6-4-2015

Tools Managing Seed URIs (Detecting Off-Topic Pages)

Yasmin AlNoamany
Old Dominion University

Michele C. Weigle
Old Dominion University, mweigle@odu.edu

Michael L. Nelson
Old Dominion University, mnelson@odu.edu

OLD DOMINION UNIVERSITY

Tools for Managing Seed URIs
(Detecting Off-Topic Pages)

Yasmin AlNoamany, Michele C. Weigle, Michael L. Nelson

Old Dominion University

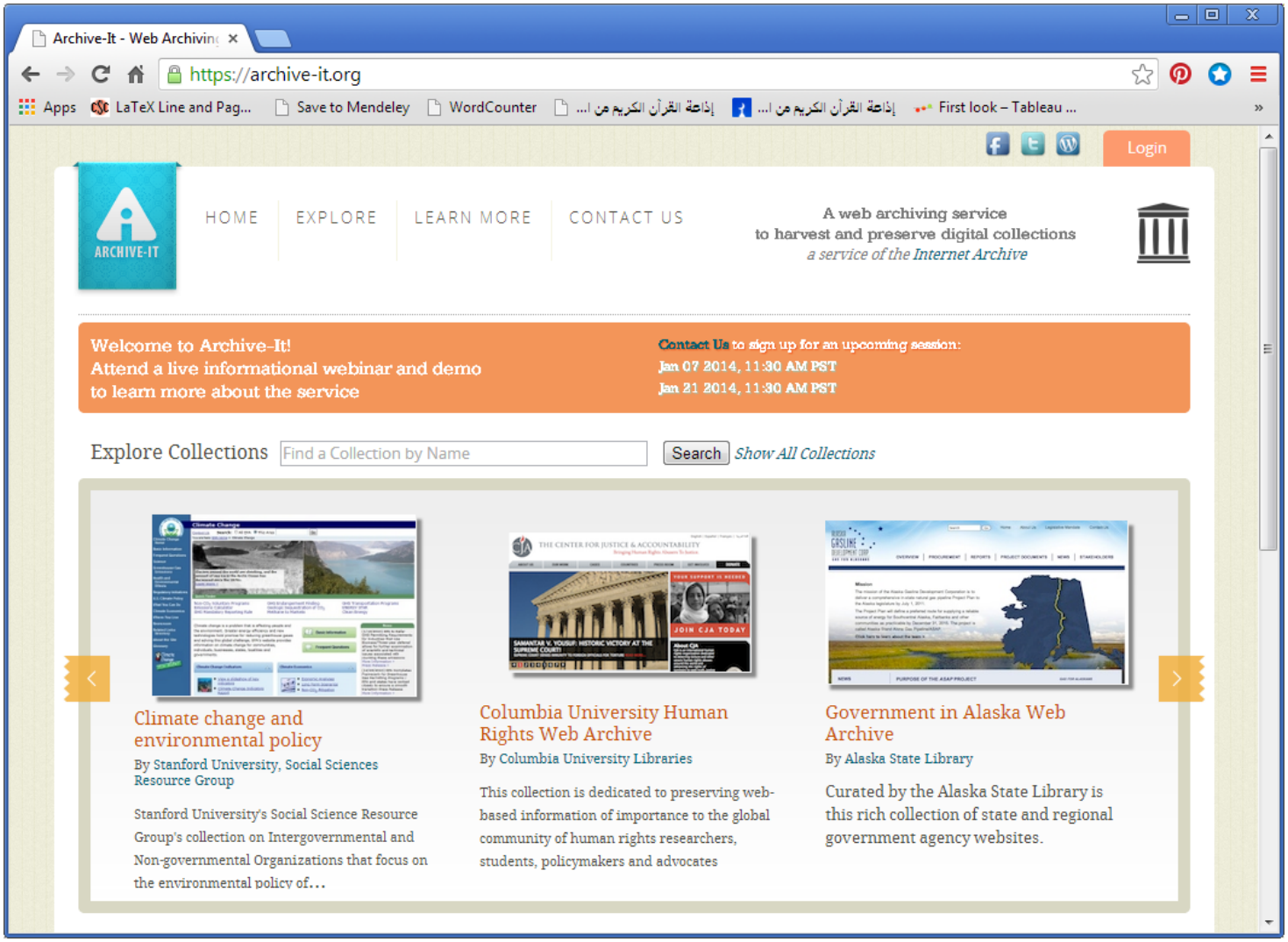
Web Science and Digital Libraries Group

http://ws-dl.cs.odu.edu/, @WebSciDL

Funded by Columbia University Libraries Web Archiving Incentive program

Web Archiving Collaboration: New Tools and Models
June 4-5, 2015

Archive-It hosts curated web collections

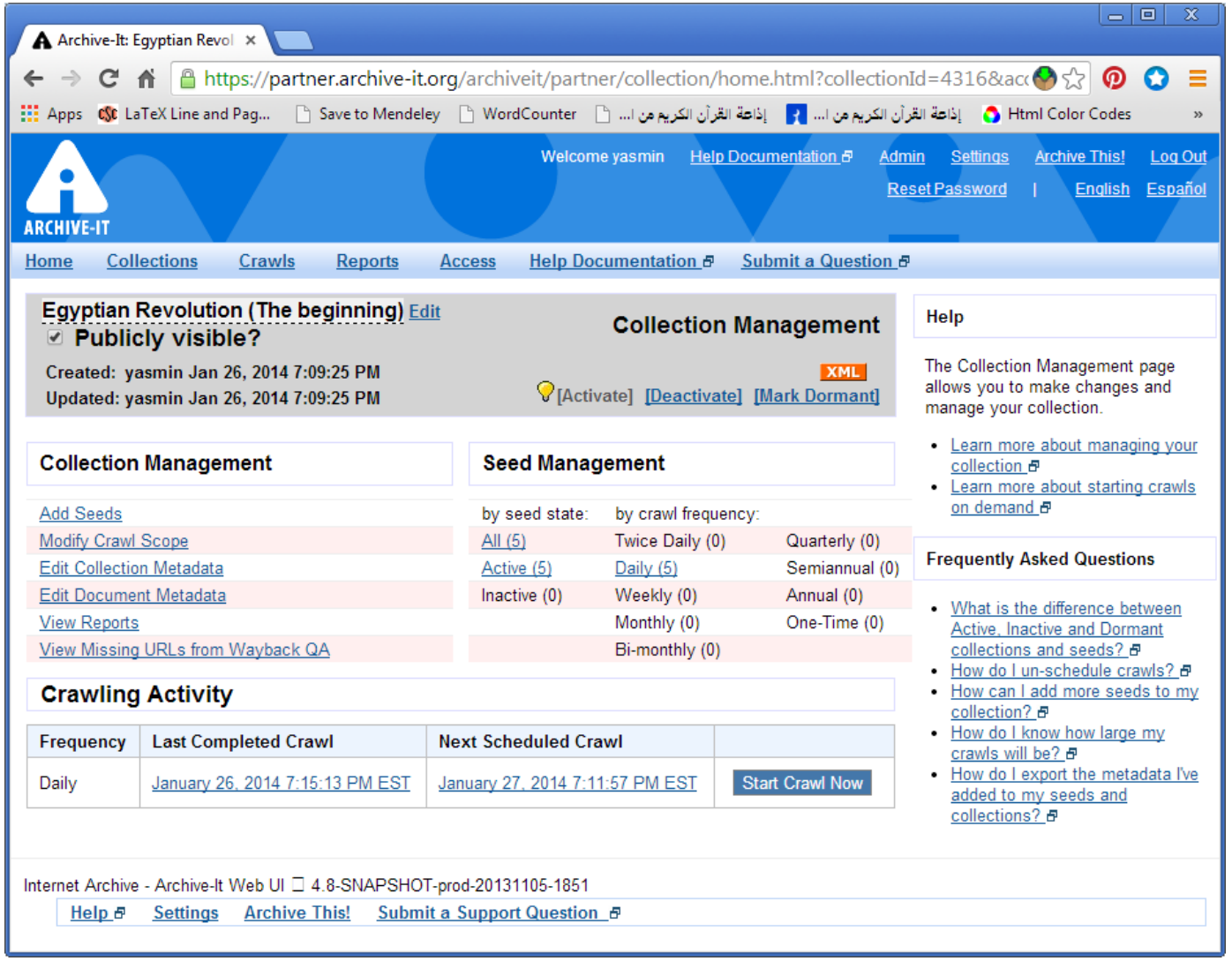


> 3,000 collections

