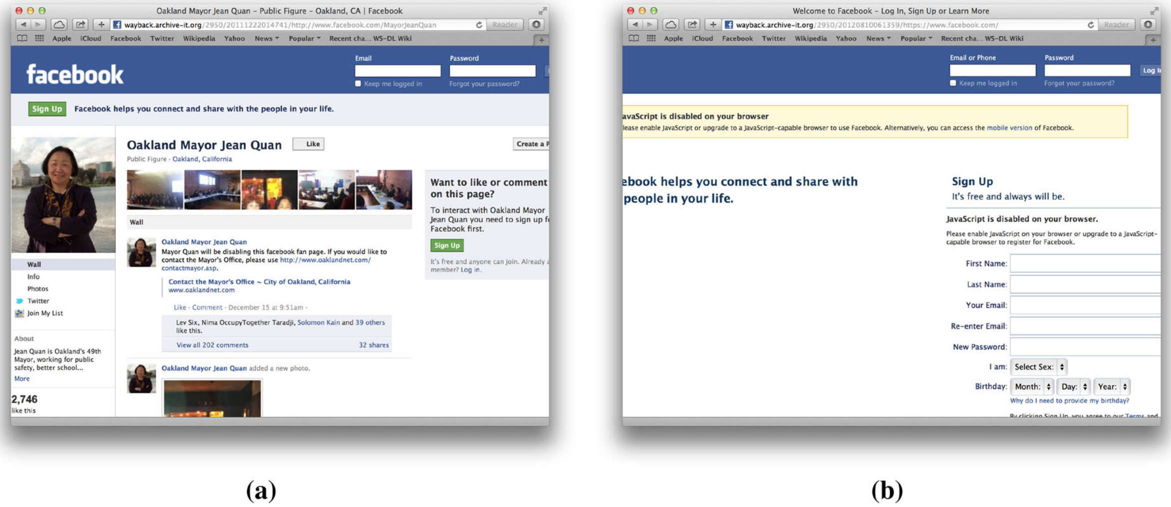
For evaluation purposes, we built our gold standard data set from three Archive-It collections, then we employ the following performance measurements: accuracy, F_1 score values, and area under the ROC curve (AUC). Experimental results show that cosine similarity at the 0.15 threshold is the most effective single method for detecting the off-topic pages with 0.983 accuracy, followed by the word count at threshold -0.85 with 0.982 accuracy. We combined the tested methods and found that the best-performing combined method across the three collections is cosine at threshold 0.10 with word count at threshold -0.85. Cosine and word count combined improved the performance over cosine alone with a 3 % increase in the F_1 score, 0.7 % increase in AUC, and 0.4 % increase in accuracy. We chose cosine at threshold 0.10 combined with word count at threshold -0.85, and evaluated the performance on a different set of Archive-It collections. Based on manual assessment of the detected off-topic pages, the average precision of the proposed technique for the tested collections is 0.9.

2 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 [37]. 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 Fig. 2). In this case, the Web page returns a 200 HTTP response, but with unwanted content [8].

There are also many cases in which the archived page redirects to another page which is not relevant, but still not spam. In Fig. 3, the Facebook page contained relevant content in the beginning (Fig. 3a), then later redirects to the homepage of Facebook as shown in Fig. 3b. The example in Fig. 3 shows how a page in a collection goes off-topic, even though the particular Web site has not been lost.

Figure 4 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


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
About
More
Sign Up
(a)
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 I
Keep me logged in
Forgot your password?
Select Sex:
Month:
Day:
Year:
(b)

Facebook helps you connect and share with the people in your life.

Oakland Mayor Jean Quan

Public Figure · Oakland, California

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.

It's free and anyone can join. Already a member? Log in.

Jean Quan is Oakland's 49th Mayor, working for public safety, better school...

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?

Fig. 3 Example of a Facebook page from the Occupy Movement collection that went off-topic. **a** December 22, 2011: Facebook page was relevant to the Occupy collection. **b** August 10, 2012: URI redirects to http://www.facebook.com

(Fig. 4a). 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 Fig. 4. In the example, the page went off-topic because of a database error on May 24, 2012 (Fig. 4b), and then it returned on-topic again. After that, the page went off-topic because of financial issues (Fig. 4c). The page continued off-topic for a long period (from March 27, 2013 until July 2, 2013) because the site was under construction (Fig. 4d). The page went on-topic again for a period of time, then the site was hacked (Fig. 4e), and the domain was lost by late 2014 (Fig. 4f).

The Web page hamdeensabahy.com has 266 archived versions, or 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 Fig. 4b–f do not contribute to the Egypt Revolution collection in the same way as the on-topic archived Web site in Fig. 4a.

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 archivist keeps them in the collection, once identified, these off-topic pages can be excluded from story summarization services [1,2] or thumbnail generation [6].

3 Background

Despite the fact that Web archives present a great potential for knowledge discovery, there has been relatively little research that is explicitly aimed at mining content stored in Web archives [23]. In this section, we highlight the research that has been conducted on mining the past Web. First, we define the terminology that will be adopted throughout the rest of the paper.

