Final Report for Tools for Managing Seed URIs

Michael L. Nelson, Michele C. Weigle
Old Dominion University, Norfolk VA 23529 USA
{mln,mweigle}@cs.odu.edu

Mon Sep 7 16:28:35 EDT 2015

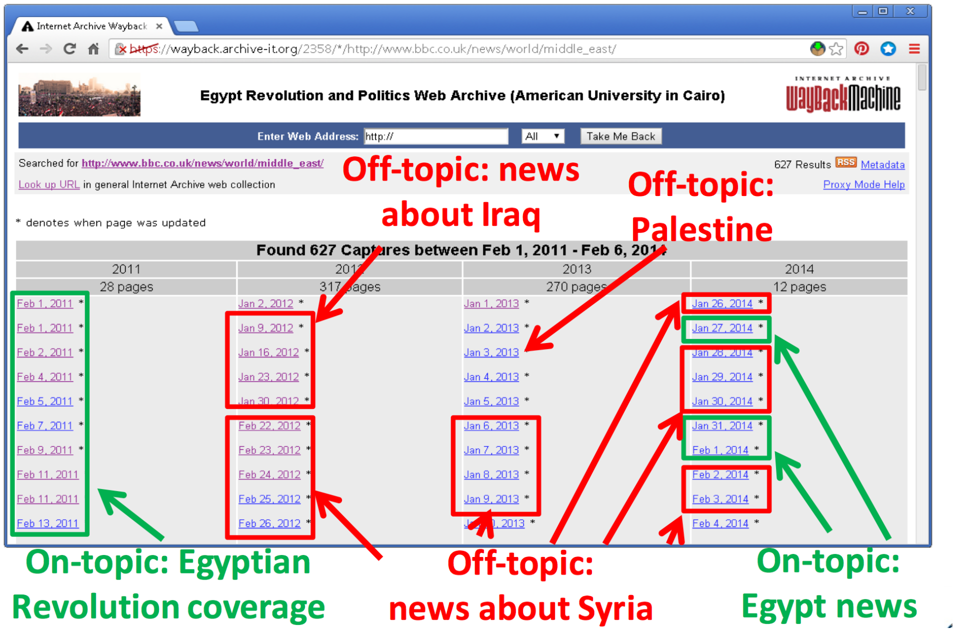
1 Overview

Perhaps the largest and most influential actor in web preservation is the Internet Archive (IA). Archive-It (`http://Archive-It.org`) is a collection development service deployed by the Internet Archive in 2006. Archive-It is currently used by over 180 institutions in 44 states, and features over 4B archived web pages in nearly 2000 separate collections. Archive-It partners receive an account at Archive-It and build themed collections of archived web pages hosted Archive-It's machines. This is done by the user 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). The Heritrix crawler at Archive-It then recrawls these seeds at the specified frequency and depth to build a collection of archived web pages that the curator believes best exemplifies the topic of the collection. Archive-It then provides faceted browsing and search services on the resulting collection.

Heritrix is a powerful web crawling utility, and Archive-It has deployed tools that allow collection curators to perform quality control on their crawls. However, the tools are currently focused on issues such as the mechanics of HTTP (e.g., how many HTML files vs. PDFs, how many 404 missing URIs) and domain information (e.g., how many .uk sites vs. .com sites). There are currently no content-based tools that allow curators to judge the quality of their crawl and detect when seed URIs (and other crawled pages) are off topic (low precision) or discover candidate seed URIs that are not currently included (low recall).

Some sites are initially on-topic, but then go off-topic. A frequent reason is that the domain expired and the previous owner forgot to re-register the domain. The lapsed domains are then purchased by spammers who desire all the incoming traffic that the site accrued while it was "legitimate". In contrast, some sites captured in crawls were never on-topic and were included by accident, often hastily included when a rapidly developing event is being archived. However, it is important to remember that on- and off-topic is defined relative to the collection itself. Figure 1 shows a `bbc.co.uk` page about the latest news from the Middle East. The stories are always about developments in the Middle East, but not all of the pages are about the Egyptian Revolution (Archive-It collection 2358).

