Detecting quality problems in archived websites using image similarity

Brenda Reyes Ayala, James Sun, Jennifer McDevitt, Xiaohui Liu[1]

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

June 15, 2021

Web Archiving Conference (WAC 2021)

Overview I

Why should we care about similarity?

The visual representation of information on the web has become an important part of judging the quality of archived web pages [2].

Definition of visual correspondence in a web archive

the similarity in appearance between the original website and the archived website [4]

Research Questions and Purpose

How effective are different similarity measures at measuring the visual correspondence between an archived website and its live counterpart?

We examine how the visual correspondence of an archived website can be measured using popular image similarity measures. Using these measures we evaluate how visual correspondence can be used as an indication of overall archive quality.

Steps for calculating the visual similarity

Created set of tools called "wa screenshot compare", currently freely available as a Github repository at `https://github.com/reyesayala/wa_screenshot_compare`

1. Create the configuration file
2. Create list of live URLs
3. Create list of archived URLs
4. Take screenshots of live URLs (if available)
5. Take screenshots of archived URLs
6. Pair each live screenshot with its archived counterpart
7. Calculate similarities

Create the configuration file

In `screenshot.ini`, specify:

- The seedlist of your collection
- Collection id and name
- Method for taking screenshots
- Names of output files and directories
- Similarity measures to use

Take screenshots of URLs

Four ways of taking screenshots:

1. Pyppeteer (Python port of Puppeteer)
2. Chrome
3. Cutycapt
4. Selenium

For archived websites, closes the Archive-It banner, which has proven to have a significant effect on similarity calculations

Calculating similarity

Based on popular image similarity measures: Structural Similarity Index (SSIM), Mean Squared Error (MSE), and “vector distance”, which produces the distance between the RGB values of each screenshot. We changed this metric slightly by subtracting every result from 100, thus giving us the percentage similarity between a pair of images.

- SSIM: calculates similarity on a scale of [-1,1]. 1 is perfect similarity.
- MSE: calculates similarity on a scale of [0, ∞]. 0 is perfect similarity.
- Vector distance: calculates similarity on a scale [0,1]. 1 is perfect similarity.

The dataset used

We used three Archive-It collections on the topic of Western Canadian cultural heritage, created and maintained by the University of Alberta Libraries:

1. Idle No More [7]: websites related to “Idle No More”, a Canadian political movement encompassing environmental concerns and the rights of indigenous communities.
2. Western Canadian Arts [8]: born-digital resources created by filmmakers in Western Canada.
3. Fort McMurray Wildfire 2016 [6]: websites related to the Fort McMurray Wildfire of 2016 in the provice of Alberta, Canada.

Characteristics of Web Archive Collections Used for Similarity Judgments

We categorized as "lost", those websites that returned an HTTP status code other than 200 and were not redirects [5]

Collection	No. Seeds	No. Seeds Available 2019 (%)	No. Seeds Available 2021(%)
Idle No More	196	182 (**92.9**)	105 (**53.6**)
Western Canadian Arts	101	95 (94.1)	87 (86.1)
Fort McMurray Wildfire 2016	52	NA	27 (51.9)

Web archive collections and their preservation status

Furthermore, we deployed MemGator[1], a Memento Aggregator, to determine if copies of the websites were also present in web archives around the world [5].

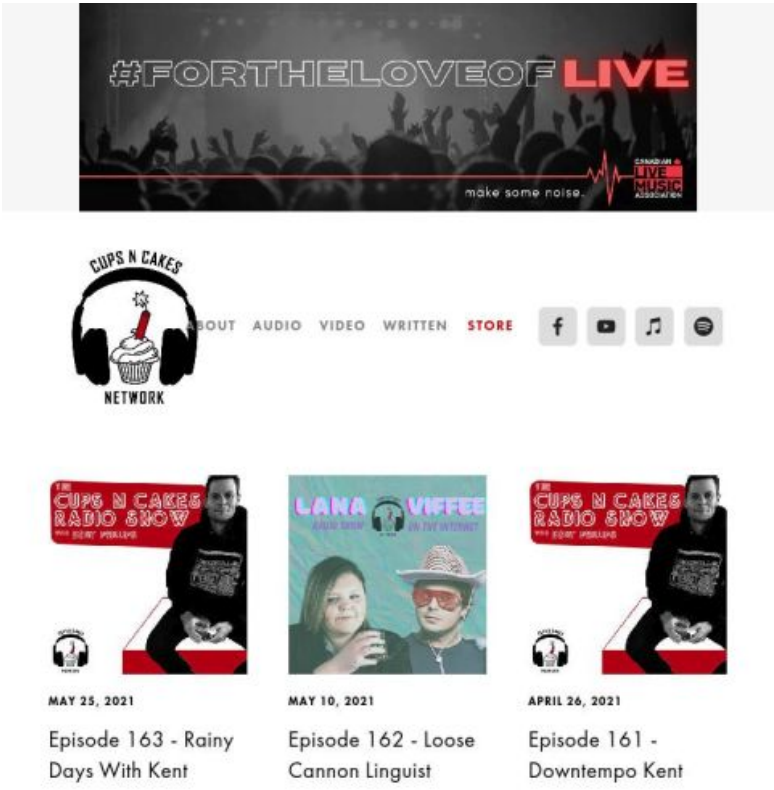
Collection	No. of AIT captures	No. of IA captures	Total No. of captures
Idle No More	7005 (32.3%)	14049 (64.8%)	21664
Western Canadian Arts	289 (8.6%)	2930 (87.7%)	3342
Fort McMurray Wildfire 2016	4693 (34.3%)	8706 (63.7%)	13677