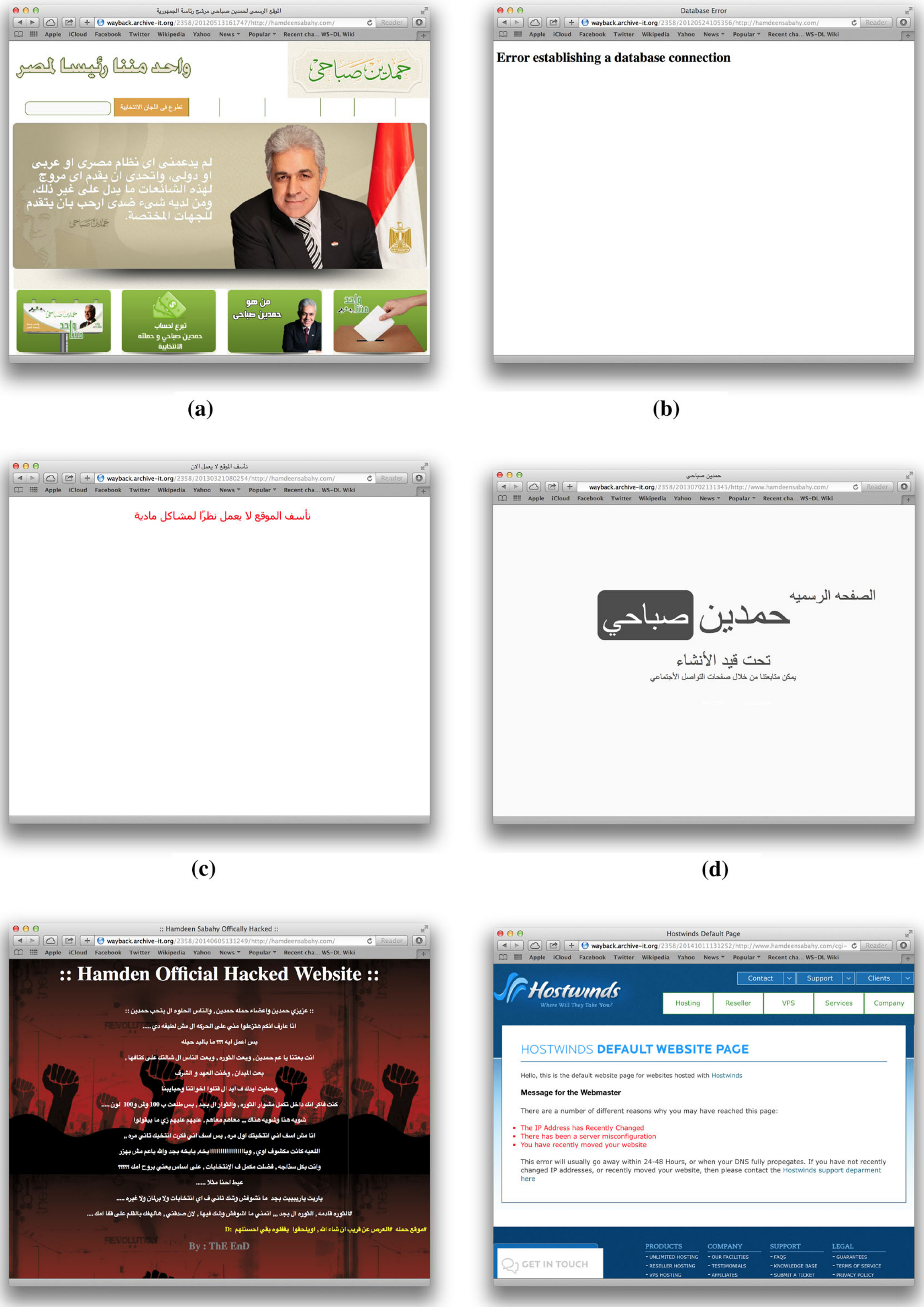
3.1 Memento terminology

Memento [55] is an HTTP protocol extension that enables time travel on the Web by linking the current resources with their prior state. Memento defines the following terms:

- URI-R identifies the original resource. It is the resource as it used to appear on the live Web. A URI-R may have 0 or more mementos (URI-Ms).
- URI-M identifies an archived snapshot of the URI-R at a specific datetime, which is called Memento-Datetime, e.g., URI-M$_i$ = URI-R@t_i. We refer to an archived snapshot as a "memento".
- URI-T identifies a TimeMap, a resource that provides a list of mementos (URI-Ms) for a URI-R with their Memento-Datetimes

3.2 Related work

Because the Web is a dynamic information space, the resources may change, disappear, and move from one location to another frequently [26,31]. Many studies have shown that the expected lifetime of a Web page is short (between 44 and 190 days) [11,24,34,37] and that Web resources disappear quickly [28,30,48]. 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



wayback.archive-it.org/2358/20120513161747/http://hamdeensabahy.com/
Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular Recent cha... WS-DL Wiki
Reader

الموقع الرسمى لحمدين صباحى مرشح رئاسة الجمهورية

حمدين صباحى

واحد مننا رئيسا لمصر

تطوع فى اللجان الانتخابية

لم يدعمنى اى نظام مصرى او عربى او دولى، واتحدى ان يقدم اى مروج لهذه الشائعات ما يدل على غير ذلك، ومن لديه شيء ضدى ارحب بان يتقدم للجهات المختصة.

تبرع لحساب حمدين صباحي و حملته الانتخابية

من هو حمدين صباحى

(a)

wayback.archive-it.org/2358/20120524105356/http://hamdeensabahy.com/

Database Error

Error establishing a database connection

(b)

wayback.archive-it.org/2358/20130321080254/http://hamdeensabahy.com/

نأسف الموقع لا يعمل الان

نأسف الموقع لا يعمل نظراً لمشاكل مادية

(c)

wayback.archive-it.org/2358/20130702131345/http://www.hamdeensabahy.com/

حمدين صباحي

الصفحه الرسميه

حمدين صباحي

تحت قيد الأنشاء

يمكن متابعتنا من خلال صفحات التواصل الأجتماعي

(d)

wayback.archive-it.org/2358/20140605131249/http://hamdeensabahy.com/

:: Hamdeen Sabahy Offically Hacked ::

:: 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 , وبا!!!!!!!!!!!!!!!!!!ايخه بايخه بجد والله ياعم مش بهزر

وانت بكل سذاجه , فضلت مكمل ف الانتخابات , على اساس يعني بروح امك ؟؟؟؟

عبط احنا مثلا ......

ياريت يارييييت بجد ما نشوفش وشك تاني ف اي انتخابات ولا برلمان ولا غيره .....

#الثوره قادمه , الثوره ال بجد ,,, اتمنى ما اشوفش وشك فيها , لان صدقني , هالهفك بالقلم على قفا امك ....

#موقع حمله #العرض عن قريب ان شاء الله , اويلحقوا يقفلوه بقي احسنلهم :D

By : ThE EnD

(e)

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 Page

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)

Fig. 4 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.**a** May 13, 2012: The page started as on-topic. **b** May 24, 2012: Off-topic due to a database error. **c** March 21, 2013: Not working because of financial problems. **d** July 2, 2013: under construction. **e** June 5, 2014: The site has been hacked. **f** October 10, 2014: The domain has expired

parties to access this information. Archiving the political process has become popular, both in terms of Web pages [19,45,50] and YouTube and blogs [13,36]. Web archives are becoming important for studies in social science and humanities research [1]. It is a challenge for Web archiving institutes to balance between the completeness and quality of archived materials while avoid wasting time and space for storing and indexing [46].

The limitation of the resources of Web archives, such as storage and site politeness rules, attracted much attention from many researchers toward optimize the processes of Web archiving lifecycle such as the selection, storage and preservation [4,5,9,15,16,43,46].

Ben Saad et al. [46] 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.

Moreover, focused crawling has become an active area of research to make such collection-building crawls more effective [9,14,17]. As Bergmark et al. mentioned [9], the goal of the focused crawl is to make a "best-first" crawl of the Web.

The previous techniques 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 [6] 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.

Mining the past Web is different from Web content mining because of the temporal dimension of the archived content [23,33,53]. Despite nearly two decades of Web history, there has not been much research conducted for mining Web archive data. The benefit of utilizing the Web archives for knowledge discovery has been discussed many times [7,21,23]. Below, we outline some of the approaches that have been used for mining the past Web using data in Web archives.

Jatowt and Tanaka [23] 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. [22] 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 with simple highlighting techniques.