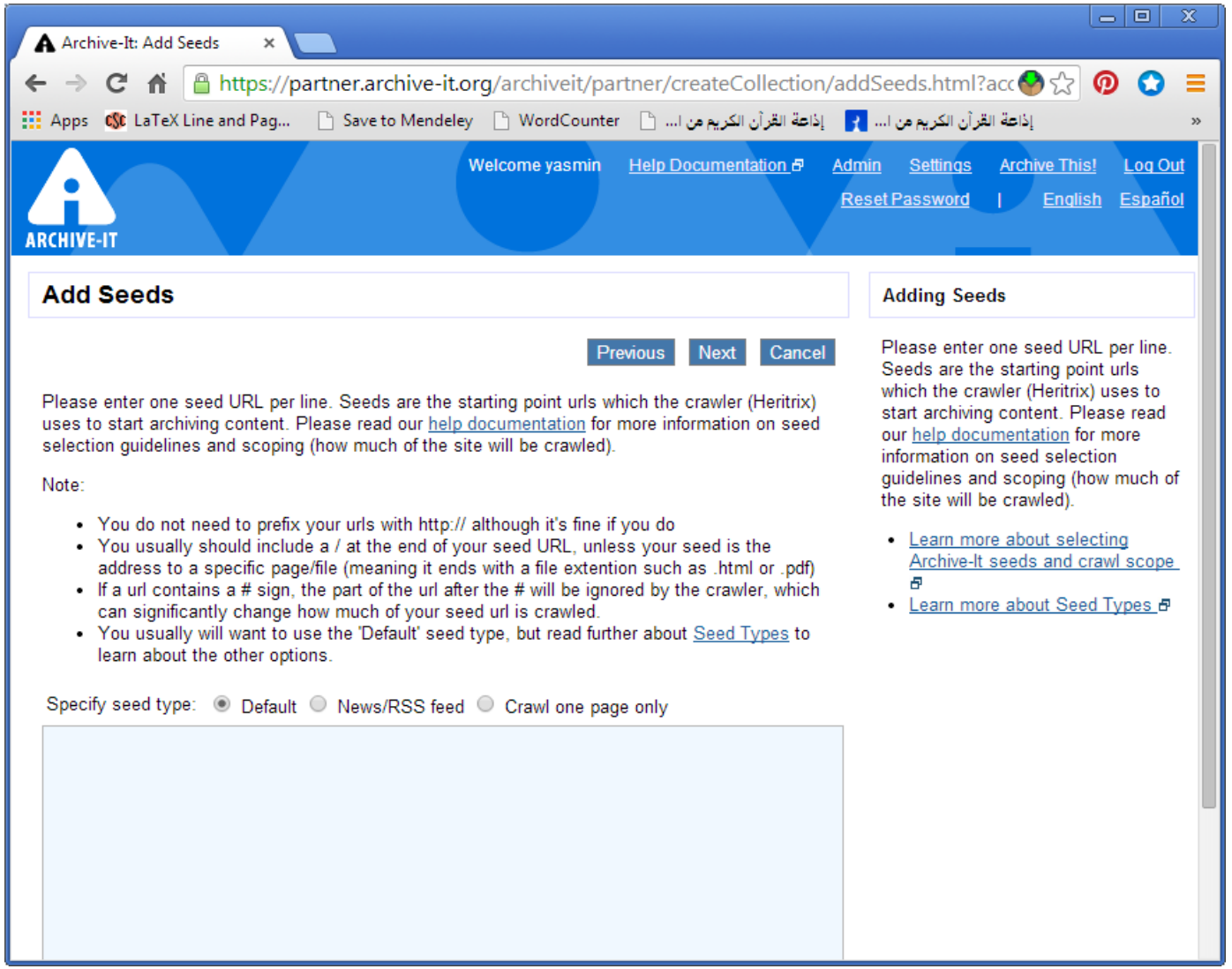
~340 institutions

> 10B archived pages

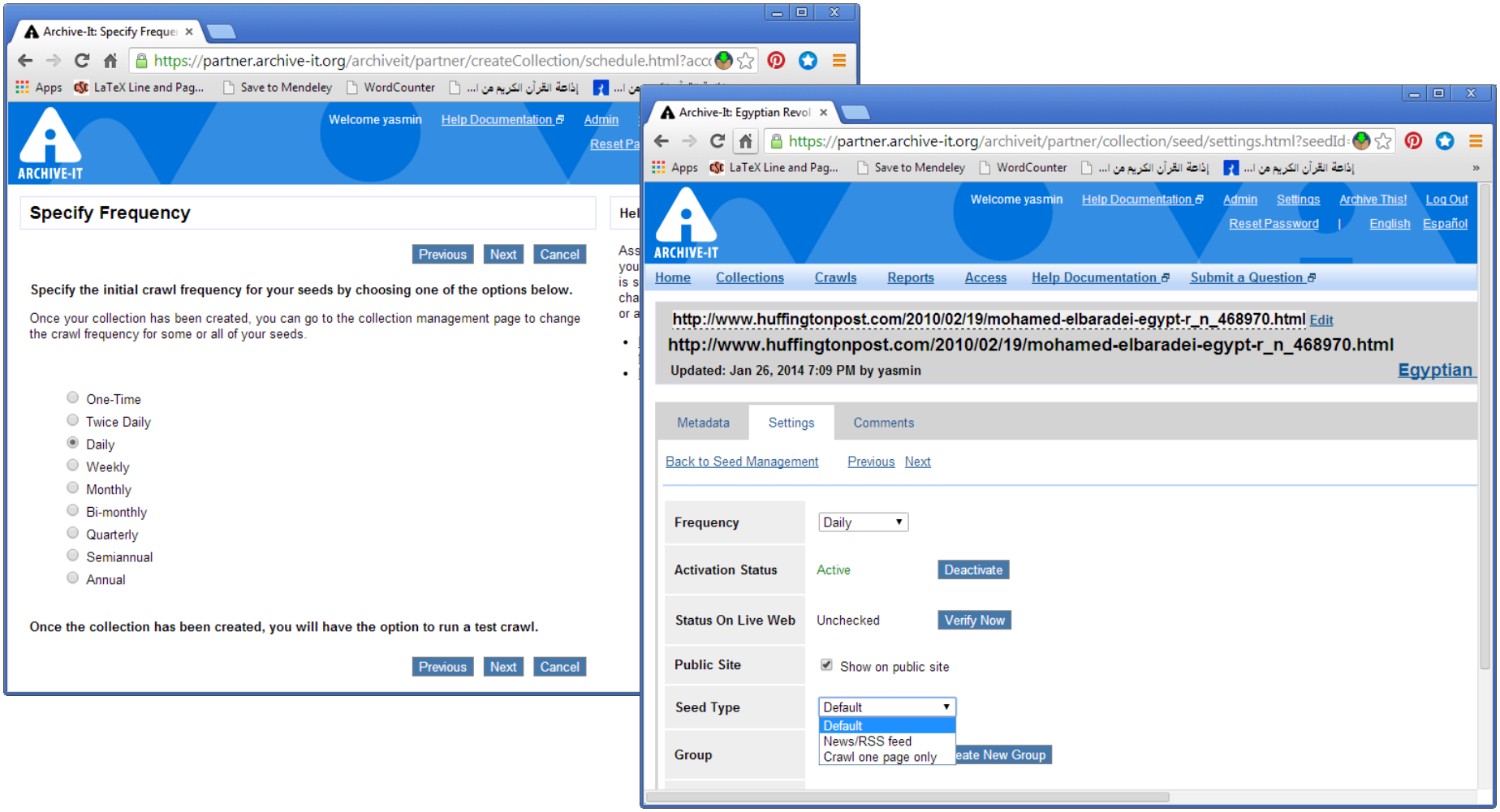
Curator's view of Archive-It



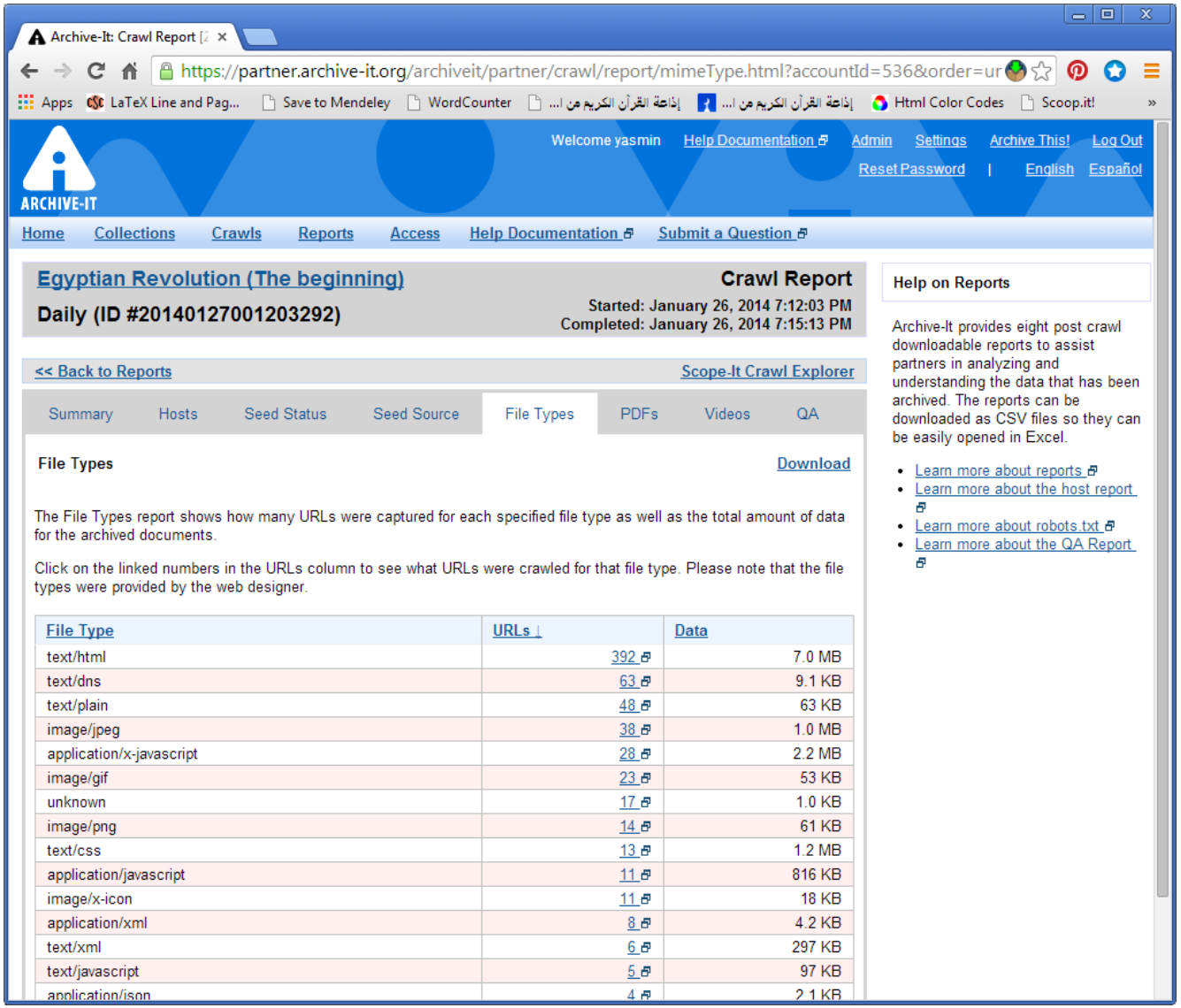
The collection curator specifies seed URIs



Curators specify the breadth and depth of the crawl



Current tools measure HTTP events, not "aboutness"



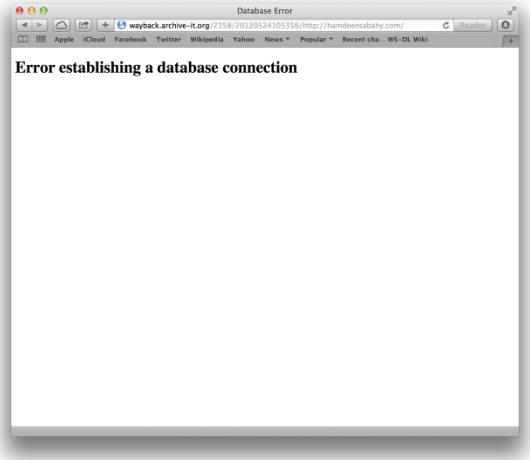
Pages can go off-topic through time

May 13, 2012: The page started as on-topic.

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

Pages can go off-topic through time

May 13, 2012: The page started as on-topic.

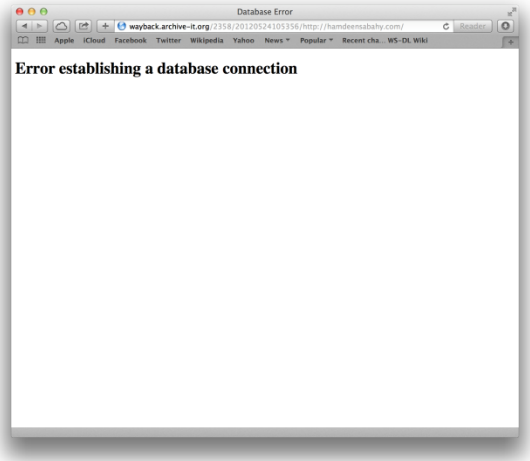


May 24, 2012: Off-topic due to a database error.

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

Pages can go off-topic through time