Issues encountered during the screenshot process

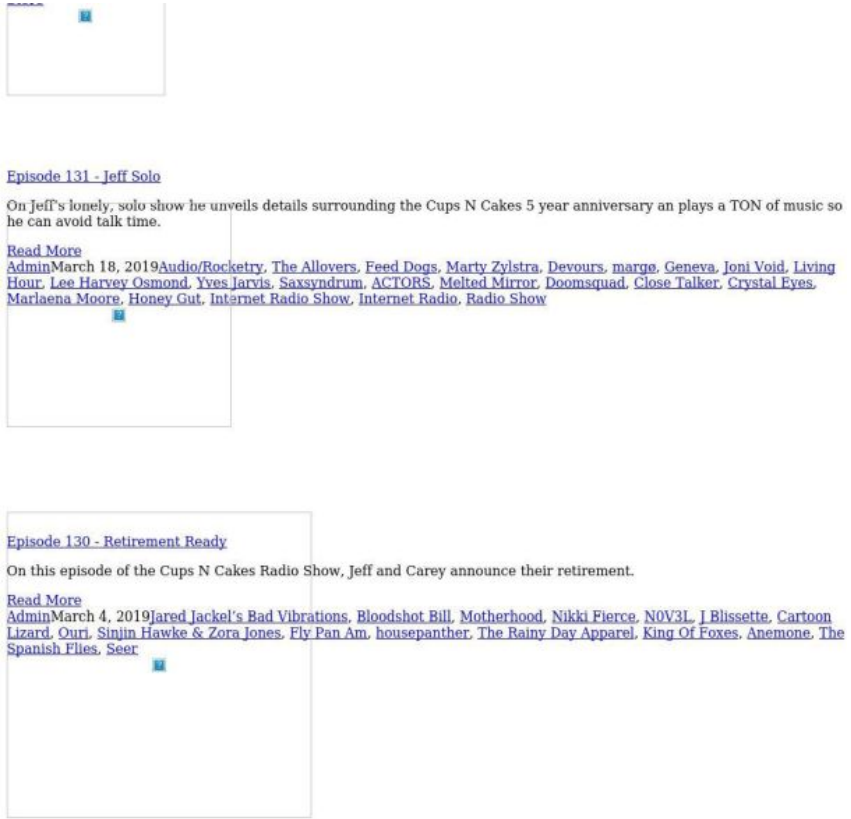
Not a trivial process despite our initial assumptions:

- Appended the text “id_” to the url of the archived websites, but this approach often breaks the CSS styling of the archived site. Solution: use “if_” instead.
- Can be slow, even if different methods are used.
- Webmasters might question why you are visiting their website so often. Time to develop your PR skills!