To help people in understanding Web content change, Teevan et al. [54] introduced DiffIE, a browser plug-in that caches the page a user visits, and then detects and highlights any changes to that page since the user's last visit. They compared the Document Object Model representation of the 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, we are not looking for a deep reading between versions, but rather flagging off-topic pages for non-consideration for other processes (e.g., story summarization [1] and thumbnail generation [6]).

AlSum and Nelson [6] proposed various summarization techniques to optimize the thumbnail creation for TimeMaps based on information retrieval techniques. They found that SimHash similarity fingerprints have the best prediction for the visualization change. They proposed three algorithms, threshold grouping, K clustering, and time normalization. They minimized the number of generated thumbnails by 9–27 % on average.

Spaniol and Weikum used Web archive data to track the evolution of entities (e.g., people, places, things) through time and visualize them [52]. 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. [21] also utilized the public archival repositories for automatically detecting the age of Web content through the past snapshots of pages.

Recently, Odijk et al. [42] introduced tools for selecting, linking, and visualizing the second World War (WWII) collection from collections of the NIOD,[6] the National Library of the Netherlands, and Wikipedia. They also link digital collections via implicit events, i.e., if two articles are close in time and similar in content, they are considered to be related. Furthermore, they provided an exploratory interface to explore the connected collections. They used Manhattan distance for textual similarity over document terms in

[6] http://www.niod.nl/en.

Table 1 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	6341

a TF-IDF weighted vector space and measured temporal similarity using a Gaussian decay function. They found that textual similarity performed better than temporal similarity, and combining textual and temporal similarity improved the nDCG score.

Web archiving research has focused on the selection, storage, and preservation of Web content and solving the challenges that face them [38]. 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. The focus of this paper is on assisting curators in identifying the pages that are off-topic in a TimeMap.

4 Data set

In this section, we describe our gold standard data set. We evaluate our techniques using the ODU mirror of Archive-It's collections. ODU has received a copy of the Archive-It collections through April 2013 in Web ARchive file format (WARC) [20]. The three collections in our data set 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 December 2011 and October 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 February 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) 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 6341 mementos.

Table 2 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	6570	6886	2304
Off-topic URI-Ms	458 (7 %)	384 (9 %)	94 (4 %)
URI-Rs with off-topic URI-Ms	67 (26 %)	34 (25 %)	33 (17 %)

Table 1 provides the details of the three collections. The time span 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.

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 2 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 data set 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 of their collections, but also save the time required for a manual

Table 3 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

check of the relevance of the URIs and the storage required for these pages.

5 TimeMap behavior

Many studies have been performed on the key aspects of document "aboutness", such as the page's title [27], tags [26], and lexical signatures [29]. Section 6.2 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 3 shows the percentage of each type of TimeMap present in our three manually labeled 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 many mementos ($URI\text{-}Ms$), and a set of mementos for a seed URI composes a TimeMap ($URI\text{-}T$). A collection C can be formally defined as the following:

$$\begin{aligned} C &= \{URI\text{-}R_1, URI\text{-}R_2, \ldots, URI\text{-}R_n\} \text{ and} \\ &\forall URI\text{-}R_i \in C, \exists URI\text{-}T \text{ where} \\ &URI\text{-}T = \{URI\text{-}M_0, URI\text{-}M_1, \ldots, URI\text{-}M_n\} \\ &\text{and } URI\text{-}M_i \text{ is } URI\text{-}R@t_i \end{aligned} \tag{1}$$

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 data set (Sect. 4), 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 (Fig. 5a):

$\forall t\ aboutness(URI\text{-}R@t) \approx aboutness(URI\text{-}R@t_0)$, and $URI\text{-}R@t_0$ is relevant to C.

This is the majority case in the gold standard data set, with at least 74 % of the TimeMaps always on-topic (Table 3).

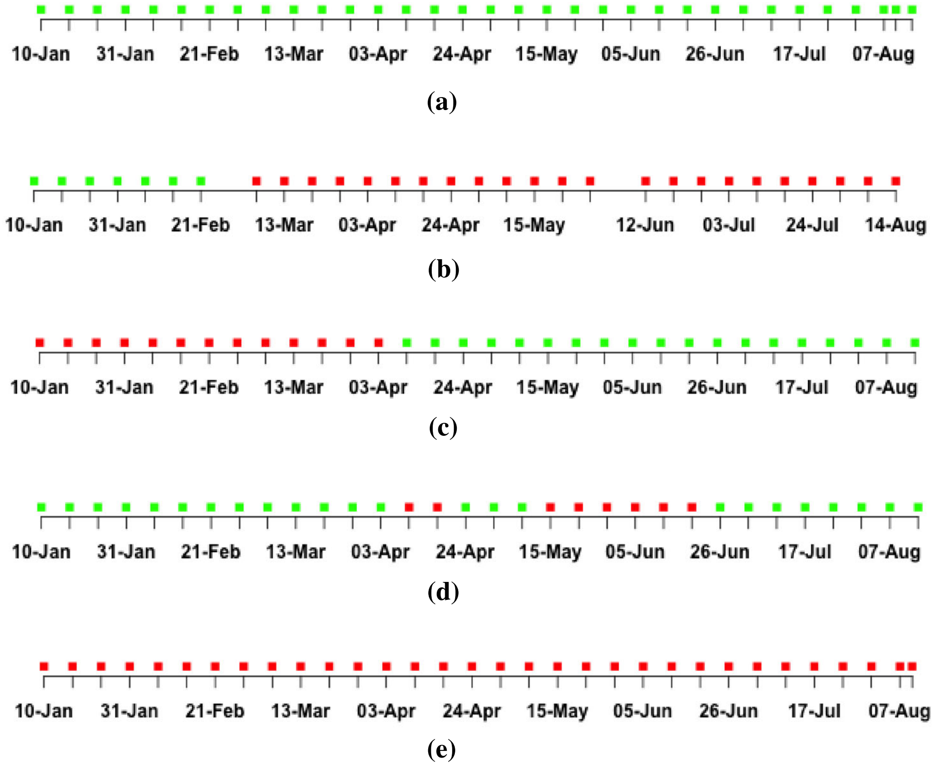


Fig. 5 Example showing different behaviors for TimeMaps (*green* = on-topic, *red* = off-topic). **a** Always On http://wayback.archive-it.org/2950/*/, http://occupypsl.org. **b** Step Function On http://wayback.archive-it.org/2950/*/, http://occupygso.tumblr.com. **c** Step Function Off http://wayback.archive-it.org/2950/*/, 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 (color figure online)

Step Function On: $URI\text{-}R@t_0$ is on-topic, but then at some t goes off-topic and continues $\forall t$ (Fig. 5b):

$\forall t \geq i$, 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 $\forall t$ (Fig. 5c):

$\forall t \geq i$, 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 where the TimeMap starts with an off-topic memento and 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 data set, 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 (Fig. 5d):

$\exists t$ where$(URI\text{-}R@t) \not\approx aboutness(URI\text{-}R@t+i)$ and $aboutness(URI\text{-}R@t) \approx aboutness(URI\text{-}R@t-j)$ where $i, j \geq 0$ and i, j are integers.