May 13, 2012: The page started as on-topic.



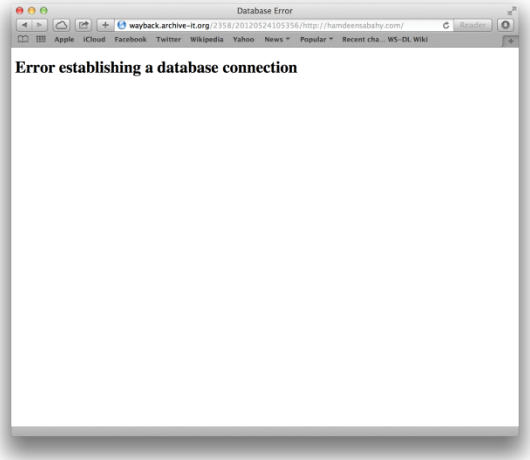
May 24, 2012: Off-topic due to a database error.

Mar. 21, 2013: Not working because of financial problems.

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

Pages can go off-topic through time

May 13, 2012: The page started as on-topic.



May 24, 2012: Off-topic due to a database error.

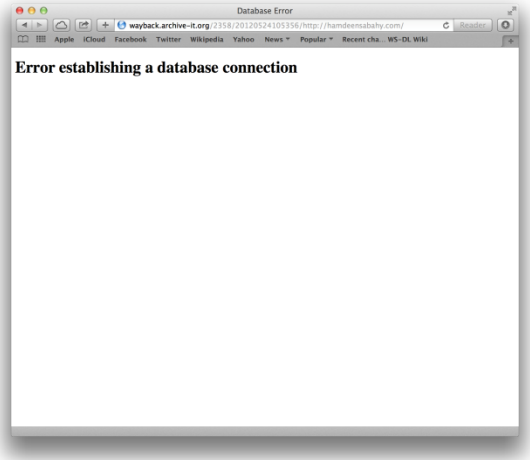
Mar. 21, 2013: Not working because of financial problems.

May 21, 2013: On-topic again

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

Pages can go off-topic through time

May 13, 2012: The page started as on-topic.



May 24, 2012: Off-topic due to a database error.

Mar. 21, 2013: Not working because of financial problems.

