Detecting Content Drift on the Web Using Web Archives and Textual Similarity (short paper)*

Brenda Reyes Ayala[1,*], Qiufeng Du[1] and Juyi Han[1]

[1]University of Alberta, School of Library and Information Studies, 11210 87 Ave, Edmonton, Alberta T6G 2G5, Canada

Abstract

Content drift, which occurs when a website's content changes and moves away from the content it originally referenced, is problem that affects both live websites and web archives. Content drift can also occur when the page has been hacked, its domain has expired, or the service has been discontinued. In this paper, we present a simple method for detecting content drift on the live web based on comparing the titles of live websites to those of their archived versions. Our assumption was that the higher the difference between the title of an archived website and that of its live counterpart, the more likely content drift had taken place. In order to test our approach, we first had human evaluators manually judge websites from three Canadian web archives to determine or not content drift had occurred. Then we extracted the titles from all websites, and used cosine similarity to compare the title of the live websites to the title of the archived websites. Our approach achieved positive results, with an accuracy of 85.2, precision of 89.3, recall of 92.1, and F-measure values of 90.7. Having simple methods such as the one presented in this paper can allow institutions or researchers to quickly and effectively detect content drift without needing many technological resources.

Keywords

web archiving, cultural heritage, relevance, reference rot, link rot, content drift

1. Introduction

The ephemeral and transient nature of the web has been well-established. For many institutions who seek to preserve their online cultural heritage, the process of web archiving can seem like a race against time as web archivists struggle to capture websites before they disappear from the web. The danger of websites disappearing from the web altogether is part of a larger problem known as *reference rot*, which has two components, as identified by [1]:

1. Link rot: The resource identified by a URI vanishes from the web. As a result, a URI reference to the resource ceases to provide access to referenced content.
2. Content drift: The resource identified by a URI changes over time. The resource's content evolves and can change to such an extent that it ceases to be representative of the content that was originally referenced.

The process of web archiving arose in part to combat link rot; however, the subtler and more insidious problem of content drift persists in both the live web and in web archives. The study

TPDL2022: 26th International Conference on Theory and Practice of Digital Libraries, 20-23 September 2022, Padua, Italy

*Corresponding author.

brenda.reyes@ualberta.ca (B. Reyes Ayala); qiufeng@ualberta.ca (Q. Du); juyi@ualberta.ca (J. Han)

0000-0002-9342-3832 (B. Reyes Ayala)

of content drift is complicated by the appearance of *soft 404s*. A soft 404 is an incorrect HTTP status code of 200 (OK) or 3xx (redirect) that masks a correct status code of 404 (Not found). It occurs when websites redirect failed URLs to a site's homepage, thus causing it to mask the standard 404 return code that occurs when there is a failure to access a web resource [2]. Content drift can also occur when the page has been hacked, its domain has expired, or the service has been discontinued. Many web archives, such as those created using the Internet Archive's Archive-It service [3], are topic-specific or thematic in that they collect and preserve many websites that cover a single topic or news event, such as Human Rights or the COVID-19 pandemic. If a live website is affected by content drift, and the site is then crawled, the resulting archived webpage, and the web archive which contains it, will also have content drift. Within web archives, webpages affected by content drift can also be referred to as being "off-topic", and are defined as those "that have changed through time to move away from the initial scope of the page, which should be relevant to the topic of the collection" [4].

It is very useful for both web archivists and researchers to be able to determine if a live website has been affected by content drift. Web archivists who preserve websites for their institutions, if they see that a URI has drifted, can decide whether or not they wish to keep crawling it. Since web archiving technologies and services usually entail a significant investment of time, money, and resources, institutions can avoid some of these costs by ceasing to crawl websites that have drifted. Detecting content drift on the live web can prevent it from occurring in web archives. Furthermore, researchers who study reference rot on the web could use our method as a way of detecting soft 404s on the live web. The presence of soft 404s can lead researchers to underestimate the occurrence of reference rot, since these pages do not return the typical 404 "Not Found" error.