We found that 9–13 % of the TimeMaps were Oscillating between on-topic and off-topic.

Always Off: This is the most challenging case, where all the mementos are off-topic (Fig. 5e):

$\forall\, t, URI\text{-}R@t$ is *not* relevant to C.

We manually identified these cases (totaling 3 seed URIs) and excluded these from the gold standard data set. 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.

6 Research approach

In this section, we explain the methodology for preparing the data set and then the methodology for applying different measures to detect the off-topic pages.

6.1 Data set preprocessing

We applied the following steps to prepare the gold standard data set:

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.[7]
3. Extract the HTML of the mementos from the WARC files (locally hosted at ODU).
4. Extract the text of the page using the Boilerpipe library [32].
5. Extract terms from the page, using scikit-learn [44] to tokenize, remove stop words, and apply stemming.

6.2 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 [35] 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, and document filtering. Cosine similarity measures the cosine of the angle between two vectors (d_1 and d_2) by taking the dot product between them [49,51]:

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\| d_1 \| \| d_2 \|}. \tag{2}$$

After text preprocessing, we calculated the TF-IDF for mementos, and 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 [35]. The Jaccard between set A and set B is formulated as follows:

$$J(A, B) = \frac{A \cap B}{A \cup B}. \tag{3}$$

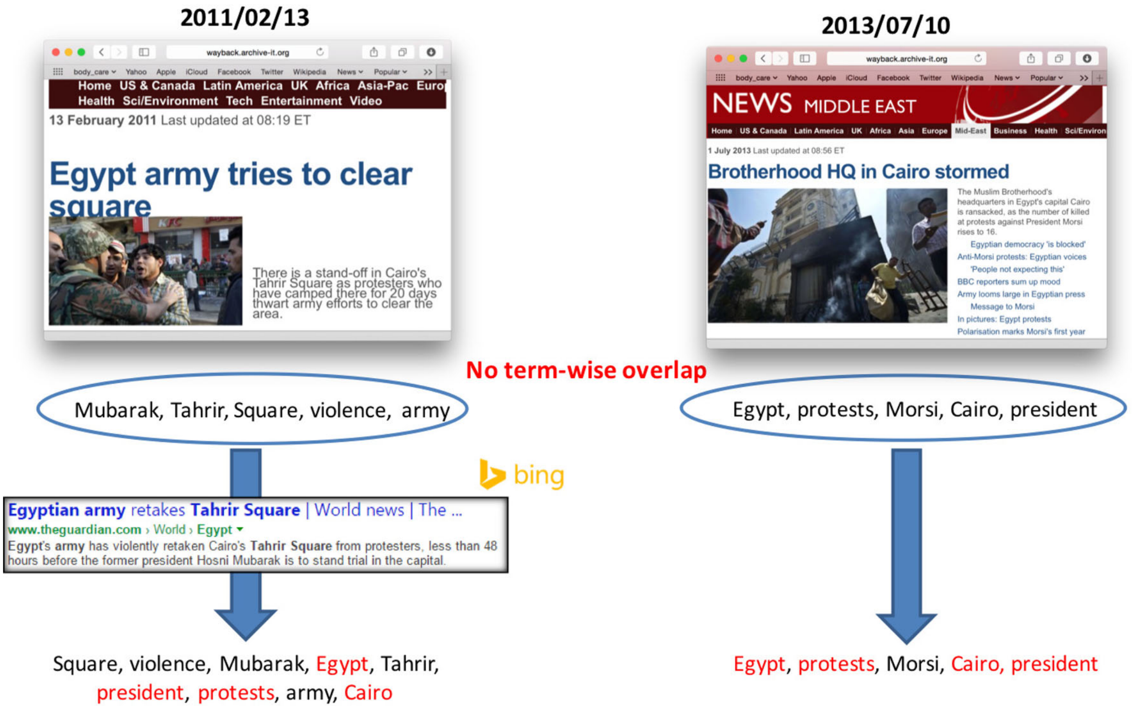
After preprocessing the text (result from step 5), we apply the Jaccard coefficient on the resulting terms to specify the similarity between the $URI\text{-}R@t$ and $URI\text{-}R@t_0$.

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 of the intersection between the

[7] http://archive.org/web/researcher/cdx_file_format.php.


2011/02/13
wayback.archive-it.org
Home US & Canada Latin America UK Africa Asia-Pac Euro
Health Sci/Environment Tech Entertainment Video
13 February 2011 Last updated at 08:19 ET
Egypt army tries to clear square
2013/07/10
wayback.archive-it.org
NEWS MIDDLE EAST
1 July 2013 Last updated at 08:56 ET
Brotherhood HQ in Cairo stormed
No term-wise overlap
Mubarak, Tahrir, Square, violence, army
Egypt, protests, Morsi, Cairo, president
bing
Egyptian army retakes Tahrir Square | World news | The ...
www.theguardian.com › World › Egypt
Square, violence, Mubarak, Egypt, Tahrir, president, protests, army, Cairo
Egypt, protests, Morsi, Cairo, president

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.

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.

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. 6 An example for increasing the semantic context by the Web-based kernel function using a search engine (SE)

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

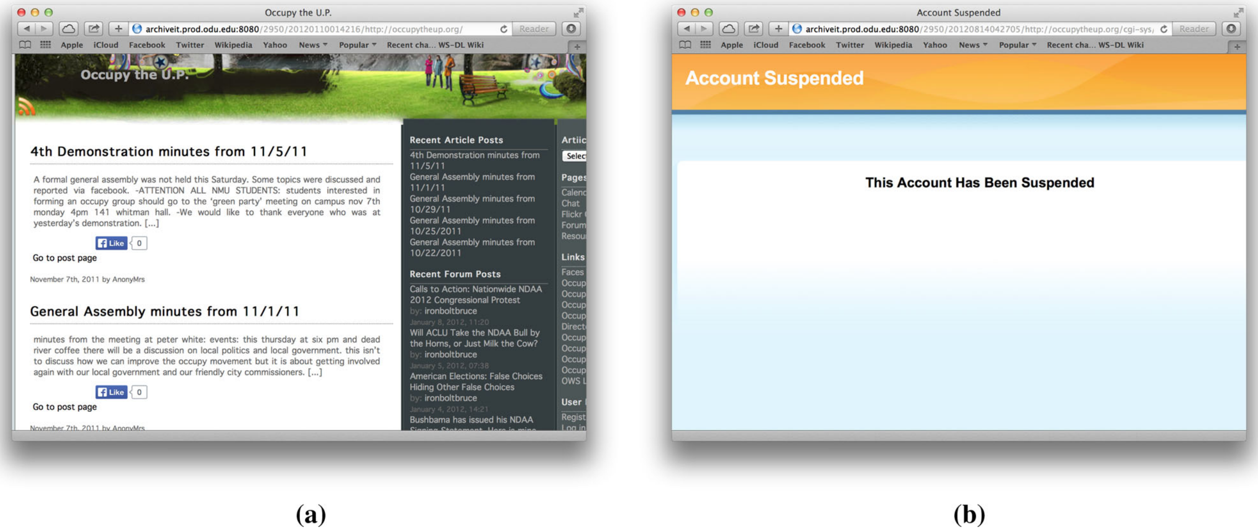
The previous methods are term-wise similarity measures, i.e., they use lexicographic term matching. But these methods may not be suitable for archived collections with a large time span or pages that contain a small amount of text. For example, the Egyptian Revolution collection is from February 2011 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 [47] 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 [12], 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 [56]. 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 ten 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 6 shows an example of how we apply the Web-based kernel function on a memento from the Egyptian Revolution collection. As the figure illustrates, we use terms "Mubarak, Tahrir, Square, violence, army" of the first memento as search keywords to generate the 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 Fig. 6 after extending the terms of the first memento is 0.4.