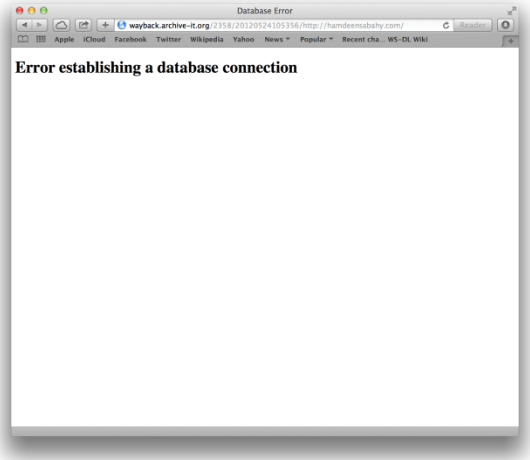
May 21, 2013: On-topic again

June 5, 2014: The site has been hacked

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

Over 60% of archived versions of hamdeensabahy.com are off-topic

May 13, 2012: The page started as on-topic.

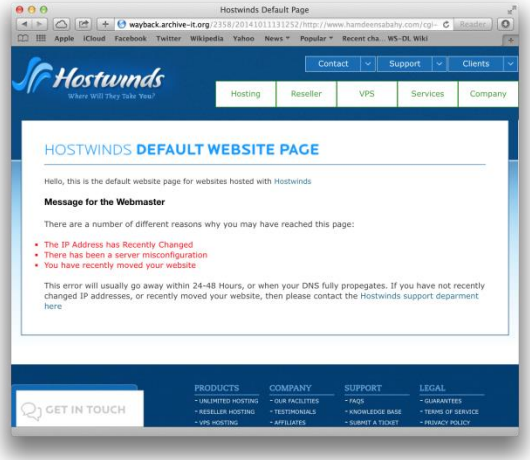


May 24, 2012: Off-topic due to a database error.

Mar. 21, 2013: Not working because of financial problems.

May 21, 2013: On-topic again

June 5, 2014: The site has been hacked



Oct. 10, 2014: The domain has expired.

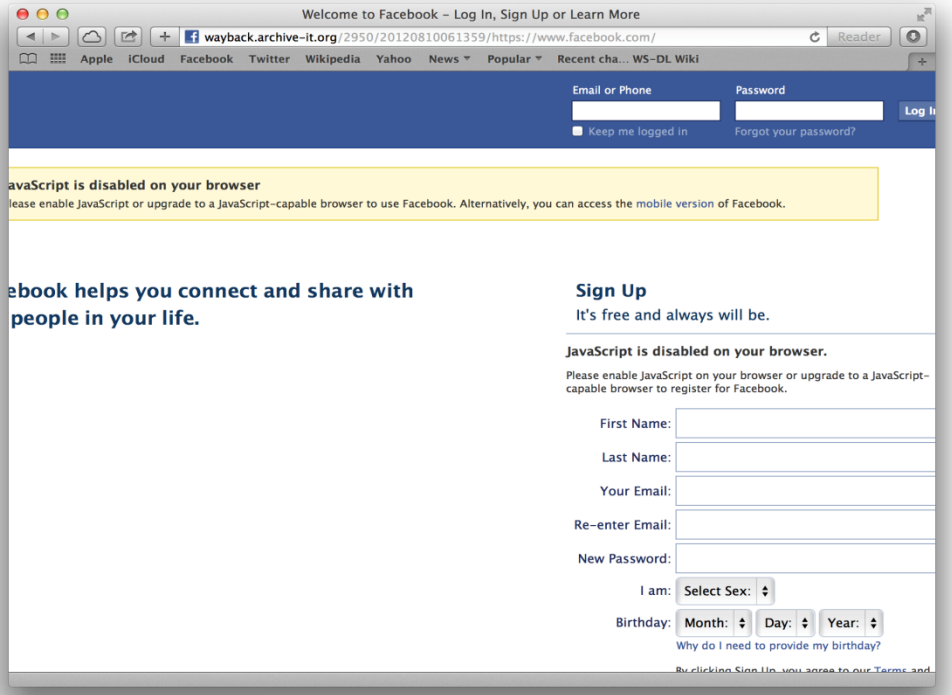
http://wayback.archive-it.org/2358/*/http://hamdeensabahy.com

