Detecting Off-Topic Pages in Web Archives

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

Old Dominion University, Department of Computer Science
Norfolk VA 23529, USA
{yasmin,mweigle,mln}@cs.odu.edu

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 off-topic pages in Web archive collections. We evaluate six different methods to detect when the page has gone off-topic through subsequent captures. 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 the off-topic pages is 0.92.

Keywords: Archived Collections, Experiments, Analysis, Document Filtering

1 Introduction

The Internet Archive [1] (IA) is the largest and oldest of the various Web archives, holding over 400 billion Web pages with archives as far back as 1996 [2]. Archive-It[1] is a collection development service operated by the Internet Archive since 2006. Archive-It is currently used by over 340 institutions in 48 states, and features over 9B archived Web pages in nearly 2800 separate collections.

Archive-It provides their partners with tools that allow them to build themed collections of archived Web pages hosted at Archive-It. 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 Internet Archive Global 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).

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

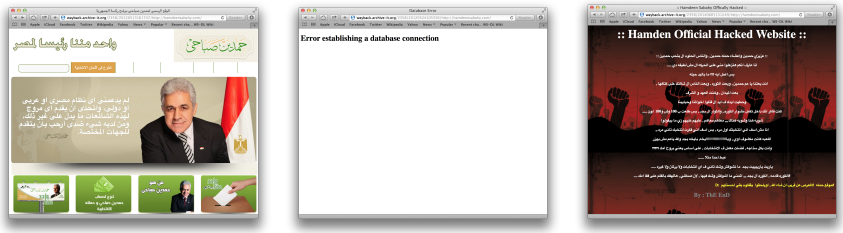
The Heritrix [3] crawler at Archive-It crawls/recrawls these seeds based on the predefined frequency and depth to build a collection of archived Web pages that the curator believes best exemplifies the topic of the collection. Archive-It has deployed tools that allow a collection's curators to perform 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 and how many 404 missing URIs) and domain information (e.g., how many .uk sites vs. .com sites). Currently, there are no content-based tools that allow curators to detect when seed URIs go off-topic.

In this paper, we evaluate different approaches for detecting off-topic pages in the archives. The approaches depend on comparing versions of the pages through time. Three methods depend on the textual content (cosine similarity, intersection of the most frequent terms, and Jaccard coefficient), one method uses the semantics of the text (Web-based kernel function using a search engine), and two methods use the size of pages (the change in number of words and the content length). For evaluation purposes, we built our gold standard data set from three Archive-It collections, then we employed the following performance measurements: accuracy, F_1 score, and area under the ROC curve (AUC). Experimental results show that cosine similarity at the 0.15 threshold is the most effective single method in detecting the off-topic pages with 0.983 accuracy. We paired several of the 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 then used this combined method 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 0.92.

2 Motivating Example

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 [4]. 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". In this case, the Web page returns a 200 HTTP response but with unwanted content [5].

Figure 1 shows a scenario of a page going off-topic for several reasons. In May 2012, `hamdeensabahy.com`, which belonged to a candidate in Egypt's 2012 presidential election, is originally relevant to the "Egypt Revolution and Politics" collection (Figure 1(a)). Then, the page went back and forth between on-topic and off-topic many times for different reasons. Note that there are on-topic pages between the off-topic ones in Figure 1. In the example, the site went off-topic



(a) May 13, 2012: The page started as on-topic. (b) May 24, 2012: Off-topic due to a database error. (c) June 5, 2014: The site has been hacked.

Fig. 1. A site for one of the candidates from Egypt's 2012 presidential election.

because of a database error on May 24, 2012 (Figure 1(b)), then it returned on-topic again. After that, the page went off-topic between late March 2013 and early July 2013. The site went on-topic again for a period of time, then it was hacked (Figure 1(c)), and then the domain was lost by late 2014. Today, `hamdeensabahy.com` is not available on the live Web.

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 Figure 1(b)-1(c) do not contribute to the Egypt Revolution collection in the same way that the on-topic archived Web site in Figure 1(a) does. Although the former can be kept in the IA's general Web archive, it is a candidate to be purged from the Egyptian Revolution collection, or at the very least it should not be considered when summarizing the collection. Sites like `hamdeensabahy.com` that currently are not available on the live Web do not contribute to the collection, and assisting curators to identify and remove such pages is the focus of this paper.