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. [...]

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. [...]

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)

Account Suspended

This Account Has Been Suspended

(b)

Fig. 7 Later versions of occupytheup.org are off-topic **a** Occupy the U.P. on January 10, 2012. **b** Expired on August 14, 2012, but no textual content

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 five 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 by $d(A, B)$, can be defined formally as follows:

$$d(A, B) = 1 - \frac{s(A)}{s(B)}, \quad (4)$$

where s is the size of document.

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 Fig. 3 went off-topic in Fig. 3b and has 62 KB, but the on-topic page in Fig. 3a is nearly similarly sized with 84 KB. Using bytes is recommended when the size of the collection is large and the curator wants to detect potential off-topic pages in a short time.

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, has transient errors, or has 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 7 has an example where the account is suspended and the size of the page is almost 0 bytes.

7 Evaluation

In this section, we define how we evaluate the methods presented in Sect. 6.2 on our gold standard data set. Based on these results, we define a threshold th for each method for when a memento becomes off-topic.

7.1 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 numbers of off-topic pages that are predicted as on-topic
- Accuracy (ACC), the fractions of the classifications that are correct.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})} \quad (5)$$

Table 4 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

Bold values indicate the best performing method based on the highest F1

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.

$$F_1 = \frac{2\text{TP}}{(2\text{TP} + \text{FP} + \text{FN})} \tag{6}$$

- The ROC AUC score, a single number that computes the area under the receiver operating characteristic (ROC) [18] curve, which is also denoted by AUC.

7.2 Results

We tested each method with 21 thresholds (378 tests for three collections) on our gold standard data set to estimate which threshold for each method is able to separate the off-topic from the on-topic pages. 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 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 $th = -0.95$.

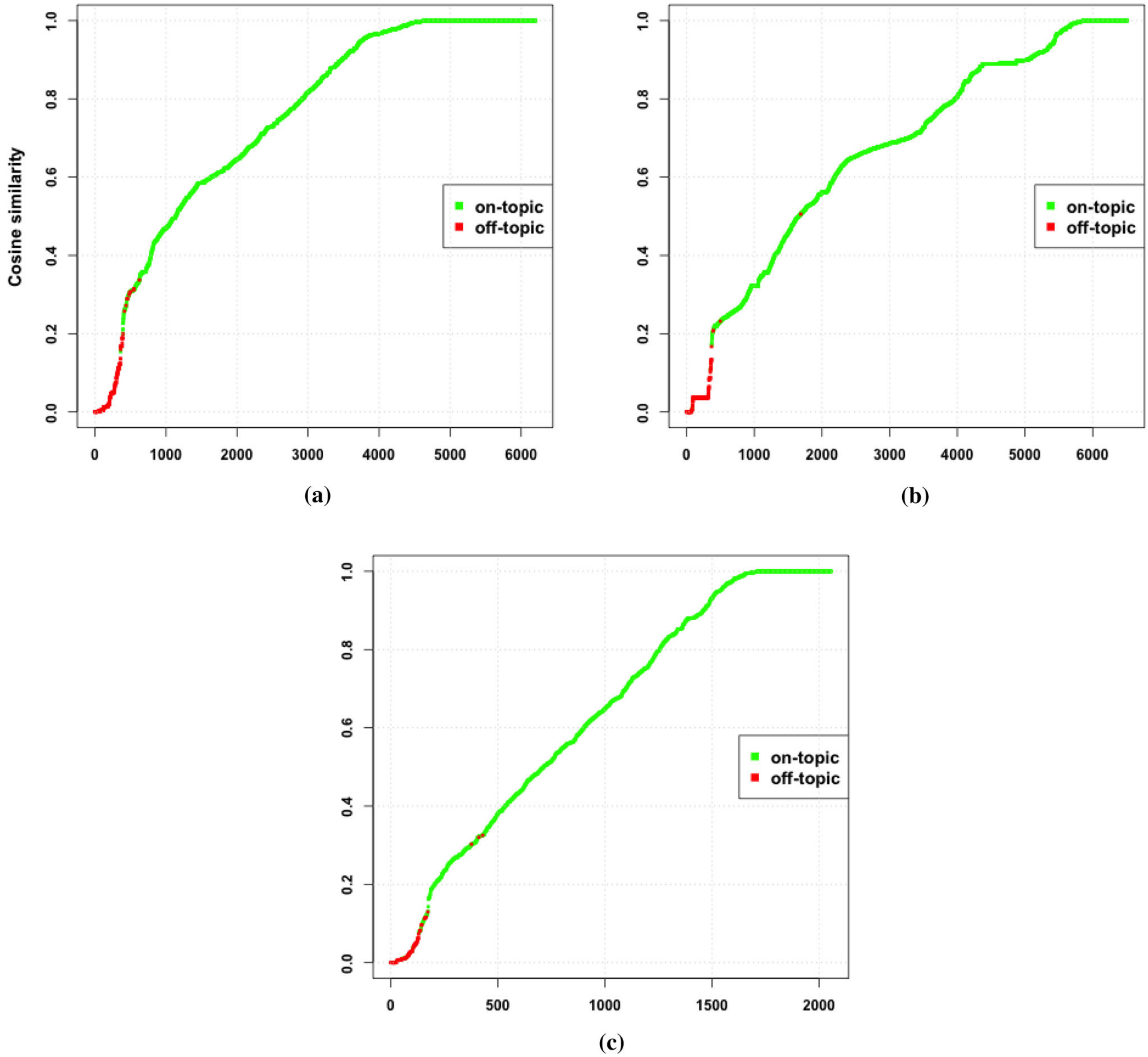
Table 4 contains the summary of running the similarity approaches on the three collections. The table shows the best result based on the F_1 score at the underlying threshold measures averaged on all three collections. From the table, the best-performing measure is Cosine with average ACC = 0.983, F_1 = 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 8 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 $th = 0.40$. Figure 9 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., −(80–100) % 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 0.20, 0.15, and 0.10, respectively.