Social media pages can go off-topic

Dec. 22, 2011: Facebook page was relevant to the Occupy collection

http://wayback.archive-it.org/2950/*/http://www.facebook.com/MayorJeanQuan

Social media pages can go off-topic



Dec. 22, 2011: Facebook page was relevant to the Occupy collection

Aug. 10, 2012: URI redirects to www.facebook.com

http://wayback.archive-it.org/2950/*/http://www.facebook.com/MayorJeanQuan

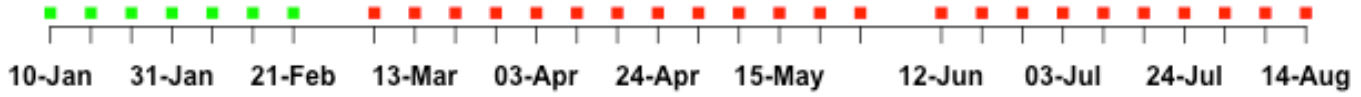
Classifying web page behavior over time

A TimeMap is the list of a URI-R's mementos

We identified 5 classes of TimeMaps

1. Always On

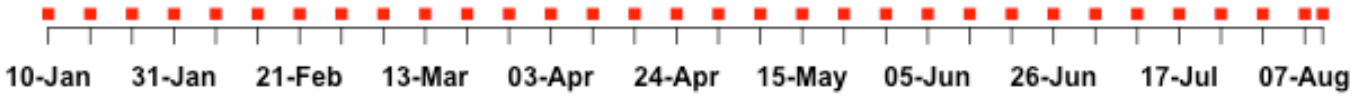
2. Step Function On



3. Step Function Off

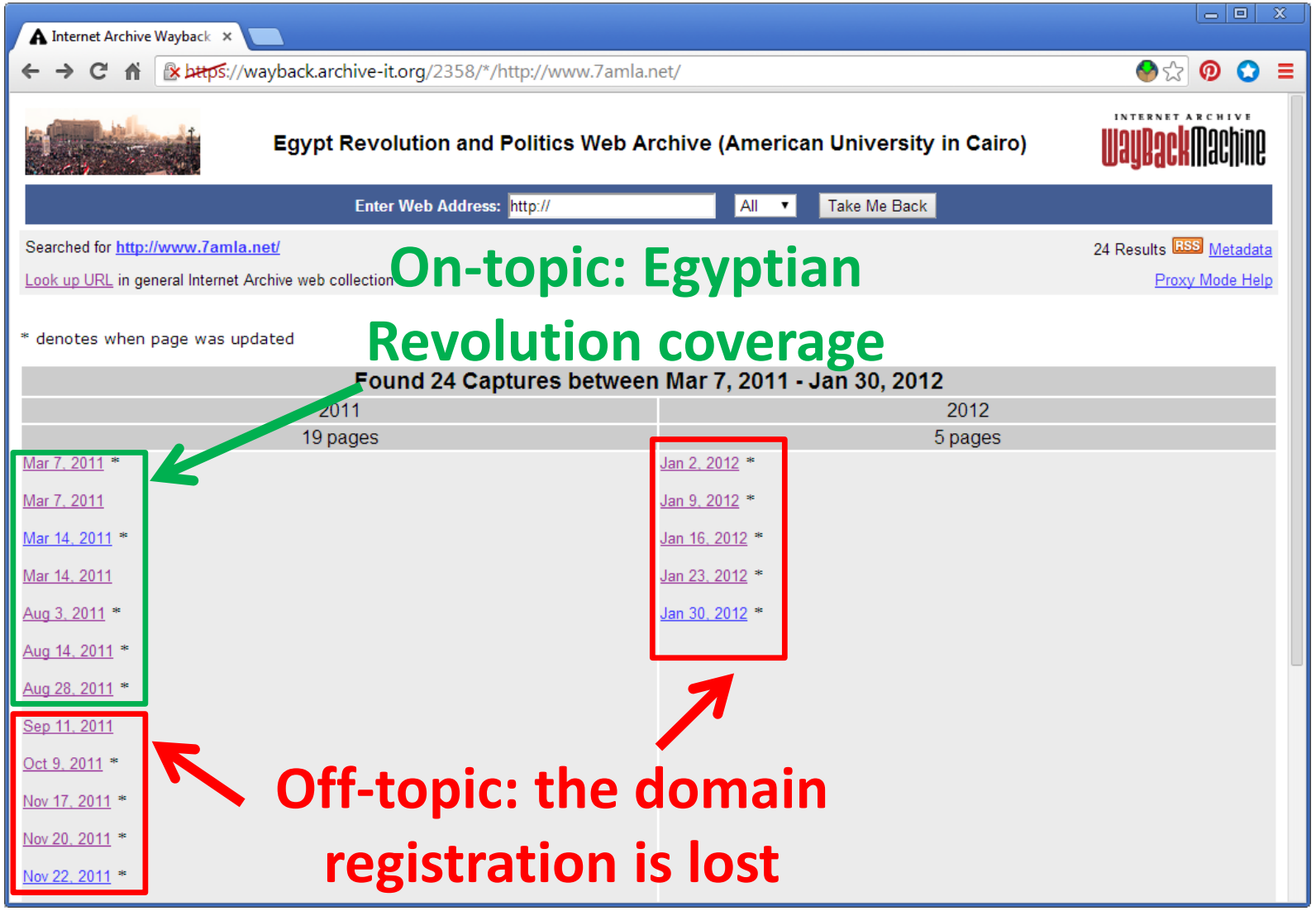
4. Oscillating

5. Always Off



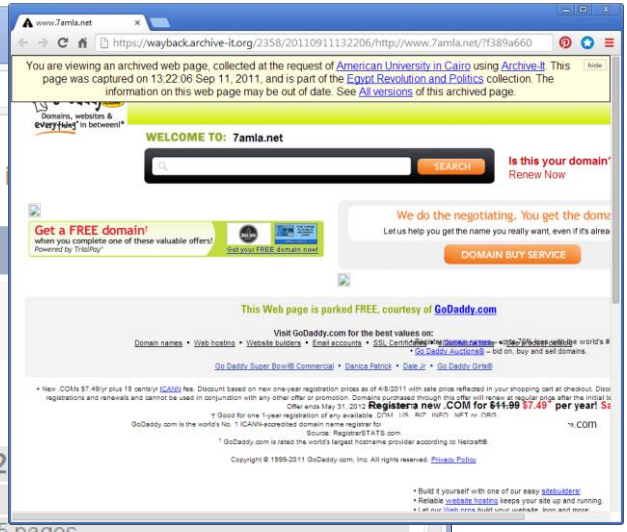
1. wayback.archive-it.org/2950/*/http://occupypsl.org
2. wayback.archive-it.org/2950/*/http://occupygso.tumblr.com
3. wayback.archive-it.org/2950/*/http://occupyashland.com
4. wayback.archive-it.org/2950/*/http://www.indyows.org
5. wayback.archive-it.org/2950/*/http://occupy605.com

A web page goes off-topic (Step Function On)



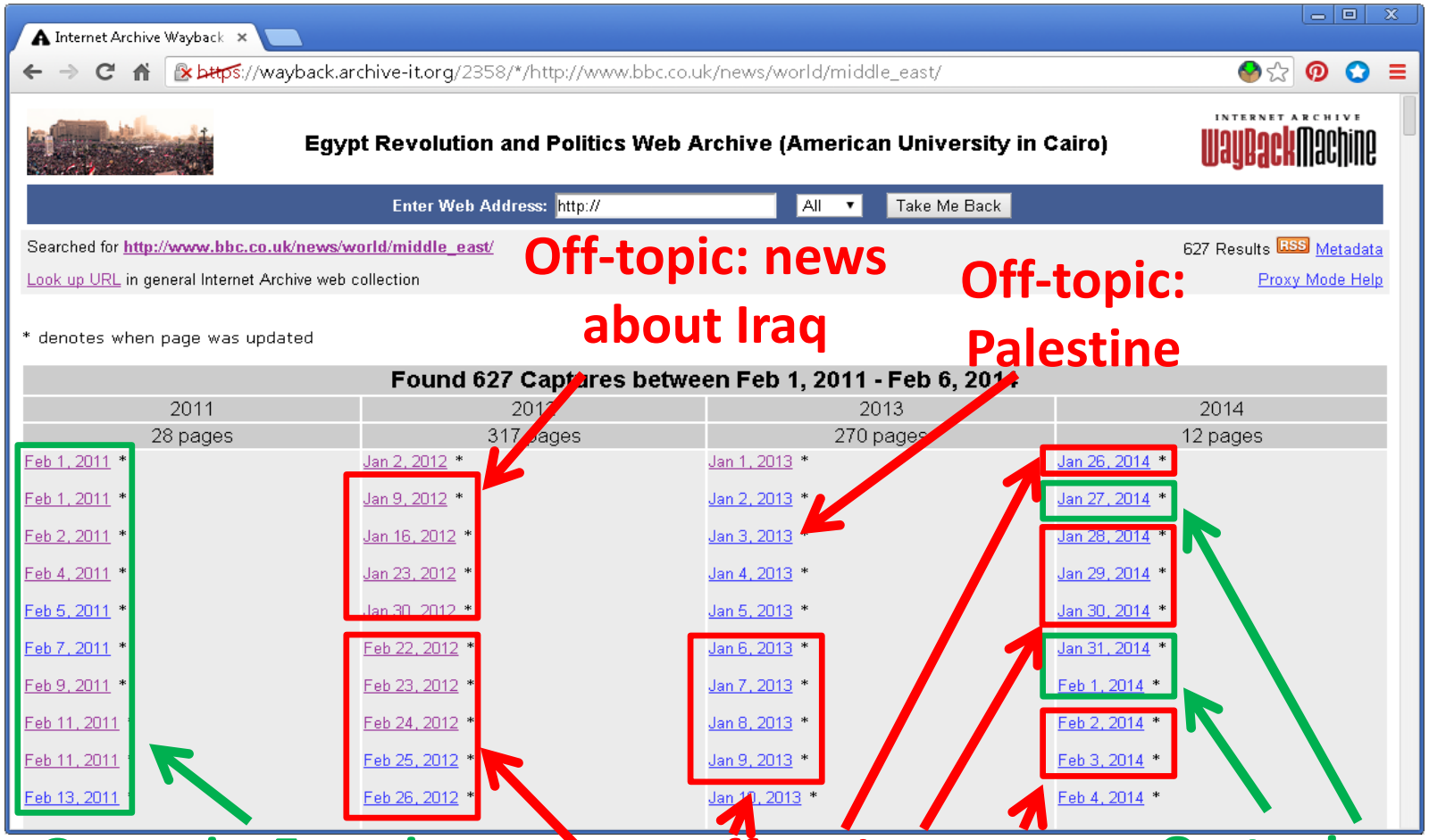
http://wayback.archive-it.org/2358/*/http://www.7amla.net

A web page goes off-topic (Step Function On)



