Detecting content drift on the web using web archives and textual similarity

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

[1]School of Library and Information Studies, University of Alberta
Edmonton, Alberta, Canada
brenda dot reyes at ualberta dot ca

September 20, 2022

Linked Archives 2022: International Workshop on Archives and Linked Data
TPDL2022: 26th International Conference on Theory and Practice of Digital Libraries, Padua, Italy

Overview I

1. Introduction

2. Previous Work

3. Methodology

4. Results and Discussion

5. Conclusion

References

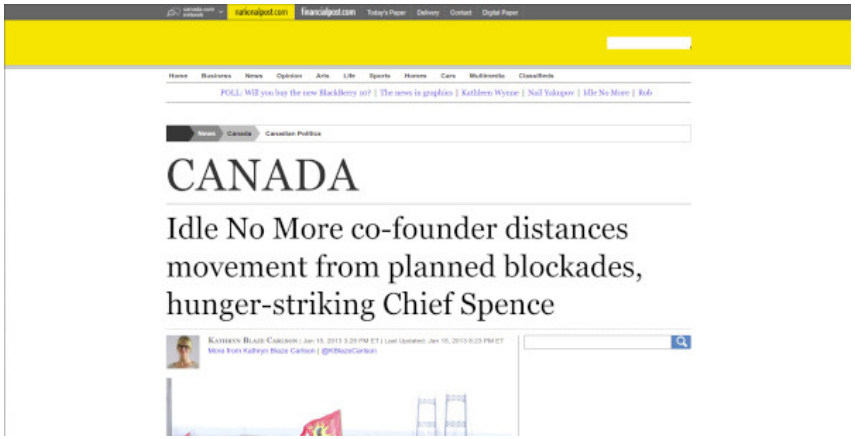
Reference rot

Has two components:

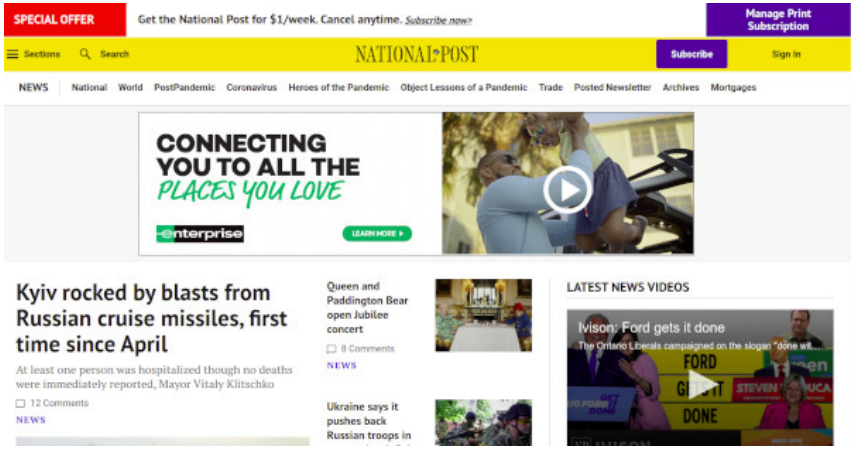
1. Link rot: The resource identified by a URI vanishes from the web. As a result, a URI reference to the resource ceases to provide access to referenced content.
2. Content drift: The resource identified by a URI changes over time. The resource's content evolves and can change to such an extent that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

(Jones et al., 2016)

Example of content drift



Archived website from January 29, 2013



Live website as of June 5, 2022

Figure: Screenshots of a URL of news article. 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.

Detecting content drift is difficult

- Just detecting 404 error codes is not enough
- *Soft 404s* occur 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 (Meneses, Furuta, & Shipman, 2012)

Purpose and Research Question

Purpose

to find an approach to detecting content drift that simulates a human evaluator inspecting a website and determining if content drift has occurred

Research Question

Is a large change in a website's title indicative of content drift?

Previous work on reference rot

Reference rot, particularly link rot, has been well-documented over time:

- Approximately 67% of URIs became inaccessible after a four-year period (Koehler, 2002)
- In the arXiv repository, 45% of the URLs referenced still exist, but are not preserved (Sanderson, Phillips, & Van de Sompel, 2011)
- Link rot in Electronic Theses and Dissertations (ETDs) increased from 23% in 1999 to 80% in 2012 (Phillips, Alemneh, & Reyes Ayala, 2014)
- In a collection of 3.5 million scholarly articles, one out of five articles suffered from link rot (Klein et al., 2014)
- In New York Times articles (1996-2019), out of over 2 million hyperlinks, 25% were gone and over 13% of links that were still reachable had suffered content drift (Zittrain, Bowers, & Stanton, 2021)

Previous work on content drift and web archives

(Jones et al., 2016) examined content drift in the same collection used by (Klein et al., 2014)

1. Extracted URIs referenced in the collection
2. Obtained the archived versions (stored in WARC files) of these URIs whenever available
3. Extracted the text of the archived websites
4. Used textual similarity measures to compare their content to their live web counterparts and detect content drift

Researchers were able to detect content drift very accurately. For over 75% of references the content had drifted away from what it was when referenced

But detecting content drift with WARC files has drawbacks...