We took the average of the three collections at $th = 0.20$, $th = 0.15$, and $th = 0.10$; 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.



(a)

(b)

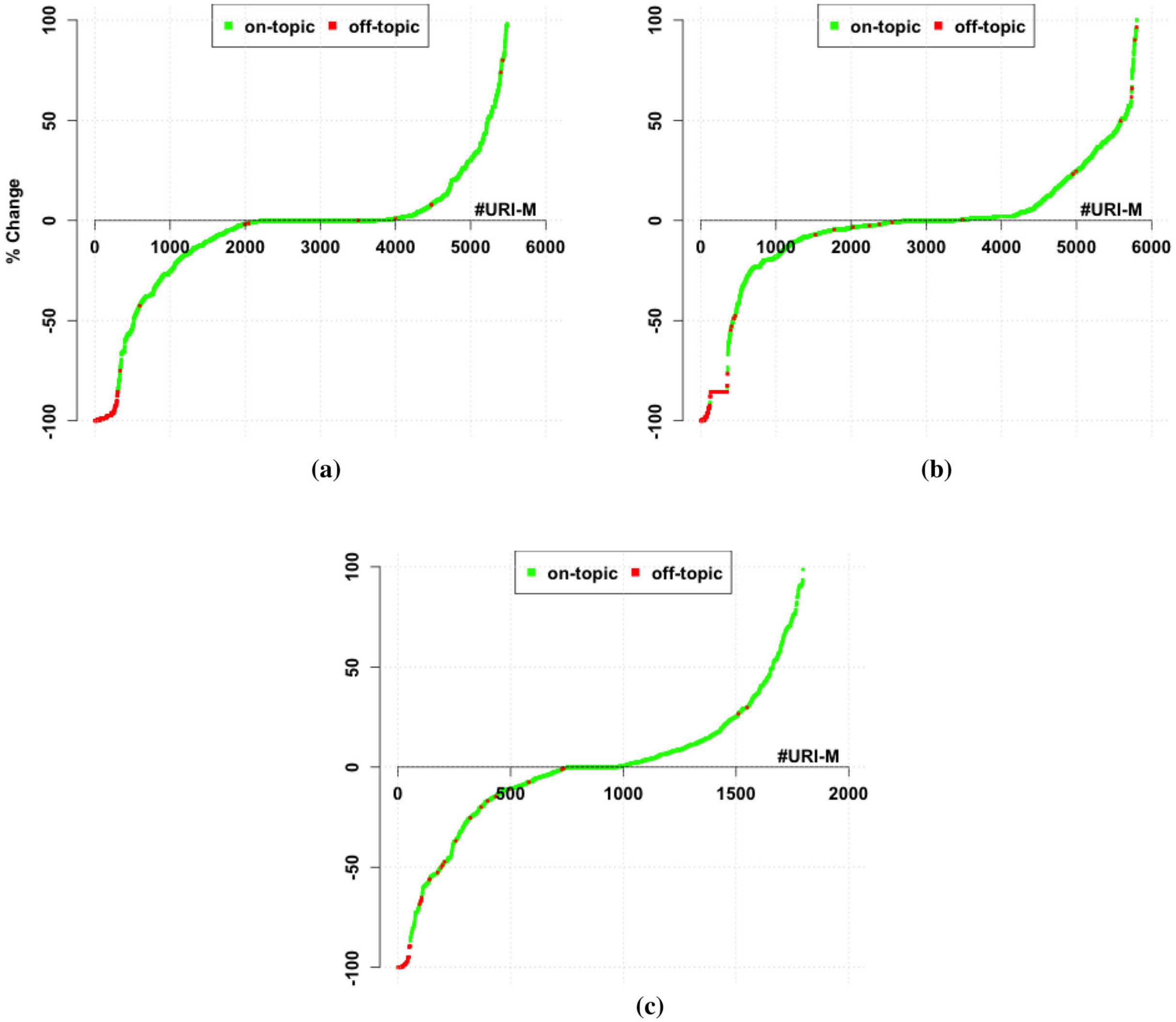
(c)

Fig. 8 How cosine similarity separates the off-topic from the on-topic pages **a** Occupy Movement Collection. **b** Egypt Revolution Collection. **c** Human Rights Collection

7.3 Combining the similarity measures

We tested 6615 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 5. 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 $th = (0.10, -0.85)$. 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 $th = (0.10, 0.00)$ 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 $th = (0.10, -0.85)$ 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.



(a)

(b)

(c)

Fig. 9 How change of page size (based on word count) separates the off-topic from the on-topic pages. **a** Occupy Movement Collection. **b** Egypt Revolution Collection. **c** Human Rights Collection

Table 5 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

Bold values indicate the best performing method based on the highest F1

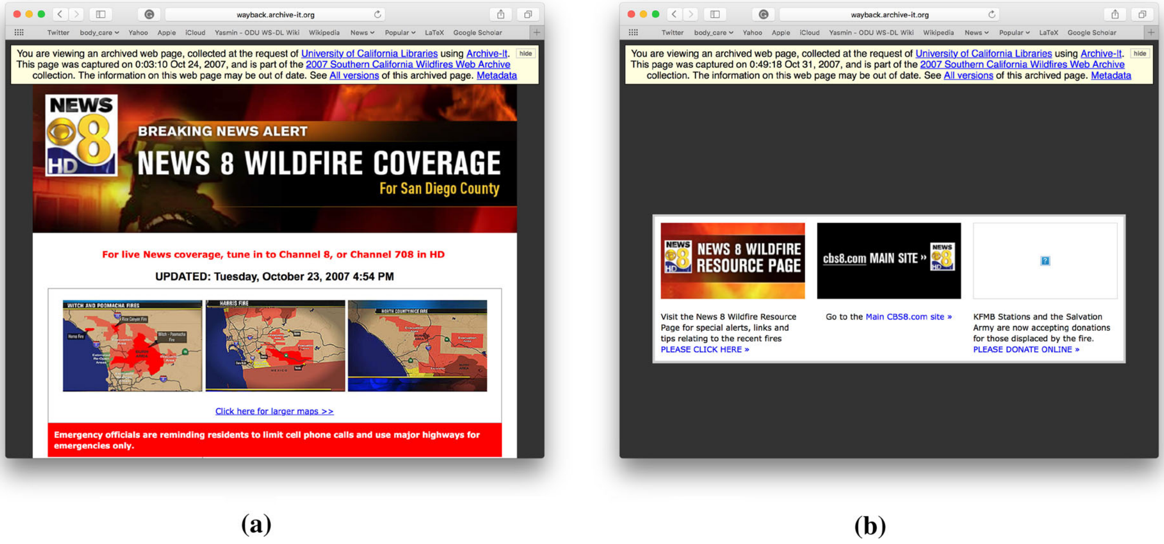
8 Evaluating Archive-It collections

We used the best-performing method (Cosine, WordCount) on the labeled data set 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 6 contains the details of the 18 tested collections, such as the collection's ID and time span that comprise 4019 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,[8] the 2007 Southern California Wildfires Web

[8] https://archive-it.org/collections/4887/.