The purpose of this work is to develop a command line tool that operates on the WARC files (i.e., the files directly created by Heritrix crawler) or optionally the public web interface of Archive-It for detecting when seed URIs have gone off topic. In this project we: 1) created a gold standard data set of on- and off-topic URIs for three Archive-It collections, 2) published in a refereed conference proceedings an evaluation of various methods of detecting off-topic pages, and 3) wrote an open source command line tool that implements the best methods discovered in the research documented in the paper.



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)

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

Searched for http://www.bbc.co.uk/news/world/middle_east/

Look up URL in general Internet Archive web collection

627 Results RSS Metadata

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 Iraq
Off-topic: Palestine
On-topic: Egyptian Revolution coverage
Off-topic: news about Syria
On-topic: Egypt news

Figure 1: Always About the Middle East, but Not Always About the Egyptian Revoluion

2 Work Accomplished

The work reported here closely aligns with the Ph.D. research of Yasmin AlNoamany, and as such represents a section of continuing work. Although the official work period covered September 2014 – March 2015, Yasmin's research continues with a likely Ph.D. defense in 2016.

2.1 Task 1: Establish Baseline of Current Collections:

For this task we established a gold standard data set drawn from three collections:

- Human Rights `https://archive-it.org/collections/1068`
- Occupy Movement 2011/2012 `https://archive-it.org/collections/2950`
- Egypt Revolution and Politics `https://archive-it.org/collections/2358`

The URLs above are to the public versions of the collections even though our analysis was down on the ODU dark archive of Archive-It's collection.

We looked at both seed URIs ("original URIs" or URI-Rs in Memento parlance) and mementos (URI-Ms) and manually determined which ones were on or off topic. We calculated:

- Occupy Movement 2011/2012: on-topic: 188 of 255 URI-Rs and 6112 of 6570 URI-Ms
- Human Rights: on-topic: 172 of 206 URI-Rs and 2213 of 2346 URI-Ms
- Egypt Revolution and Politics: on-topic: 106 of 143 URI-Rs and 6417 of 7082 URI-Ms

Always On:

```
on  on  on  on  on  on  on  on
t0  t1  t2  t3  t4  t5  t6  t7
```

Step Function On:

```
on  on  on  on  off off off off
t0  t1  t2  t3  t4  t5  t6  t7
```

Step Function Off:

```
off off off on  on  on  on  on
t0  t1  t2  t3  t4  t5  t6  t7
```

Oscillating:

```
on  on  off off on  off off on
t0  t1  t2  t3  t4  t5  t6  t7
```

Always Off:

```
off off off off off off off off
t0  t1  t2  t3  t4  t5  t6  t7
```

Figure 2: Abstract TimeMap Types

With these three collections forming our gold standard data set, we can assess the performance of various methods for detecting when URIs go off-topic. The gold standard data set is available on `github.com` in the following format:

```
id date URI label
...
19 20110602192838 http://wayback.archive-it.org/1068/20110602192838/http://amnestymauritius.org/ 1
19 20110902210710 http://wayback.archive-it.org/1068/20110902210710/http://amnestymauritius.org/ 1
19 20111202210619 http://wayback.archive-it.org/1068/20111202210619/http://amnestymauritius.org/ 0
19 20120303005657 http://wayback.archive-it.org/1068/20120303005657/http://amnestymauritius.org/ 0
...
```

Where “id” is the consecutively numbered seed URI (URI-R), and “1” and “0” represent a manual assessment of on- and off-topic, respectively.

2.2 Task 2: General Trends in Archive-It Collections:

We found more variety in how things went off topic than we originally expected. We defined the following classifications, with abstract examples of on-topic and off-topic pages.