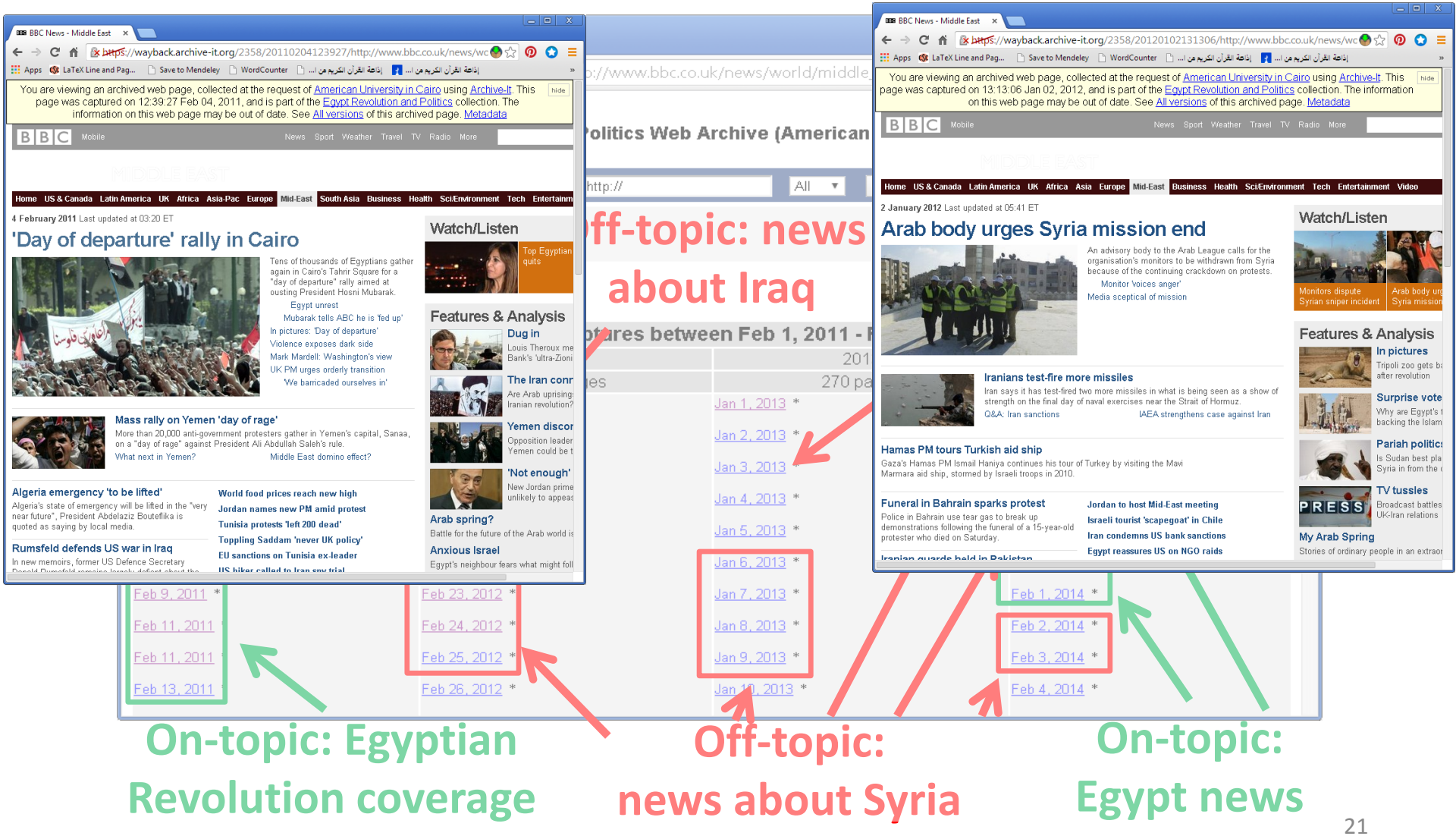
http://wayback.archive-it.org/2358/*/http://www.7amla.net

A web page goes off-topic and on-topic many times (Oscillating)



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

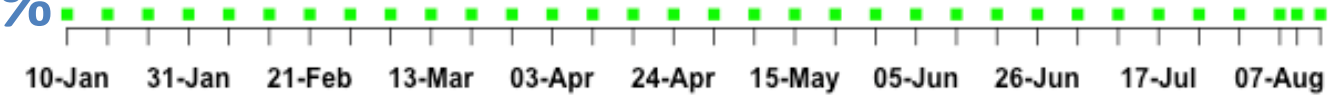
A web page goes off-topic and on-topic many times (Oscillating)



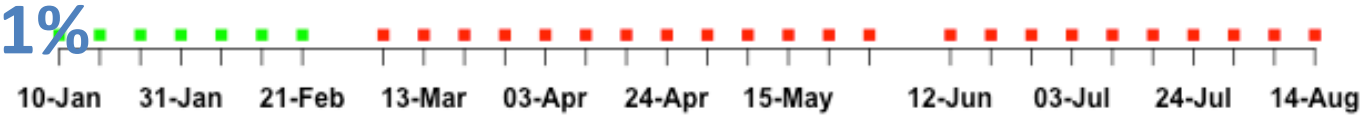
http://wayback.archive-it.org/2358/*/http://www.bbc.co.uk/news/world/middle_east/

Most TimeMaps are Always On

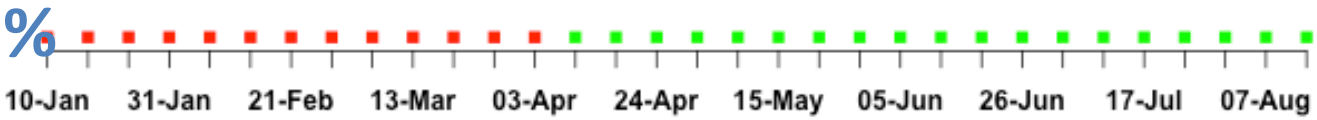
1. Always On **74%**



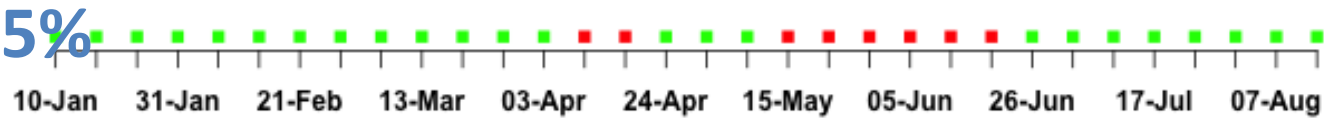
2. Step Function On **8-11%**



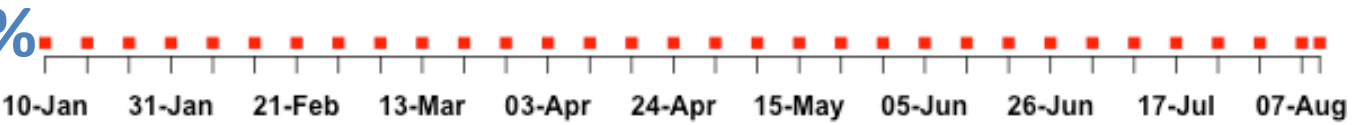
3. Step Function Off **0-2%**



4. Oscillating **6-15%**



5. Always Off **~0%**



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

Methods for detecting off-topic pages

From Archive-It collection to terms

1. Obtain the seed URIs from the front-end interface of Archive-It
2. Obtain the TimeMap of the seed URIs from the CDX file*
3. Extract the HTML of the mementos from the WARC files*
4. Extract the text of the page using the Boilerpipe library
5. Extract terms from the page, using scikit-learn to tokenize, remove stop words, and apply stemming

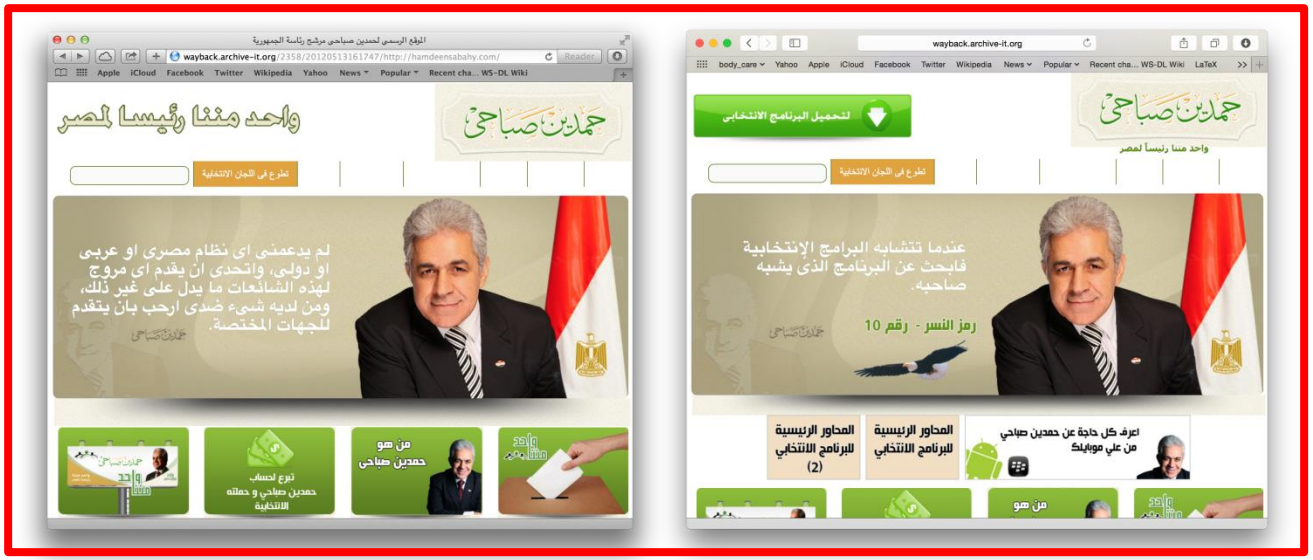
***locally hosted at ODU**

We investigated 6 similarity metrics

- Textual Content
  - cosine similarity of TF-IDF
  - intersection of the 20 most frequent terms
  - Jaccard similarity coefficient
- Semantics
  - Web-based kernel function using a search engine (SE)
- Structural
  - the change in number of words
  - the change in content length