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 [6]. 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 [7] is an HTTP protocol extension which enables time travel on the Web by linking the current resources with their prior state. Memento defines

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
Time span	12/03/2011-10/09/2012	02/01/2011-04/18/2013	05/15/2008-03/21/2013
Total URI-Rs	728	182	560
Total URI-Ms	21,268	18,434	6,341
Sampled URI-Rs	255 (35%)	136 (75%)	198 (35%)
Sampled URI-Ms	6,570	6,886	2,304
Off-topic URI-Ms	458 (7%)	384 (9%)	94 (4%)
URI-Rs with off-topic URI-Ms	67 (26%)	34 (25%)	33 (17%)

Table 1. Description of the Archive-It collections, including manual labeling of on and off-topic URI-Ms.

several terms that we will use throughout. A 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). A URI-M identifies an archived snapshot of the URI-R at a specific datetime, which is called Memento-Datetime, e.g., $URI\text{-}M_i = URI\text{-}R@t_i$. A 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

Mining the past Web is different from Web content mining because of the temporal dimension of the archived content [6, 8]. 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 [9, 6, 10]. Below, we outline some of the approaches that have been used for mining the past Web using data in Web archives.

Jatowt and Tanaka [6] 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. [11] proposed an interactive visualization system called Page History Explorer (PHE), an application for providing overviews of 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.

Francisco-Revilla et al. [12] described the Walden's Paths Path Manager system, which checks a list of Web pages for relevant changes using document signatures of paragraphs, headings, links, and keywords. Ben Saad et al. [13] claimed that using patterns is an effective way to predict changes, and then used this prediction to optimize the archiving process by crawling only important pages.

Spaniol and Weikum used Web archive data to track the evolution of entities (e.g., people, places, things) through time and visualize them [14]. 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. [10] also utilized the public archival repositories for automatically detecting the age of Web content through the past snapshots of pages.

Most of the previous papers used the change of web pages for optimizing the crawl or visualization. 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 assisting curators in identifying the pages that are off-topic in a TimeMap, so these pages can be excluded from the collections.

4 Data Set

In this section we describe our gold standard dataset. We evaluate our techniques using the ODU mirror of Archive-It's collections. ODU has received a copy of the Archive-It collections through April 2013 in Web ARchive file format (WARC) [15]. The three collections in our dataset differ in terms of the number of URI-Rs, number of URI-Ms, and time span over which the Web pages have been archived (ending in April 2013). The three collections, which include pages in English, Arabic, French, Russian, and Spanish, are described below. The details are provided in Table 1.

Occupy Movement 2011/2012 covers the Occupy Movement protests and the international branches of the Occupy Wall Street movement.

Egypt Revolution and Politics covers the January 25th Egyptian Revolution and Egyptian politics, contains different kinds of Web sites (e.g., social media, blogs, news, etc.).

Human Rights covers documentation and research about human rights that has been created by non-governmental organizations, national human rights institutions, and individuals.

We randomly sampled 588 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. We manually labeled these 15,760 mementos as on-topic or off-topic. The bottom portion of Table 1 contains the results of this labeling.

5 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.

5.1 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 TimeMap of the seed URIs from the CDX file[2], (3) extract the HTML of the mementos from the WARC files (hosted at ODU), (4) extract the text of the page using the Boilerpipe library [16], and (5) extract terms from the page, using scikit-learn [17] to tokenize, remove stop words, and apply stemming.

5.2 Methods for Detecting Off-Topic Pages

In this section, we use different similarity measures between pages to detect when the *aboutness(URI-R)* over time changes and to define a threshold that separates the on-topic and the off-topic pages.

Cosine similarity. Cosine similarity is one of the most commonly used similarity measures in information retrieval. After text preprocessing, we calculated the TF-IDF, then we applied cosine similarity to compare the *aboutness(URI-R@t_0)* with *aboutness(URI-R@t)*.