In this paper, we present an approach to detecting content drift on the live web. The purpose of this research is to find an approach to detecting content drift that simulates a human evaluator inspecting a website and determining if content drift has occurred. In the past, researchers have developed accurate methods for detecting content drift (discussed in Section 2); however, many of these approaches are computationally intensive, and, in the case of web archives, may require access to the Web ARChive (WARC) files that store the archived pages [5]. We take a different approach based on a very simple assumption: that changes in the title of a website are indicative of changes in its content, and thus large changes in the title of a website may be indicated of content drift. This approach is quick, and is suitable for researchers or other institutions that many not have access to the WARC files they are studying or have the computational power required to perform these calculations.

2. Previous work

The prevalence of reference rot in publications and the web is a topic that has been well-documented. As early as 2001, the authors of [6] noted that improved citation practices were necessary to minimize the future loss of information. One of the first studies to quantify reference rot was by [7], who monitored the status of a random set of URIs over four years. His results showed that approximately 67% of URIs became inaccessible after a four-year period.

The increased use of URLs and URIs in online publications has in part fueled the increase in

reference rot over time. According to a study [8], Electronic Theses and Dissertations (ETDs) that include URL references have increased over the past 14 years from 23% in 1999 to 80% in 2012. In a study of the persistence of web resources in the arXiv repository and the University of North Texas (UNT) Digital library, [9] found that 45% of the URLs referenced from arXiv still exist, but are not preserved for future generations, and 28% of resources referenced by UNT papers have been lost. In a 2014 paper, [10] investigated how reference rot impacted the ability to revisit the context of scholarly articles after they had been published. The authors extracted the URIs referenced in a collection of 3.5 million scholarly articles from Science, Technology, and Medicine (STM) fields, and observed one out of five articles suffered from reference rot, meaning it is impossible to revisit the web context that surrounds them some time after their publication. In 2021, researchers at Harvard Law School examined New York Times articles from 1996 to 2019. They found that link rot had increased linearly over time, and that out of over 2 million hyperlinks, 25% were inaccessible and that over 13% of links that were still reachable had suffered content drift [11].

Other researchers have employed web archives to study the nature of reference rot. A study of reference in the UK web archives of the British Library found that, of over 1,000 archived URIs, 40% were gone from the live web after two years, while another (40%) were had been affected by content drift [12]. In 2012, [2] developed a Naïve Bayes classifier that could detect soft 404 pages with a precision of 99% and a recall of 92%.

In [4], the authors compiled three different Archive-It collections and experimented with several methods of detecting these off-topic webpages and with how to define thresholds that separate the on-topic from the off-topic pages. This involved comparing the text (after pre-processing, stemming and stopword removal) of the archived website when it was first captured ($URI - R@t_0$) with the text archived website that was captured at a later time ($URI - R@t$). The authors tested a variety of methods and found that the cosine similarity method proved the best at detecting off-topic web pages, with an average accuracy of 0.983, and F-measure (harmonic mean of precision and recall) of 0.881, and an Area Under the Curve (AUC) measure of 0.961. The second-best performing measure was word count. The author also experimented with combining several similarity measures in an attempt to increase performance. The combination of the cosine similarity and word count methods yielded the best results, with an accuracy equal to 0.987, $F = 0.906$, and $AUC = 0.968$ [4].

[1] examined content drift in the same collection used by [10]. They first extracted the URIs referenced in the collection of articles, then obtained the archived version (snapshots) of these URIs whenever available. The text from these archived websites was extracted, and textual similarity measures were used to compare their content to their live web counterparts. The authors found that representative snapshots exist for about 30% of all URI references, and that for over 75% of references the content had drifted away from what it was when referenced. A high degree of both link rot and content drift was detected in the scholarly collection.

3. Methodology

As was seen in Section 2, researchers have used web archives to study the notion of reference rot on the web. Past work on detecting content drift in web archives has focused on comparing

Table 1
Details of the collections

Collection	No. seeds	No. captures	No. of judged captures	% Content drift
INM	73	863	784	9.6%
FMW	37	618	618	33.2%
WCA	86	95	94	11.6%
Total	196	1576	1496	25.1%

the extracted text of some archived websites to the extracted text of other archived websites. This necessitates access to the WARC files that contain the archived websites in the first place.