Textual Content

cosine similarity, intersecting the most frequent terms, Jaccard similarity

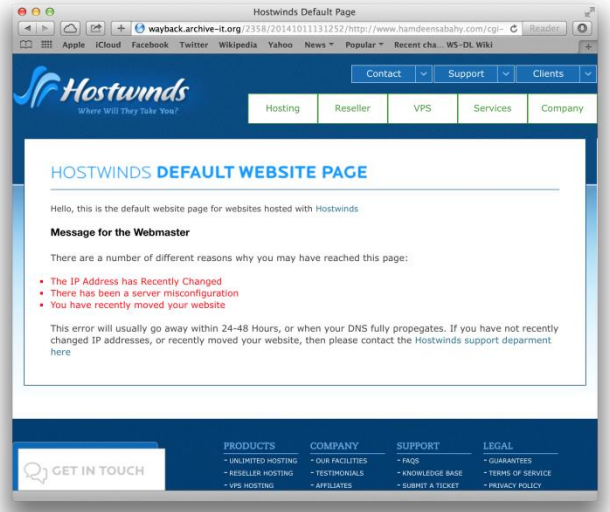


Method	Similarity
cosine	0.7
TF-Intersection	0.6
Jaccard	0.5

Textual Content

cosine similarity, intersecting the most frequent terms, Jaccard similarity

Method	Similarity
cosine	0.7
TF-Intersection	0.6
Jaccard	0.5



Method	Similarity
cosine	0.0
TF-Intersection	0.0
Jaccard	0.0

Semantics of the Text

Web based kernel function using the search engine (SE)

Feb. 2011

July 2013

Tahrir, Egypt, army

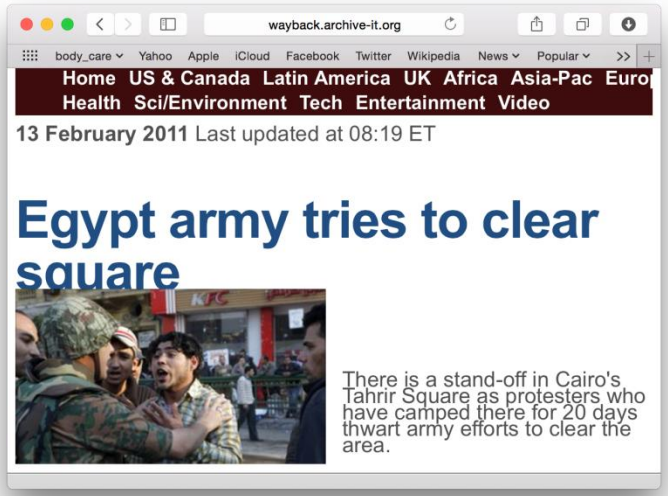
No term-wise overlap

Cairo, Morsi, protests

Semantics of the Text

Web based kernel function using the search engine (SE)

Feb. 2011



July 2013

Tahrir, Egypt, army

No term-wise overlap

Cairo, Morsi, protests

bing

Egyptian army retakes **Tahrir Square** | World news | The ...
www.theguardian.com › World › Egypt
Egypt's **army** has violently retaken Cairo's **Tahrir Square** from protesters, less than 48 hours before the former president Hosni Mubarak is to stand trial in the capital.

2012–13 Egyptian protests - Wikipedia, the free ...
en.wikipedia.org/wiki/2012–13_Egyptian_protests
The 2012-13 Egyptian **protests** were part of a large scale popular uprising in **Egypt** **against** then-President Mohamed **Morsi**. On 22 November 2012, millions of protesters ...