- WARC files are very large (often many GBs and sometimes TBs in size)
- Not everyone has access to them; access cannot always be granted depending on the country
- Pre-processing steps for extracting the text, stop-word removal, and stemming are needed
- Computationally intensive and time consuming. Large amounts of storage space, memory, and a robust research infrastructure

What to do?

The dataset

These collections were created by the University of Alberta Libraries in an effort to preserve western Canadian cultural heritage on the web (University of Alberta Library, n.d).

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

Process

1. Manually inspect each of the live websites and compare them to their archived versions
2. Classify each website as "off-topic" if it has been affected by content drift and "on-topic", if otherwise. This was our "ground truth" dataset
3. Extract the title of each live website and compare to the title of each archived website using textual similarity measures
4. Use a threshold value to determine the cut-off point for similarity scores
5. Compare our results to those produced by human evaluators

Manual evaluation of content drift

Table: Content drift in the collections, as judged by human evaluators

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

Extracting titles and using textual similarity measures

1. Used Python library *Beautiful Soup* and Selenium Webdriver
2. Removed stop words and converted each title to lowercase
3. Used cosine similarity to compare the titles of the archived websites to those of the live website
4. Decided on a threshold value of 0.6, which gave us the best performance

Threshold values

If cosine similarity $> 0.6 \rightarrow$ on-topic (no content drift has occurred)
If cosine similarity $< 0.6 \rightarrow$ off-topic (content drift has occurred)

Results

Table: 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

Discussion

- Overall, good performance was achieved, with high or medium-high values of accuracy, precision, recall, and F-measure
- High recall levels indicate most off-topic websites were detected
- Run-time was short, particularly for the smaller collections (FMW and WCA)

Running times

As seen on a system with 3GB RAM and two CPU cores running Ubuntu and Python 3.6

Table: Running times for each collection

Collection	Running time (min)
INM	356
FMW	45
WCA	8

Contributions

- It is highly consistent with human judgments 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

Conclusions and Future Work

- Simple methods can allow institutions or researchers to quickly and effectively detect content drift without needing many technological resources
- Institutions who engage in web archiving could use this method to quickly and effectively detect if content drift has occurred, and decide whether or not they wish to keep archiving that specific site, thus saving money and 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

Thanks and 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 I

Jones, S. M., Van de Sompel, H., Shankar, H., Klein, M., Tobin, R., & Grover, C. (2016, 12). Scholarly context adrift: Three out of four uri references lead to changed content. *PLOS ONE*, *11*(12), 1-32. Retrieved from https://doi.org/10.1371/journal.pone.0167475 doi: 10.1371/journal.pone.0167475

Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., & Tobin, R. (2014, December 26). Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLoS ONE*, *9*(12), e115253+. doi: 10.1371/journal.pone.0115253

References II

Koehler, W. (2002). Web page change and persistencea four-year longitudinal study. *Journal of the American Society for Information Science and Technology*, *53*(2), 162-171. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.10018 doi: 10.1002/asi.10018

Meneses, L., Furuta, R., & Shipman, F. (2012). 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* (pp. 197–208). Berlin, Heidelberg: Springer Berlin Heidelberg.

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

References III

Sanderson, R., Phillips, M. E., & Van de Sompel, H. (2011, June). Analyzing the persistence of referenced web resources with memento. Austin, TX, USA. Retrieved from http://digital.library.unt.edu/ark:/67531/metadc39318/

University of Alberta. (2016, March). *Fort McMurray wildfire 2016 collection.* Retrieved from https://archive-it.org/collections/7368

University of Alberta. (n.da, March). *Idle No More collection.* Retrieved from https://archive-it.org/collections/3490

University of Alberta. (n.db, March). *Western Canadian Arts collection.* Retrieved from https://archive-it.org/collections/6296

University of Alberta Library. (n.d). *Digital preservation services.* Retrieved from https://www.library.ualberta.ca/digital-initiatives/preservation

References IV

Zittrain, J., Bowers, J., & Stanton, C. (2021, apr). *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. doi: http://dx.doi.org/10.2139/ssrn.3833133