We had expected to find Always On, Step Function On, and a few Always Off (included by accident). We were surprised to discover the frequency of Oscillating. The Always Off is hard to discover (since our assumption is that the first memento is always on topic). For the purpose of our analysis and the resulting tools, we assumed that Always Off does not occur (it does occur, but the frequency is nearly 0%).

Whereas figure 2 illustrates abstract TimeMaps, figure 3 shows actual values measured for seed URIs from the “Occupy Movement” (2950) collection:

1. `wayback.archive-it.org/2950/*/http://occupypsl.org`
2. `wayback.archive-it.org/2950/*/http://occupygso.tumblr.com`

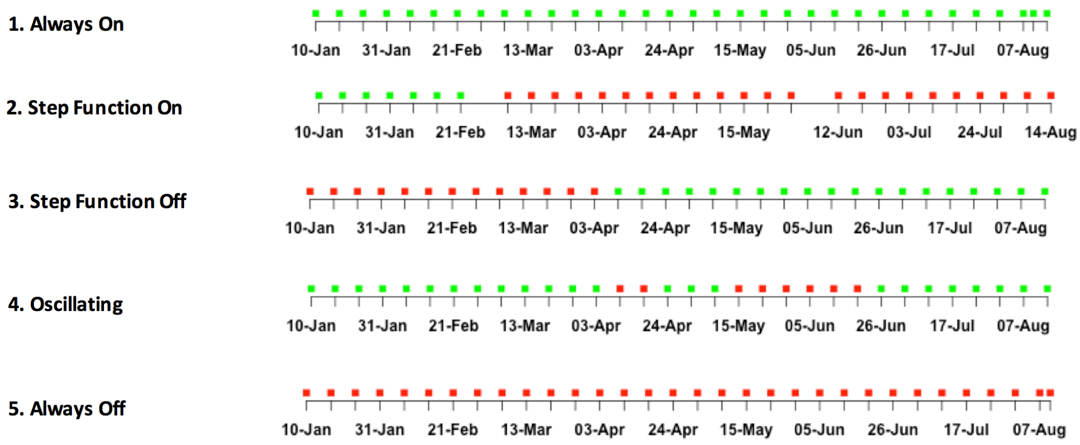


Figure 3: TimeMap Types for the "Occupy Movement" (2950) Collection (Green=On-Topic, Red=Off-Topic).

2.3 Task 3: Tools for Pruning the Collection:

We investigated many different methods (and combinations of methods) for detecting off-topic mementos, including:

- Intersection of the Top 20 terms
- Jaccard similarity
- Cosine similarity
- Percentage change in unique terms
- Percentage change in bytes
- Search engine kernel

Detailed in our TPDL 2015 paper (see section 5), we ran 6,615 tests on our gold standard data set (15 method combinations × 21 threshold values × 21 threshold values). We achieved our highest F1 score (0.906) with cosine similarity $\leq$ 0.10 OR change in bytes = -0.85 (i.e., an 85% shrinkage in size).

We developed a tool that can detect off-topic pages for entire Archive-It collections or just individual TimeMaps. See the `github.com` page for full details, but the command line tool in operation looks like:

[3] `wayback.archive-it.org/2950/*/http://occupyashland.com`

[4] `wayback.archive-it.org/2950/*/http://www.indyows.org`

[5] `wayback.archive-it.org/2950/*/http://occupy605.com`

```
% echo "collection 1826, with cosine threshold of 0.15" > /dev/null
% python detect_off_topic.py -i 1826 -th 0.15
extracting seed list
...
http://agroecol.umd.edu/Research/index.cfm
http://casademaryland.org
...
50 URIs are extracted from collection https://archive-it.org/collections/1826
Downloading timemap using uri
  http://wayback.archive-it.org/1826/timemap/link/http://agroecol.umd.edu/Research/index.cfm
Downloading timemap using uri
  http://wayback.archive-it.org/1826/timemap/link/http://casademaryland.org
...
Downloading 4 mementos out of 306
Downloading 14 mementos out of 306
...
Detecting off-topic mementos using Cosine Similarity method

Similarity memento_uri
0.0 http://wayback.archive-it.org/1826/20131220205908/http://www.mncppc.org/commission_home.html/
0.0 http://wayback.archive-it.org/1826/20141118195815/http://www.mncppc.org/commission_home.html


% echo "a single TimeMap with a byte count threshold of -0.85" > /dev/null
% python detect_off_topic.py -t
  https://wayback.archive-it.org/2358/timemap/link/http://hamdeensabahy.com/  -m wcount -th -0.85

Downloading 0 mementos out of 270
http://wayback.archive-it.org/2358/20140524131241/http://www.hamdeensabahy.com/
http://wayback.archive-it.org/2358/20130321080254/http://hamdeensabahy.com/
http://wayback.archive-it.org/2358/20130621131337/http://www.hamdeensabahy.com/
http://wayback.archive-it.org/2358/20140602131307/http://hamdeensabahy.com/
http://wayback.archive-it.org/2358/20140528131258/http://www.hamdeensabahy.com/
http://wayback.archive-it.org/2358/20130617131324/http://www.hamdeensabahy.com/

...
Downloading 4 mementos out of 270
...
Extracting text from the html
...
Detecting off-topic mementos using Word Count method

Similarity memento_uri
-0.979434447301 http://wayback.archive-it.org/2358/20121213102904/http://hamdeensabahy.com/

-0.966580976864 http://wayback.archive-it.org/2358/20130321080254/http://hamdeensabahy.com/

-0.94087403599 http://wayback.archive-it.org/2358/20130526131402/http://www.hamdeensabahy.com/

-0.94087403599 http://wayback.archive-it.org/2358/20130527143614/http://www.hamdeensabahy.com/
```

We also ran this tool on 11 different Archive-It collections (not in the gold standard data set) and found that while four collections did not have off-topic mementos (all but one of the four collections covered short duration events and thus did not have the opportunity to go off-topic), the other seven collections had many off-topic URI-Rs and mementos (i.e., the off-topic pages were not just the result of a single "bad" page). The precision for identifying these off-topic pages ranged from 1.00 to 0.72, with a mean precision of 0.92 (see the TPDL paper for details).

2.4 Task 4: Package and Release Tools:

The final phase of the project is to release the tools and gold standard data set to the web archiving community, via `github.com`. The URLs for the code, datasets, published evaluations, etc. are provided in sections 3 and 5.

3 Outcomes and Deliverables

All source code, examples, libraries, gold standard dataset for Archive-It collections 1068, 2358, and 2950, and other related deliverables are available at https://github.com/yasmina85/OffTopic-Detection. This repository is available for forking, branching, updates, etc. We welcome any additional feedback and/or contributions.

4 Obstacles

We did not encounter any obstacles or issues that affected our ability to complete the proposed work.

5 Presentations and Publications

- Michele Weigle briefly covered the work in progress in her presentation "Tools for Managing the Past Web" at the 2014 Archive-It Partners Meeting.

  http://ws-dl.blogspot.com/2014/11/2014-11-20-archive-it-partners-meeting.html

- Michele Weigle presented Yasmin AlNoamany's slides "Tools for Managing Seed URIs (Detecting Off-Topic Pages)" at the Web Archiving Collaboration conference at Columbia University, June 4 and 5, 2015.

  http://ws-dl.blogspot.com/2015/06/2015-06-09-web-archiving-collaboration.html

- Yasmin AlNoamany gave an invited presentation titled "Detecting Off-Topic Pages in Web Archives" at the Internet Archive, August 20, 2015.

  http://ws-dl.blogspot.com/2015/08/2015-08-20-odu-l3s-stanford-and.html

- Yasmin AlNoamany, Michele C. Weigle, and Michael L. Nelson, Detecting Off-Topic Pages in Web Archives, Proceedings of Theory and Practice of Digital Libraries (TPDL) 2015, September 2015.

  http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-off-topic.pdf