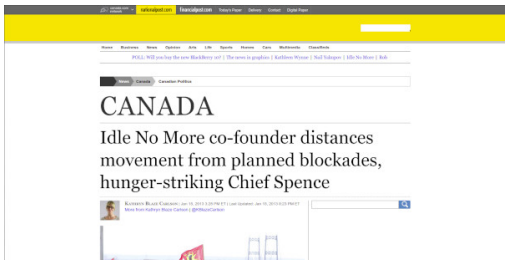
WARC is a container file format that can store a very large number of data objects, including audio and video resources, inside of a single, compressed file. In web archiving, WARC files are used to store content that has been harvested from the web via web crawlers [5]. Most WARC files are quite large, ranging from a few gigabytes to many terabytes in size, which make them cumbersome and slow to traverse and analyze. Extracting content from WARC files requires pre-processing steps for extracting the text, stop-word removal, and stemming. As noted in [1], extracting text from HTML and PDF files proved substantially time-consuming and arduous, and necessitated the writing of custom code even beyond the pre-existing code that is already available to do so. As [13] states, analyzing WARC files often means working with 100s of terabytes of data. This necessitates large amounts of storage space, memory, and a robust research infrastructure, which many cultural heritage institutions do not have access to.

3.1. The dataset

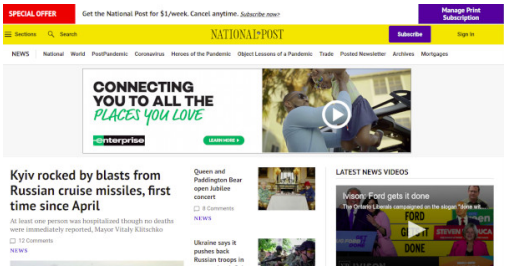
We used the following three web archive collections to gather test data. These were created and maintained by the University of Alberta using the Archive-It subscription service. These collections were created by the University of Alberta Libraries in an effort to preserve western Canadian cultural heritage on the web [14].

1. Idle No More (INM): websites related to "Idle No More", a Canadian political movement encompassing environmental concerns and the rights of indigenous communities [15].
2. Fort McMurray Wildfire 2016 (FMW): websites related to the Fort McMurray Wildfire of 2016 in the province of Alberta, Canada [16].
3. Western Canadian Arts (WCA): born-digital resources created by filmmakers in Western Canada [17].

The collections Idle No More and Fort McMurray Wildfires consist mostly of news articles and social media posts primarily from Twitter. As a result, we expected both of these collections to suffer significantly more from content drift due to frequent webpage changes. Since the WCA collection includes the personal websites of many artists, we expected this particular collection to have much less content drift.



Archived website from January 29, 2013



Live website as of June 5, 2022

Figure 1: Screenshots of a URL of news article from INM collection. The archived website originally contained a National Post article about the movement. The live website now redirects to the homepage of the newspaper, showing content drift has occurred.

3.2. Evaluation of content drift

In order to determine the amount of content drift in our web archive collections, all three authors manually inspected each of the live websites and compared them to their archived versions. Each capture for each website was evaluated by inspecting the content, as well its look and feel. A website was labeled "off-topic" if it had been affected by content drift and "on-topic", otherwise. Most captures were judged, except in a few cases where the archived version was of very poor quality, and evaluators were not able to determine if content drift had taken place.

Figure 1 shows an example of content drift from the INM collection. The archived website originally contained a news article about the movement, but the live website now redirects to the homepage of the newspaper, showing content drift has occurred. Were it not for an archived copy of the page, the article would have been lost forever. The details for each collection are given in Table 1.

Overall, about a quarter of the data set has been affected by content drift, with the FMW collection experiencing 33.2% content drift. The INM collection had undergone surprisingly little content drift (9.6%). The judgements as to whether a website was on-topic or off-topic were used as the ground truth for our next steps.

3.3. Extracting titles

In the title extraction process, the Python Library “Beautiful Soup”[1] was initially applied. However, Beautiful Soup encountered some errors with Twitter URLs. To solve this problem, the Selenium Webdriver [2] was used as a backup plan. Due to the fact that extracting titles with Selenium is much more time-consuming than with Beautiful Soup, this method is only triggered when errors occurred. There were many Twitter URLs that redirected to different pages. To avoid getting the wrong title, the program waits five seconds after the URL is loaded to extract the title.