Jaccard similarity coefficient. The Jaccard similarity coefficient is the size of the intersection of two sets divided by the size of their union. After preprocessing the text (step 5), we apply the Jaccard coefficient on the resulting terms to specify the similarity between the *URI-R@t* and *URI-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-R@t*, and then compared the top 20 most frequent terms of the *URI-R@t* with the top 20 most frequent terms of the *URI-R@t_0*. The size of the intersection between the top 20 terms of *URI-R@t* and *URI-R@t_0* represents the similarity between the mementos. We name this method TF-Intersection.

Web-based kernel function. The previous methods are term-wise similarity measures, i.e., they use lexicographic term matching. But these methods may not suitable for archived collections with a large time span or pages that contain a small amount of text. For example, the Egyptian Revolution collection is from February 2011 until April 2013. Suppose a page in February 2011 has terms like "Mubarak, Tahrir, Square" and a page in April 2013 has terms like "Morsi, Egypt". 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

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

on the semantic context of the documents. The work by Sahami and Heilman [18] inspired us to augment the text of *URI-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, which have been well-studied in information retrieval[19]. We used the contextually descriptive snippet text returned with results from the Bing Search API. We call this method "SEKernel".

We augment the terms of *URI-R*@t_0 with semantic context from the search engine as follows:

1. Format a query q from the top 5 words of the first memento (*URI-R*@t_0).
2. Issue q to the search engine SE.
3. Extract the terms p from the top 10 snippets returned for q.
4. Add the terms of the snippets p to the terms of the original text of the first memento to have a new list of terms, $ST = p \cup$URI-R@t_0.
5. $\forall t \geq 1$, calculate the Jaccard coefficient between the expanded document ST and the terms of *URI-R*@t.

If we apply this method on the previous example, we use terms "Mubarak, Tahrir, Square" as search keywords to generate semantic context. The resulting snippet will have new terms like "Egypt, President", which term-wise overlaps with the page that contains "Morsi, Egypt".

Change in size We noticed that the size 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, tahrir, the, square) in the page. For example, assume *URI-R*@t_0 contains 100 words and *URI-R*@t contains 5 words. This represents a 95% decrease in the number of words between *URI-R*@t_0 and *URI-R*@t.

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 HTML forming the page template is still large. There are many cases where the page goes off-topic and the size of the page decreases or in some cases reaches 0 bytes, e.g., the account is suspended, transient errors, or no content in the page. One of the advantages of using the structural-based methods over the textual-based methods is that structural-based methods are language independent. Many of the collections are multi-lingual, and each language needs special processing. The structural methods are suitable for those collections.

6 Evaluation

In this section, we define how we evaluate the methods presented in Section 5.1 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.

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
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

Table 2. Evaluating the similarity approaches, averaged over the three collections.

6.1 Evaluation Metrics

We used multiple metrics to evaluate the performance of the similarity measures. We considered false positives (FP), the number of on-topic pages predicted as off-topic; false negatives (FN), the number of off-topic pages predicted as on-topic; accuracy (ACC), the fraction of the classifications that are correct; F_1 score (also known as F-measure), the weighted average of precision and recall; and the ROC AUC score (AUC), a single number that computes the area under the receiver operating characteristic curve [20].

6.2 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. In order to determine the best threshold, we used the evaluation metrics described in the previous section. To say that *URI-R@t* is off-topic at $th = 0.15$ means that the similarity between *URI-R@t* and *URI-R@t*$_0$ is < 0.15. On-topic means the similarity between *URI-R@t* and *URI-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 change of size measures represent the decrease in size, so -1 means the page has a 100% decrease from *URI-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-R@t* and *URI-R@t*$_0$ indicates that the page might be off-topic. Therefore, if the change in size between *URI-R@t* and *URI-R@t*$_0$ is less than −95%, that means *URI-R@t* is off-topic at $th = -95\%$.

Table 2 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.

There was consistency among the values of th with the best performance of TF-Intersection, Jaccard, and SEKernel methods for the three collections, e.g., for all the 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 Occupy Movement collection, Egypt Revolution collection, and Human Rights collection are 0.2, 0.15, 0.1 respectively. We took the average of the three collections at $th = 0.2$, $th = 0.15$, and $th = 0.1$, then based on the best F_1 score, we specified the threshold that has the best average performance.

Specifying a threshold for detecting the off-topic pages is not easy due to differences in the nature of the collections. For example, long running collections such as the Human Rights collection (2009–present) have more opportunities for 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.