Table 6 The results of evaluating Archive-It collections through the assessment of the detected off-topic pages using (Cosine, WordCount) methods at $th = (0.10, -0.85)$

Collection	ID	Timespan	URI-Rs	URI-Ms	Off-topic URI-Ms	Affected URI-Rs	TP	FP	P
Global food crisis	2893	2011/10/19–2012/10/24	65	3063	22	7	22	0	1.00
Government in Alaska	1084	2006/12/01–2013/04/13	68	506	16	4	16	0	1.00
Virginia tech shootings	2966	2011/12/08–2012/01/03	239	1670	24	2	24	0	1.00
Wikileaks 2010 document release collection	2017	2010/07/27–2012/08/27	35	2360	114	8	107	0	1.00
DIBAM	1019	2008/02/22–2008/03/24	25	106	4	1	4	0	1.00
Global health events	4887	2014/10/01–2015/10/21	165	3518	56	8	53	3	0.95
2003 California recall election	5947	2003/09/24–2003/12/12	178	2312	270	36	254	16	0.94
Jasmine revolution - tunisia 2011	2323	2011/01/19–2012/12/24	231	4076	107	31	107	7	0.94
Academics at baylor	3497	2013/01/28–2016/04/26	232	414	26	13	20	6	0.77
IT historical resource sites	1827	2010/02/23–2012/10/04	1459	10, 283	59	34	45	14	0.76
Human rights documentation initiative	1475	2009/04/29–2011/10/31	147	1530	54	20	39	15	0.72
2007 southern california wildfires Web archive	5810	2007/10/23–2007/11/02	156	2416	335	68	215	120	0.64
Maryland state document collection	1826	2010/03/04–2012/12/03	69	184	0	0	−−	−−	–
April 16 Archive	694	2007/05/23–2008/04/28	35	118	0	0	−−	−−	–
Brazilian school shooting	2535	2011/04/09–2011/04/14	476	1092	0	0	−−	−−	–
Russia plane crash sept 7, 2011	2823	2011/09/08–2011/09/15	65	447	0	0	−−	−−	–
Burke library New York city religions 340	1945	2011/11/16–2013/02/11	107	208	0	0	−−	−−	–
Hurricane irene (Aug 2011)	2816	2011/09/02–2011/09/26	71	102	0	0	−−	−−	–



(a)

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.

(b)

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 »

Fig. 10 Example of a significant change in cbs8.com: from October 24, 2007 to October 31, 2007 **a** cbs8.com on October 24, 2007. **b** cbs8.com on October 31, 2007

Archive,[9] the Academics at Baylor,[10] and the 2003 California Recall Election,[11] which we recently obtained from Archive-It.

The results of evaluating (Cosine, WordCount) at $th = (0.10, -0.85)$ are shown in Table 6. 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 precision $\mathrm{P} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$ 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 five 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 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 Figs. 10 and 11.

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.

In summary, (Cosine, WordCount) at $th = (0.10, -0.85)$ performed well on Archive-It collections with average $P = 0.9$.

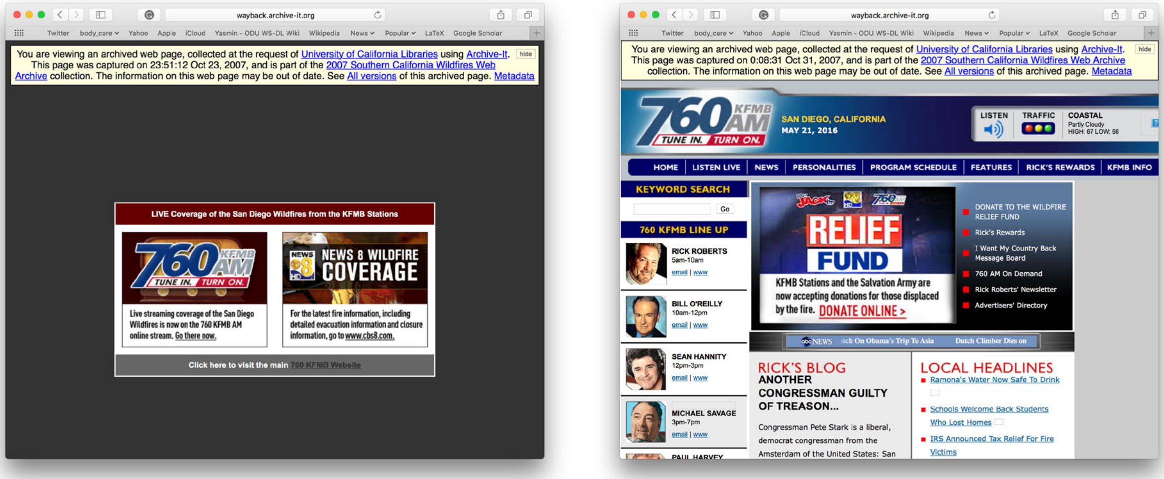
9 Conclusions and future work

In this paper, we presented approaches for assisting curators in identifying off-topic mementos of individual TimeMaps in the archive. We presented six methods for measuring the similarity between pages: cosine similarity, Jaccard similarity, intersection of the most 20 frequent terms, Web-based kernel function, change in the 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 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 also found that adding semantics to text using SEKernel enhanced the performance over Jaccard. We combined the suggested methods and found that, based on the F_1 score and the AUC, (Cosine, WordCount) at $th = (0.10, -0.85)$ enhances the performance to

[9] https://archive-it.org/collections/5810/.

[10] https://archive-it.org/collections/3497/.

[11] 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 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)

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

Partly Cloudy
HIGH: 67 LOW: 56

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)

Fig. 11 Example of a significant change in 760kfmb.com: from October 23, 2007 to October 31, 2007 **a** 760kfmb.com on October 23, 2007. **b** 760kfmb.com on October 31, 2007

have the highest F_1 score at 0.906 and the highest AUC at 0.968. We tested the selected thresholds and methods on different Archive-It collections. We tested the performance of (Cosine, WordCount) at $th = (0.10, -0.85)$ by applying them on 18 Archive-It collections, which comprise 4019 URI-Rs in which they have 36,785 URI-Ms. We manually assessed the relevancy of the detected off-topic pages. In summary, the suggested approach, (Cosine, WordCount) at $th = (0.10, -0.85)$, showed good results at detecting the off-topic pages with 0.89 precision. The presented approaches will help curators to judge their crawls and also 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 many applications such as story summarization and thumbnail generation.

We 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.