[1]https://www.crummy.com/software/BeautifulSoup/
[2]https://www.selenium.dev/documentation/webdriver/

Table 2
Confusion Matrix

Collection	TP	**FN**	FP	TN	Total
INM	585	**24**	124	51	784
FMW	412	**68**	1	137	618
WCA	80	**1**	4	9	94
Overall	1077	**93**	129	197	1496

3.4. Using similarity measures

After retrieving the titles, we removed stop words and converted each title to lowercase. We then used a well-known textual similarity metric to compare the titles of the archived websites to those of the live website: the cosine similarity. We based our choice of cosine similarity partly because of its previous, sucessful usage of it in [4] and [1]. Cosine similarity is one commonly-used metric that is not sensitive to high-frequency words. Cosine similarity measures the angle between two vectors using the formula $k(x, y) = \frac{\mathbf{x}^*\mathbf{y}}{\| \mathbf{x} \| * \| \mathbf{y} \|}$. The values calculated by cosine similarity range between 0, for vectors that do not share any terms, to 1, for vectors that are identical, to -1, for vectors that point in opposite directions [18].

Our assumption was that the higher the difference between the title of an archived website and that of its live counterpart, the more likely content drift had taken place (off-topic). However, we needed a threshold value to classify each URL pair as "on-topic" or "off-topic" We initially experimented with higher threshold values of 0.7 and 0.8 after the work of [4]. However, we found that, over time, some websites had lengthened or shortened their titles slightly, but had remained on-topic, thus resulting in some false positives. We eventually decided on a threshold value of 0.6, which gave us the best performance for cosine similarity. Running the code to both extract the website titles and perform the similarity calculations took several hours for each collection, and the resulting text files were 232 KB (INM), 180 KB (FMW), and 20 KB (WCA) in size. This was a much smaller footprint than the much larger and slower WARC files.

4. Results and Discussion

In this section, we present the results of the similarity calculations between the titles of the archived websites and those of their live counterparts. Table 2 presents the confusion matrix values for the cosine similarity calculations. The following metrics are provided:

1. True Positives (TP): URLs that were on-topic and judged to be on-topic
2. False Negatives (FN): URLs that were off-topic but judged to be on-topic
3. False Positive (FP): URLs that were on-topic but judged to be off-topic
4. True Negatives (TN): URLs that were off-topic and judged to be off-topic

Since our intent is to be able to detect off-topic URLs, we were particularly interested in keeping the number of false negatives (FNs) as low as possible.

Table 3
Evaluation results for the collections

Collection	Accuracy	Precision	**Recall**	F-measure
INM	81.1	82.5	**96.1**	88.8
FMW	88.8	99.8	**85.8**	92.3
WCA	94.7	95.2	**98.8**	97
Overall	85.2	89.3	**92.1**	90.7

Table 3 presents the evaluation metrics and results for the cosine similarity calculations, as compared to human judgements of content drift. Accuracy is the fraction of both on-topic and off-topic URLs that were correctly classified, or $\mathrm{Accuracy} = \frac{(\mathrm{TP+TN})}{(\mathrm{TP+FP+FN+TN})}$. Precision is the fraction of retrieved URLs that were correctly classified as being on-topic, defined as $\mathrm{Precision} = \frac{(\mathrm{TP})}{(\mathrm{TP+FP})}$, while recall is the fraction of on-topic URLs that were retrieved, defined as $\mathrm{Recall} = \frac{(\mathrm{TP})}{(\mathrm{TP+FN})}$. The F-measure is the harmonic mean of precision and recall, or $F-measure = \frac{2\mathrm{TP}}{(2\mathrm{TP+FP+FN})}$.

Overall, good performance was achieved, with high or medium-high values of accuracy, precision, recall, and F-measure. Because the recall is defined as the fraction of off-topic websites that were detected, we were particularly interested in achieving high levels of it.

5. Conclusion

In this paper, we presented a simple method for detecting content drift on the live web based on comparing the titles of live websites to those of their archived versions. Our assumption was that the higher the difference between the title of an archived website and that of its live counterpart, the more likely content drift had taken place. Our proposed method achieved high values of accuracy, precision, recall, and F-measure, and has the following advantages:

- It is highly consistent with human judgements of content drift.
- It is quicker and less computationally intensive than other methods which require the extraction and comparison of the full text of archived websites.
- It does not require access to the WARC files which contain the archived websites, which are large and require much storage space.