Example: An archived website with quality problems



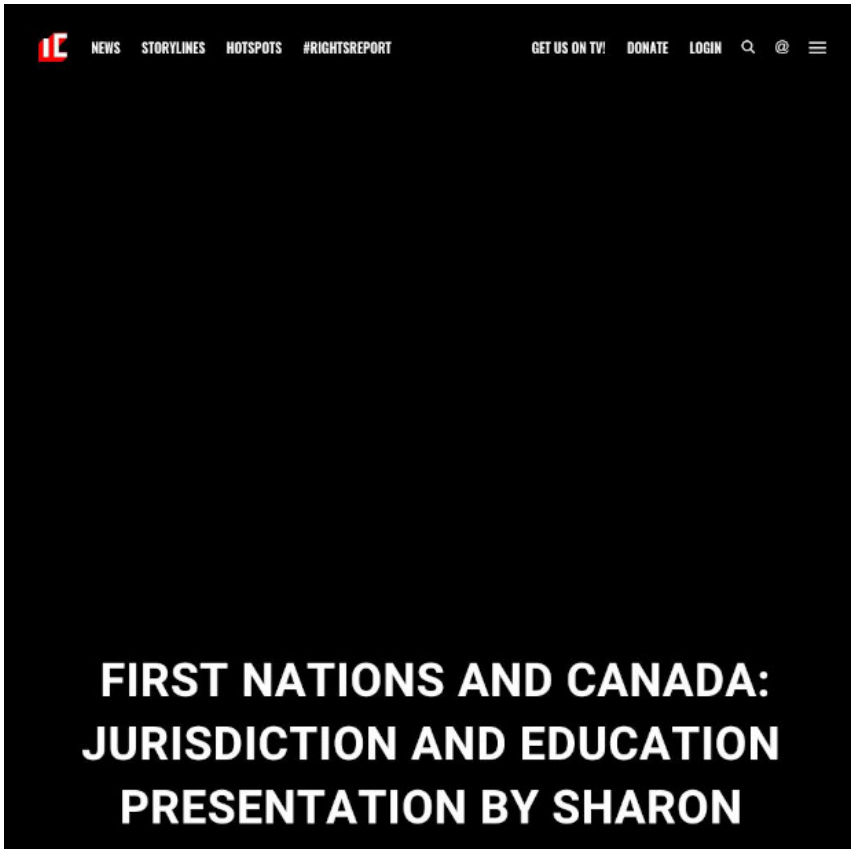
Current website



Archived website

Figure: Comparison of images for the website "Cups n Cakes". SSIM = 0.41, MSE = 73169.53, Vector Distance = 56.94

Example: An archived website with quality problems (2)



Current website

Archived website

Figure: Comparison of images for the website "Intercontinental Cry". SSIM = 0.07, MSE = 163673.63, Vector Distance = 13.66

Example: A "high" quality archived website

Current website

Archived website

Figure: Comparison of images for the website "Darryl Nepinak". SSIM = 0.97, MSE = 972.53 , Vector Distance = 98.8

Correlation between similarity measures in web archives

Performed a correlation analysis on all our similarity scores (3003 scores) for the three web archives collections to determine if there were relationships between different similarity measures.

SSIM - MSE	MSE - Vector	SSIM - Vector
-0.39	-0.98	0.38

- Weak negative correlation between SSIM and MSE score.
- Very strong negative correlation between MSE and Vector distance.
- Weak positive correlation between SSIM and vector distance scores.

Correlation between similarity measures in web archives (2)

1. Our initial hypothesis that there would be a strong correlation between SSIM and vector scores was not proven.
2. Almost perfect negative relations between MSE and vector distance suggests that one measure might be easily substituted for another.
3. Because we found MSE scores relatively difficult to interpret, we recommended the use of vector distance or SSIM as measures of similarity.

But vector distance can fail

Qualitative analyses of the screenshots indicated that vector scores could be misleading. Vector distances sometimes produced a misleadingly high score when the quality is actually bad.

Current website

Archived website

Figure: Comparison of images for the website "National Observer". SSIM = Not a Number (NaN), MSE = 9670.91, Vector Distance = 93.44

Other reasons for low similarity scores

Similarity scores can also be low for other reasons:

1. Content drift: leads to "soft 404s" [3]
2. Website redesign

What we learned

It is important to conduct visual quality assessments early in the web archiving process, while the websites collected are still online and accessible for comparison.

Conclusions

- It is possible to apply image similarity metrics in order to measure the visual correspondence (and thus visual quality) of archived websites.
- Not all image similarity metrics are created equal, and some can produce misleading results.
- Low SSIM scores can be indicative of possible QA problems.

Next steps

- How well do similarity measures match up with human judgments of visual correspondence in a web archive?
- Can similarity measures adequately detect the severity of QA problems? Ex: Missing a style sheet vs. an entirely blank webpage.

Acknowledgements

This project is supported in part by funding from the Social Sciences and Humanities Research Council of Canada

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

[1] Alam, S. & Nelson, M.L., (2016). MemGator - A portable concurrent memento aggregator: Cross-platform CLI and server binaries in Go. *Proceedings from the 2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, 243-244. doi: 10.1145/2910896.2925452

[2] Gyllstrom, K., Eickhoff, C., de Vries, A.P. & Moens, M. (2012). The downside of markup: Examining the harmful effects of CSS and Javascript on indexing today's web. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 1990-1994. doi: 10.1145/2396761.2398558

[3] Meneses, L., Furuta, R., & Shipman, F. (2012). Identifying 'Soft 404' error pages: Analyzing the lexical signatures of documents in distributed collections. In Zaphiris, P., Buchanan, G., Rasmussen, E., & F. Loizides (Eds.), *Theory and Practice of Digital Libraries* (pp. 197-208). Berlin, Germany: Springer Berlin Heidelberg.

[4] Reyes Ayala, B. (2020). Correspondence as the primary measure of quality for web archives: A grounded theory study. In M. Hall, T. Merčun, & T. Risse (Eds.), *Digital Libraries for Open Knowledge* (pp. 73-86). Cham, Switzerland: Springer International Publishing.

[5] Saiyera, T., Reyes Ayala, B., & Du, Q. (2021, June). Assessing the loss of Western Canadian digital heritage. *Proceedings of the Annual Conference of CAIS / Actes du congrÃ¨s annuel de l'ACSI*. Edmonton, Alberta, Canada: University of Alberta Libraries. doi: 10.29173/cais1218

[6] University of Alberta. (2016). Fort McMurray Wildfire 2016 collection. Retrieved from https://archive-it.org/collections/7368

[7] University of Alberta. (n.d). Idle No More collection. Retrieved from https://archive-it.org/collections/3490

[8] University of Alberta. (n.d). Western Canadian Arts collection. Retrieved from https://archive-it.org/collections/6296