This is a preliminary investigation of automatically detecting the off-topic pages from Web archives. There is more research to be done in exploring the thresholds and methods. For example, the nature of collection, such as the time span, might affect choosing the threshold. Users will be able to adjust the methods and thresholds as command-line parameters. Furthermore, there are other methods, such as probabilistic topic models [10] that can be applied to the collection that would then allow for comparison of the aboutness of the mementos to the aboutness of the collection. We generated a gold standard data set 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.

Our future work will continue to improve detection by using larger data sets and more collections with different features. The methods presented here 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 a model of the topic of the collection, in part to more easily detect the off-topic pages in the "Always Off" and "Step Function Off" TimeMaps.

Acknowledgments This work was supported in part by the AMF and the IMLS LG-71-15-0077-15. We thank Kristine Hanna from the Internet Archive for help in obtaining the data set. We also thank the anonymous reviewers for their insights regarding future directions to this work.

References

1. AlNoamany, Y.: Using Web Archives to Enrich the Live Web Experience Through Storytelling. Dissertation, Old Dominion University (2016)
2. 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 (2015). doi:10.1007/978-3-319-24592-8_20
3. 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
4. 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
5. AlSum, A., Nelson, M.L.: ArcLink: Optimization Techniques to Build and Retrieve the Temporal Web Graph. Tech. Rep. (2013). arXiv:1305.5959
6. AlSum, A., Nelson, M.L.: Thumbnail Summarization Techniques for Web Archives. In: Proceedings of the 36th European Conference on Information Retrieval, ECIR 2014, pp. 299–310 (2014). doi:10.1007/978-3-319-06028-6_25
7. 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
8. 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. ACM Press (2004). doi:10.1145/988672.988716
9. 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)
10. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
11. Brewington, B., Cybenko, G.: Keeping up with the changing web. Computer **33**(5), 52–58 (2000). doi:10.1109/2.841784
12. 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)
13. 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
14. Chakrabarti, S., Van den Berg, M., Dom, B.: Focused crawling: a new approach to topic-specific web resource discovery. Comput. Netw. **31**(11), 1623–1640 (1999). doi:10.1016/S1389-1286(99)00052-3
15. Cho, J., Garcia-Molina, H.: Estimating frequency of change. ACM Trans. Internet Technol. **3**(3), 256–290 (2003). doi:10.1145/857166.857170
16. Cho, J., Garcia-Molina, H., Page, L.: Efficient crawling through URL ordering. Comput. Netw. ISDN Syst. **30**(1–7), 161–172 (1998). doi:10.1016/S0169-7552(98)00108-1
17. Farag, M.M.G., Fox, E.A.: Intelligent Event Focused Crawling. In: Proceedings of the 11th International ISCRAM Conference, pp. 18–21 (2014)
18. Fawcett, T.: An introduction to ROC analysis. Pattern Recognit. Lett. **27**(8), 861–874 (2006). doi:10.1016/j.patrec.2005.10.010
19. Foot, K., Schneider, S.: Web Campaigning (Acting with Technology). The MIT Press, Cambridge (2006)
20. ISO 28500:2009—Information and documentation–WARC file format. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717 (2009)
21. 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)
22. Jatowt, A., Kawai, Y., Tanaka, K.: Page history explorer: visualizing and comparing page histories. IEICE Trans. Inf. Syst. **94**(3), 564–577 (2011)
23. Jatowt, A., Tanaka, K.: Towards mining past content of Web pages. New Rev. Hypermed. Multimed. **13**(1), 77–86 (2007). doi:10.1080/13614560701478897
24. Kahle, B.: Preserving the internet. Sci. Am. **276**(3), 82–83 (1997)
25. Kahle, B.: Wayback Machine Hits 400,000,000,000! http://blog.archive.org/2014/05/09/wayback-machine-hits-400000000000 (2014)
26. Klein, M., Nelson, M.L.: 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, vol. 6966, pp. 27–39. Springer, Berlin Heidelberg (2011). doi:10.1007/978-3-642-24469-8_5
27. Klein, M., Shipman, J., Nelson, M.L.: Is this a good title? In: Proceedings of the 21st ACM conference on Hypertext and Hypermedia, HT '10, pp. 3–12. ACM (2010). doi:10.1145/1810617.1810621
28. 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
29. 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
30. Koehler, W.: Web page change and persistence—a four-year longitudinal study. J. Am. Soc. Inf. Sci. Technol. **53**(2), 162–171 (2002)
31. Koehler, W.: A longitudinal study of web pages continued: a consideration of document persistence. Inf. Res. **9**(2), 2–9 (2004)
32. 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
33. Kosala, R., Blockeel, H.: Web mining research: a survey. SIGKDD Explor. Newslett. **2**(1), 1–15 (2000). doi:10.1145/360402.360406
34. 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
35. Manning, C.D., Raghavan, P., Schütze, H., Schutze, H.: Introduction to information retrieval. Cambridge University Press (2008). doi:10.1017/CBO9780511809071
36. 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
37. Marshall, C., McCown, F., Nelson, M.: Evaluating Personal Archiving Strategies for Internet-based Information. In: Proceedings of Archiving 2007, vol. 2007, pp. 151–156 (2007)
38. Masanès, J.: Web Archiving. Springer, Cham (2006)

39. 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. http://iwaw.europarchive.org/04/Mohr.pdf (2004)
40. 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)
41. Nelson, M.L.: A Plan For Curating "Obsolete Data or Resources". Tech. Rep. (2012). arXiv:1209.2664
42. Odijk, D., Grbacea, C., Schoegje, T., Hollink, L., de Boer, V., Ribbens, K., van Ossenbruggen, J.: Supporting exploration of historical perspectives across collections. In: Proceedings of the 19th International Conference on Theory and Practice of Digital Libraries. Lecture Notes in Computer Science, vol. 9316, pp. 238–251. Springer-Verlag (2015). doi:10.1007/978-3-319-24592-8_18
43. Olston, C., Pandey, S.: Recrawl scheduling based on information longevity. In: Proceeding of the 17th International World Wide Web Conference, WWW '08, p. 437. ACM Press (2008). doi:10.1145/1367497.1367557
44. 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. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
45. 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)
46. 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). doi:10.1145/1998076.1998098
47. 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
48. 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)
49. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**(11), 613–620 (1975). doi:10.1145/361219.361220
50. 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)
51. Singhal, A.: Modern information retrieval: a brief overview. Bull. IEEE Comput. Soc. Tech. Comm. Data Eng. **24**(4), 35–42 (2001)
52. 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
53. Teevan, J., Dumais, S.T., Liebling, D.J.: A longitudinal study of how highlighting web content change affects people's web interactions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10, pp. 1353–1356. ACM (2010). doi:10.1145/1753326.1753530
54. 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
55. 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)
56. Yin, Z., Shokouhi, M., Craswell, N.: Query expansion using external evidence. In: Advances in Information Retrieval, pp. 362–374. Springer (2009)