Egypt, Tahrir, president, protests, army, Cairo

Method	Similarity
SE-Kernel	0.7

Egypt, protests, Morsi, Cairo, president

Technique inspired by Sahami and Heilman, WWW 2006

Structural Methods

no. of words, content-length

100

109

Method	% change
WordCount	0.09

Structural Methods

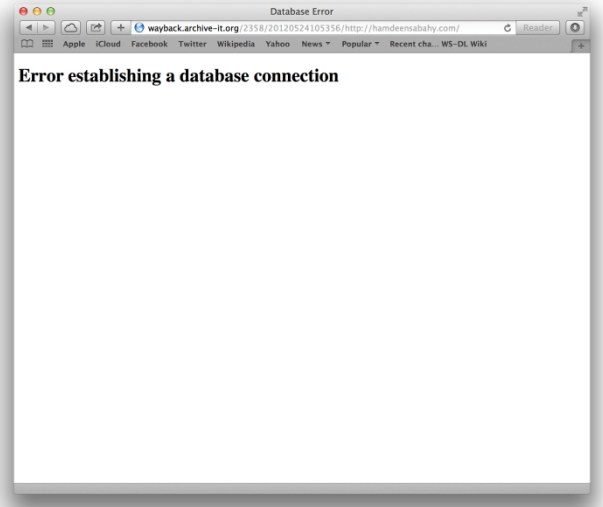
no. of words, content-length

100

109

Method	% change
WordCount	0.09

100

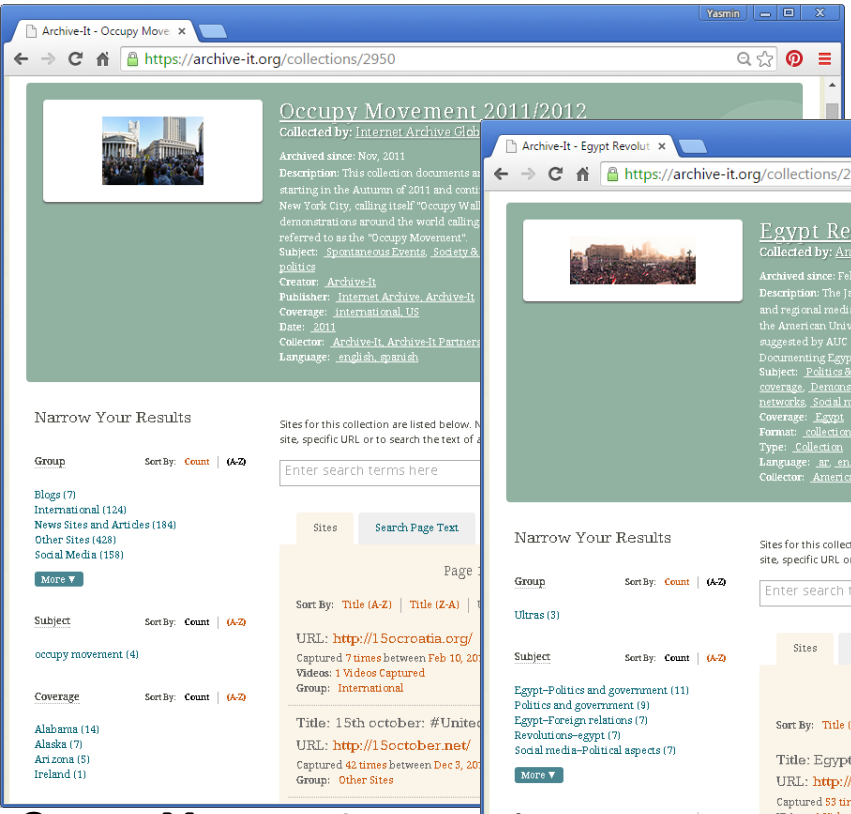


5

Method	% change
WordCount	-0.95

We built a gold standard data set to evaluate the methods

We manually labeled 15,760 mementos



Occupy Movement

URI-Rs: 255

URI-Ms: 6,570

Off-topic URI-Ms: 458

Egypt Revolution and Politics

URI-Rs: 136

URI-Ms: 6,886

Off-topic URI-Ms: 384

Columbia Univ. Human Rights collection

URI-Rs: 198

URI-Ms: 2,304

Off-topic URI-Ms: 94

Example of manually labeled set

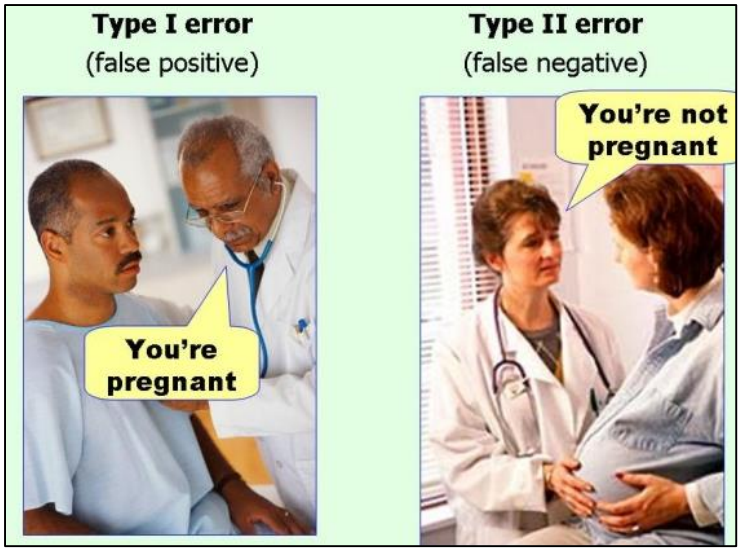
id	date	URI	label
9	20120124014240	http://wayback.archive-it.org/2950/20120124014240/http://occupysarasota.com/	1
9	20120131014118	http://wayback.archive-it.org/2950/20120131014118/http://occupysarasota.com/	1
9	20120207014119	http://wayback.archive-it.org/2950/20120207014119/http://occupysarasota.com/	1
9	20120501041141	http://wayback.archive-it.org/2950/20120501041141/http://occupysarasota.com/	0
9	20120508032644	http://wayback.archive-it.org/2950/20120508032644/http://occupysarasota.com/	0
9	20120515034720	http://wayback.archive-it.org/2950/20120515034720/http://occupysarasota.com/	0

Future work: convert to annotated/extended TimeMap format

Evaluated 6 methods at 21 thresholds

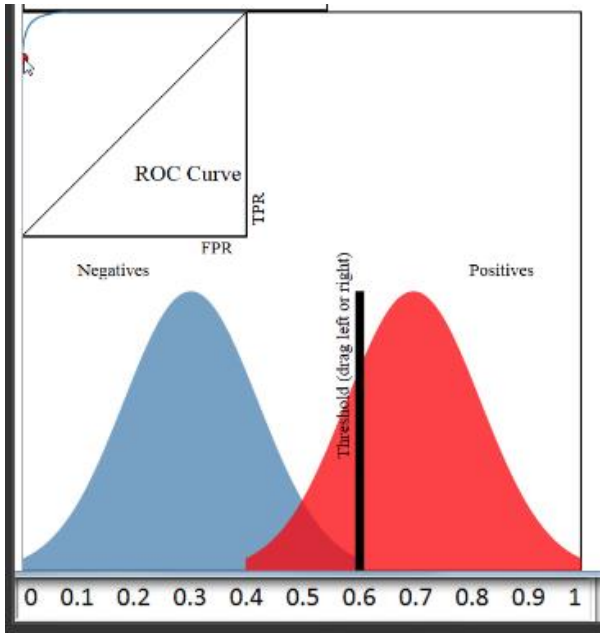
- Assumed first memento was on-topic
- Combined two methods ('OR') to find best combination method
  - 15 combinations
  - 6,615 tests (15 combinations x 21 thresholds x 21 thresholds)
- Averaged the results at each threshold over the three collections

Evaluated based on 5 metrics



- False positives (FP)
  - on-topic labeled as off-topic
- False negatives (FN)
  - off-topic labeled as on-topic
- Accuracy (ACC)
  - proportion of correct classifications
  - (TP + TN)/(TP + FP + FN + TN)

- F_1 score
  - weighted average of precision and recall
  - 2TP/(2TP + FP + FN)
- AUC
  - area under the ROC curve
  - ROC - plots false positive rate vs. true positive rate



Cosine Similarity performed well

Similarity Measure	Threshold	FP	FN	FP+FN	ACC	F1 ▼	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
Cosine	0.15	31	22	53	0.983	0.881	0.961
WordCount|SEKernel	-0.80|0.00	14	27	42	0.985	0.818	0.885
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

Finding off-topic pages in other Archive-It collections

Applied best method to 11 Archive-It collections

- Cosine|Word Count with 0.10|-0.85 thresholds

- Collection Characteristics
  - governmental, event-based, theme-based
  - time spans of 1 week - 7 years
  - 35 - 1459 URI-Rs
  - 118 - 10,283 URI-Ms

Average precision of 0.92 on 11 Archive-It collections

ID	Collection	URI-Rs	URI-Ms	Off-topic URI-Ms	Affected URI-Rs	TP	FP	P
2893	Global Food Crisis	65	3063	22	7	22	0	1.000
1084	Government in Alaska	68	506	16	4	16	0	1.000
2966	Virginia Tech Shootings	239	1670	24	2	24	0	1.000
2017	Wikileaks 2010 Document	35	2360	107	8	107	0	1.000
2323	Jasmine Revolution 2011	231	4076	114	31	107	7	0.939
1827	IT Historical Resource	1459	10,283	59	34	45	14	0.763
1475	Human Rights Document	147	1530	54	20	39	15	0.722
1826	Maryland State Document	69	184	0	0	-	-	-
694	April 16 Archive	35	118	0	0	-	-	-
2535	Brazilian School Shooting	476	1092	0	0	-	-	-
2823	Russia Plane Crash	65	447	0	0	-	-	-

Summary

- We investigated six methods for measuring similarity between mementos in a TimeMap:
  - cosine similarity of TF-IDF
  - Jaccard similarity
  - intersection of the 20 most frequent terms
  - Web-based kernel function
  - change in number of words
  - change in content length
- We tested the approaches on a gold standard data set from three Archive-It collections
- We evaluated best approach on 11 diverse Archive-It collections

Findings

- Combining cosine similarity at threshold 0.10 and change in size using word count at threshold –0.85 gives the best performance

- Cosine similarity at threshold = 0.15 is the best single method

- Using the combined method, we achieved 0.92 average precision on 11 Archive-It collections

Tool for detecting off-topic pages

- A python command-line tool for suggesting off-topic pages in web archives
  - Cosine Similarity
  - default threshold is 0.15
  - operates on live TimeMaps

Available at

https://github.com/yasmina85/OffTopic-Detection

Detecting off-topic pages in an Archive-It collection (Maryland State Docs)

```
% 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

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/
```

This was run live after we did the evaluation, so now there are off-topic mementos

Detecting off-topic pages in a single TimeMap

```
% python detect_off_topic.py -t https://wayback.archive-
it.org/2358/timemap/link/http://hamdeensabahy.com/

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/
…
Downloading 270 mementos out of 270
…
Extracting text from the html
…
Detecting off-topic mementos

Similarity          memento_uri
0.0509170839413     http://wayback.archive-
it.org/2358/20140524131241/http://www.hamdeensabahy.com/
0.0                 http://wayback.archive-
it.org/2358/20130321080254/http://hamdeensabahy.com/
0.0368021561791     http://wayback.archive-
it.org/2358/20130621131337/http://www.hamdeensabahy.com/
0.12899637517       http://wayback.archive-
it.org/2358/20140602131307/http://hamdeensabahy.com/
…
```

We're continuing work on this

- Enhancements to the detection tool
  - add the other similarity methods (WordCount first)
  - allow input of local CDX and WARC files
- Investigate characteristics of collections and TimeMaps that affect choosing thresholds
- Detect off-topic seeds (URI-Rs) in a collection
  - determine collection aboutness

Tools for Managing Seed URIs
(Detecting Off-Topic Pages)

Yasmin AlNoamany, Michele C. Weigle, Michael L. Nelson

Old Dominion University

Web Science and Digital Libraries Group

http://ws-dl.cs.odu.edu/, @WebSciDL

Python Tool: https://github.com/yasmina85/OffTopic-Detection

Web Archiving Collaboration: New Tools and Models
June 4-5, 2015