Having simple methods such as the one presented in this paper can allow institutions or researchers to quickly and effectively detect content drift without needing many technological resources. In the future, we wish to apply this method for detecting content drift to larger web archives, and seek to refine and improve its performance without sacrificing its speed and simplicity.

Acknowledgments

Thanks to Shawn M. Jones and Michael L. Nelson for the some of the ideas that inspired this

work. The research in this paper was supported in part by funding from the Social Sciences and Humanities Research Council of Canada.

References

[1] S. M. Jones, H. Van de Sompel, H. Shankar, M. Klein, R. Tobin, C. Grover, Scholarly context adrift: Three out of four uri references lead to changed content, PLOS ONE 11 (2016) 1–32. URL: https://doi.org/10.1371/journal.pone.0167475. doi:10.1371/journal.pone.0167475.

[2] L. Meneses, R. Furuta, F. Shipman, Identifying "soft 404" error pages: Analyzing the lexical signatures of documents in distributed collections, in: P. Zaphiris, G. Buchanan, E. Rasmussen, F. Loizides (Eds.), Theory and Practice of Digital Libraries, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 197–208.

[3] Archive-It, Learn more, 2020. URL: https://archive-it.org/learn-more.

[4] Y. Alnoamany, M. C. Weigle, M. L. Nelson, Detecting off-topic pages within timemaps in web archives, International Journal on Digital Libraries 17 (2016) 203–221.

[5] International Internet Preservation Consortium, The warc format 1.1, n.d. URL: https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/.

[6] S. Lawrence, D. Pennock, G. Flake, R. Krovetz, F. Coetzee, E. Glover, F. Nielsen, A. Kruger, C. Giles, Persistence of web references in scientific research, Computer 34 (2001) 26–31. doi:10.1109/2.901164.

[7] W. Koehler, Web page change and persistence—a four-year longitudinal study, Journal of the American Society for Information Science and Technology 53 (2002) 162–171. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.10018. doi:10.1002/asi.10018. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.10018.

[8] M. Phillips, D. Alemneh, B. Reyes Ayala, Analysis of url references in etds: a case study at the university of north texas, Library Management 35 (2014).

[9] R. Sanderson, M. E. Phillips, H. Van de Sompel, Analyzing the persistence of referenced web resources with memento, Austin, TX, USA, 2011. URL: http://digital.library.unt.edu/ark:/67531/metadc39318/.

[10] M. Klein, H. Van de Sompel, R. Sanderson, H. Shankar, L. Balakireva, K. Zhou, R. Tobin, Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot, PLoS ONE 9 (2014) e115253+. doi:10.1371/journal.pone.0115253.

[11] J. Zittrain, J. Bowers, C. Stanton, The Paper of Record Meets an Ephemeral Web: An Examination of Linkrot and Content Drift within The New York Times, Research Report, Berkman Klein Center for Internet & Society at Harvard University, 2021. doi:http://dx.doi.org/10.2139/ssrn.3833133.

[12] A. N. Jackson, Ten years of the uk web archive: what have we saved?, 2015. URL: https://anjackson.net/2015/04/27/what-have-we-saved-iipc-ga-2015/.

[13] I. Milligan, Demystifying the warc: Research use of web archives, 2022. URL: https://archive.org/details/demystifying-the-warc-research-use-of-web-archives-slides.

[14] University of Alberta Library, Digital preservation services, n.d. URL: https://www.library.ualberta.ca/digital-initiatives/preservation.

[15] University of Alberta, Idle No More collection, n.d. URL: https://archive-it.org/collections/3490.
[16] University of Alberta, Fort McMurray wildfire 2016 collection, 2016. URL: https://archive-it.org/collections/7368.
[17] University of Alberta, Western Canadian Arts collection, n.d. URL: https://archive-it.org/collections/6296.
[18] D. Jurafsky, J. H. Martin, Speech and Language Processing, 2nd ed., Prentice Hall, Upper Saddle River, NJ, 2008.