6.3 Combining the Similarity Measures

We tested 6,615 pairwise combinations (15 method combinations $\times 21 \times 21$ 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 the bottom portion of Table 2. We present the three best average combinations of the similarity measures based on F_1 score and AUC. Performance increases with combining Cosine and WordCount (Cosine|WordCount) at $th = 0.1| - 0.85$. There is a 36% decrease in errors (FP+FN) than the best performing single measure, Cosine. Furthermore, Cosine|WordCount has a 3% increase in the F_1 score over Cosine. Cosine|SEKernel at $th = 0.1|0.0$ has a 2% increase in F_1 over Cosine, while WordCount|SEKernel at $th = -0.80|0.00$ has lower performance than Cosine.

In summary Cosine|WordCount gives the best performance at $th = 0.1|-0.85$ across all the single and combined methods. Moreover, combining WordCount with Cosine does not cause much overhead in processing, because WordCount uses tokenized words and needs no extra text processing.

7 Evaluating Archive-It Collections

We applied the best performing method (Cosine|WordCount) with the suggested thresholds ($0.1| - 0.85$) on unlabeled Archive-It collections. We chose different

Collection	ID	Time Span	URI-Rs	URI-Ms	Affected URI-Rs	TP	FP	P
Global Food Crisis	2893	10/19/2011-10/24/2012	65	3,063	7	22	0	1.00
Government in Alaska	1084	12/01/2006-04/13/2013	68	506	4	16	0	1.00
Virginia Tech Shootings	2966	12/08/2011-01/03/2012	239	1,670	2	24	0	1.00
Wikileaks 2010 Document Release Collection	2017	07/27/2010-08/27/2012	35	2,360	8	107	0	1.00
Jasmine Revolution - Tunisia 2011	2323	01/19/2011-12/24/2012	231	4,076	31	107	7	0.94
IT Historical Resource Sites	1827	2/23/2010-10/04/2012	1,459	10,283	34	45	14	0.76
Human Rights Documentation Initiative	1475	04/29/2009-10/31/2011	147	1,530	20	39	15	0.72
Maryland State Document Collection	1826	03/04/2010-12/03/2012	69	184	0	-	-	-
April 16 Archive	694	05/23/2007-04/28/2008	35	118	0	-	-	-
Brazilian School Shooting	2535	04/09/2011-04/14/2011	476	1,092	0	-	-	-
Russia Plane Crash Sept 7,2011	2823	09/08/2011-09/15/2011	65	447	0	-	-	-

Table 3. The results of evaluating Archive-It collections through the assessment of the detected off-topic pages using Cosine|WordCount methods at $th = 0.10| - 0.85$.

types of collections, e.g., governmental collections (Maryland State Document Collection, Government in Alaska), event-based collections (Jasmine Revolution - Tunisia 2011, Virginia Tech Shootings), and theme-based collections (Wikileaks 2010 Document Release Collection, Human Rights Documentation Initiative). Table 3 contains the details of the 11 tested collections. We extracted the tested collections from the ODU mirror of Archive-It's collections. The number of URI-Rs in the table represents those URI-Rs with more than one memento.

The results of evaluating Cosine|WordCount are shown in Table 3. For the reported results for each TimeMap for each method, we manually assessed the FP and TP and then calculated the precision $P = TP/(TP + FP)$. We cannot compute recall since we cannot know how many off-topic mementos were not detected (FN). Precision is near 1.0 for five collections. Precision=0.72 for the "Human Rights Documentation" collection, with 15 FP. Those 15 URI-Ms affected three TimeMaps. An example of one of the affected TimeMaps (`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 newer versions use Adobe Flash. The 14 FPs from the "IT Historical Resource Sites" collection affected 5 URIs because the content of the pages changed dramatically through time. There are four collections that have no reported off-topic pages. Two of these collections, "Brazilian School Shooting" and "Russia Plane Crash", span only one week, which is typically not enough time for pages to go off-topic. The third collection with no detected off-topic mementos is the "Maryland State Document" collection. Perhaps this collection simply had well-chosen seed URIs.

