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 - Curated Web Collections

- > 3,000 collections
- ~340 institutions
- > 10B archived pages
Archive-It Interface - Curator's View
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 hamdeensabahy.com archived versions (266) 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/*
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 types 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)

On-topic: Egyptian Revolution coverage

Off-topic: the domain registration is lost

http://wayback.archive-it.org/2358/*/http://www.7amla.net
A web page goes off-topic (Step Function On)

On-topic: Egyptian Revolution coverage

Off-topic: the domain registration is lost

http://wayback.archive-it.org/2358/*http://www.7amla.net
A web page goes off-topic and on-topic many times (Oscillating)

On-topic: Egyptian Revolution coverage

Off-topic: news about Iraq

Off-topic: Palestine

Off-topic: news about Syria

On-topic: Egypt news

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)

On-topic: Egyptian Revolution coverage

Off-topic: news about Syria

On-topic: Egypt news

Off-topic: news about Iraq

Off-topic: news about Syria

Off-topic: Palestine

On-topic: Arab body urges Syria mission end

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
Pre-processing

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
Similarity Metrics

• Textual Content
  – cosine similarity
  – intersection of the most frequent terms
  – Jaccard coefficient

• Semantics
  – Web based kernel function using the search engine (SE)

• Structural
  – the change in number of words
  – the change in content length
cosine similarity, intersecting the most frequent terms, Jaccard similarity

<table>
<thead>
<tr>
<th>Method</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>cosine</td>
<td>0.7</td>
</tr>
<tr>
<td>TF-Intersection</td>
<td>0.6</td>
</tr>
<tr>
<td>Jaccard</td>
<td>0.5</td>
</tr>
</tbody>
</table>
cosine similarity, intersecting the most frequent terms, Jaccard similarity
Semantics of the Text
Web based kernel function using the search engine (SE)

Feb. 2011

Tahrir, Egypt, army

July 2013

Cairo, Morsi, protests

No term-wise overlap
Semantics of the Text
Web based kernel function using the search engine (SE)

Feb. 2011

Egypt army tries to clear square

Tahrir, Egypt, army

No term-wise overlap

July 2013

Brotherhood HQ in Cairo stormed

Cairo, Morsi, protests

Method | Similarity
---|---
SE-Kernel | 0.7

Egypt, Tahrir, president, protests, army, Cairo

Egypt, protests, Morsi, Cairo, president

Technique inspired by Sahami and Heilman, WWW 2006
Structural Methods

no. of words, content-length

<table>
<thead>
<tr>
<th>Method</th>
<th>% change</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordCount</td>
<td>0.09</td>
</tr>
</tbody>
</table>
Structural Methods

no. of words, content-length

<table>
<thead>
<tr>
<th>Method</th>
<th>% change</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordCount</td>
<td>0.09</td>
</tr>
<tr>
<td>WordCount</td>
<td>-0.95</td>
</tr>
</tbody>
</table>
We built a gold standard data set to evaluate the methods
We manually labeled 15,760 mementos

Occupy Movement
URI-Rs: 255
URI-Ms: 6,670
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
Manually labeled set is available

https://github.com/yasmina85/OffTopic-Detection/tree/master/collections_dataset

<table>
<thead>
<tr>
<th>id</th>
<th>date</th>
<th>URI</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>20120124014240</td>
<td><a href="http://wayback.archive-it.org/2950/20120124014240/http://occupysarasota.com/">http://wayback.archive-it.org/2950/20120124014240/http://occupysarasota.com/</a></td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>20120131014118</td>
<td><a href="http://wayback.archive-it.org/2950/20120131014118/http://occupysarasota.com/">http://wayback.archive-it.org/2950/20120131014118/http://occupysarasota.com/</a></td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>20120207014119</td>
<td><a href="http://wayback.archive-it.org/2950/20120207014119/http://occupysarasota.com/">http://wayback.archive-it.org/2950/20120207014119/http://occupysarasota.com/</a></td>
<td>1</td>
</tr>
</tbody>
</table>

Future work: convert to annotated/extended TimeMap format
Evaluation 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
Evaluation Setup

• Assumed first memento was on-topic

• Tested each method on each labeled memento at 21 different thresholds

• 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
Cosine Similarity performed well

<table>
<thead>
<tr>
<th>Similarity Measure</th>
<th>Threshold</th>
<th>FP</th>
<th>FN</th>
<th>FP+FN</th>
<th>ACC</th>
<th>F1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine</td>
<td>WordCount</td>
<td>0.10</td>
<td>-0.85</td>
<td>24</td>
<td>10</td>
<td>34</td>
<td>0.987</td>
</tr>
<tr>
<td>Cosine</td>
<td>SEKernel</td>
<td>0.10</td>
<td>0.00</td>
<td>6</td>
<td>35</td>
<td>40</td>
<td>0.990</td>
</tr>
<tr>
<td>Cosine</td>
<td>0.15</td>
<td>31</td>
<td>22</td>
<td>53</td>
<td>0.983</td>
<td>0.881</td>
<td>0.961</td>
</tr>
<tr>
<td>WordCount</td>
<td>SEKernel</td>
<td>-0.80</td>
<td>0.00</td>
<td>14</td>
<td>27</td>
<td>42</td>
<td>0.985</td>
</tr>
<tr>
<td>WordCount</td>
<td>-0.85</td>
<td>6</td>
<td>44</td>
<td>50</td>
<td>0.982</td>
<td>0.806</td>
<td>0.870</td>
</tr>
<tr>
<td>SEKernel</td>
<td>0.05</td>
<td>64</td>
<td>83</td>
<td>147</td>
<td>0.965</td>
<td>0.683</td>
<td>0.865</td>
</tr>
<tr>
<td>Bytes</td>
<td>-0.65</td>
<td>28</td>
<td>133</td>
<td>161</td>
<td>0.962</td>
<td>0.584</td>
<td>0.746</td>
</tr>
<tr>
<td>Jaccard</td>
<td>0.05</td>
<td>74</td>
<td>86</td>
<td>159</td>
<td>0.962</td>
<td>0.538</td>
<td>0.809</td>
</tr>
<tr>
<td>TF-Intersection</td>
<td>0.00</td>
<td>49</td>
<td>104</td>
<td>153</td>
<td>0.967</td>
<td>0.537</td>
<td>0.740</td>
</tr>
</tbody>
</table>
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

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

• We presented six methods for measuring similarity between pages:
  – cosine similarity
  – Jaccard similarity
  – intersection of the most 20 frequent terms
  – Web-based kernel function
  – change in number of words
  – 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 4 mementos out of 306
Downloading 14 mementos out of 306
...
Detecting off-topic mementos

<table>
<thead>
<tr>
<th>Similarity</th>
<th>memento_uri</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td><a href="http://wayback.archive-it.org/1826/20131220205908/http://www.mncppc.org/commission_home.html/">http://wayback.archive-it.org/1826/20131220205908/http://www.mncppc.org/commission_home.html/</a></td>
</tr>
<tr>
<td>0.0</td>
<td><a href="http://wayback.archive-it.org/1826/20141118195815/http://www.mncppc.org/commission_home.html/">http://wayback.archive-it.org/1826/20141118195815/http://www.mncppc.org/commission_home.html/</a></td>
</tr>
</tbody>
</table>

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
...
Download 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/
...
Future Work

• Enhancements to the detection tool
  – add WordCount method
  – allow input of local CDX and WARC files

• Determine collection aboutness
  – detect off-topic URI-Rs in a collection
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