In summary, Cosine|WordCount at $th = 0.1|-0.85$ performed well on Archive-It collections with average $P = 0.92$. Studying the FP cases has given us an idea about the limitations of the current approach, such as detecting URIs with little to no textual content or URIs whose topics change significantly through time.

8 Conclusions and Future Work

In this paper, we present approaches for assisting the curator in identifying off-topic mementos in the archive. 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, and change in content length. We tested the approaches on three different labeled subsets of collections from Archive-It. We found that of the single methods, the cosine similarity measure is the most effective method for detecting the off-topic pages at $th = 0.15$. The change in size based on the word count comes next at $th = -0.85$. We 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 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 11 Archive-It collections. We manually assessed the relevancy of the detected off-topic pages and found that the average precision = 0.92. In summary, evaluating the suggested approach, Cosine|WordCount at $th = 0.10| - 0.85$, for detecting the off-topic pages in the archives has shown good results with 0.92 precision. The presented approach will help curators to judge their crawls and also will obviate users from getting unexpected content when they access archived 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. The code and gold standard data set are available at `https://github.com/yasmina85/OffTopic-Detection`.

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. The next step is to compute a model of the topic of the collection, in part to more easily detect seeds that begin off-topic.

9 Acknowledgments

This work supported in part by the Andrew Mellon Foundation. We thank Kristine Hanna from Internet Archive for facilitating obtaining the data set.

References

1. 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)
2. Kahle, B.: Wayback Machine Hits 400,000,000,000! http://blog.archive.org/2014/05/09/wayback-machine-hits-400000000000/ (2014)
3. Mohr, G., Stack, M., Ranitovic, I., Avery, D., Kimpton, M.: An Introduction to Heritrix An open source archival quality web crawler. In: Proceedings of IWAW. (2004) 43–49
4. Marshall, C., McCown, F., Nelson, M.: Evaluating Personal Archiving Strategies for Internet-based Information. In: Proceedings of Archiving. (2007) 151–156
5. Bar-Yossef, Z., Broder, A.Z., Kumar, R., Tomkins, A.: Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay. In: Proceedings of WWW. (2004) 328–337
6. Jatowt, A., Tanaka, K.: Towards mining past content of Web pages. New Review of Hypermedia and Multimedia **13**(1) (July 2007) 77–86
7. Van de Sompel, H., Nelson, M.L., Sanderson, R.: RFC 7089 - HTTP framework for time-based access to resource states – Memento (2013)
8. Kosala, R., Blockeel, H.: Web Mining Research: A Survey. SIGKDD Exploration Newsletter. **2**(1) (June 2000) 1–15
9. 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 ACM/IEEE JCDL. (2006) 95–102
10. Jatowt, A., Kawai, Y., Tanaka, K.: Detecting Age of Page Content. In: Proceedings of ACM WIDM. (2007) 137–144
11. Jatowt, A., Kawai, Y., Tanaka, K.: Page History Explorer: Visualizing and Comparing Page Histories. IEICE Transactions on Information and Systems **94**(3) (2011) 564–577
12. Francisco-Revilla, L., Shipman, F., Furuta, R., Karadkar, U., Arora, A.: Managing change on the web. In: Proceedings of ACM/IEEE JCDL. (2001) 67–76
13. Ben Saad, M., Gançarski, S.: Archiving the Web using Page Changes Patterns: A Case Study. In: Proceedings of ACM/IEEE JCDL. (2012) 113–122
14. Spaniol, M., Weikum, G.: Tracking Entities in Web Archives: The LAWA Project. In: Proceedings of WWW. (2012) 287–290
15. ISO: ISO 28500:2009 - Information and documentation – WARC file format. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717 (2009)
16. Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate Detection Using Shallow Text Features. In: Proceedings of ACM WSDM. (2010) 441–450
17. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12** (2011) 2825–2830
18. Sahami, M., Heilman, T.D.: A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets. In: Proceedings of WWW. (2006) 377–386
19. Buckley, C., Salton, G., Allan, J., Singhal, A.: Automatic Query Expansion Using SMART: TREC 3. Overview of the Third Text REtrieval Conference (TREC-3) (1995) 69–80
20. Fawcett, T.: An Introduction to ROC Analysis. Pattern Recognition Letters **27**(8) (June 2006) 861–874