<Author Name>

<Institution>, <City, Prov, Country>

Quality matters: A new approach for detecting quality problems in web archives

Abstract or Résumé:

Since the practice of web archiving, or the act of preserving websites as historical, legal, and informational records, become more commonplace in the 2000s, web archives have become valuable sources for historical research. Unfortunately, many archived websites are of low quality and are missing crucial elements. In this paper, we examine the issue of quality and focus on visual correspondence, the similarity in appearance between the original website and its archived counterpart. We examine how the visual correspondence of an archived website can be measured using image similarity measures. Our results indicate that the Structural Similarity Index metric (SSIM) was able to successfully measure visual correspondence. If applied to the Quality Assurance process of an institution, this similarity metric could help web archivists quickly detect quality problems in their web archives, and fix them in order to create high-quality web archives.

1. Introduction

In 1996, the Internet Archive began using a web crawler to periodically take snapshots of websites and store them as historical records. Internet users could then access these archived websites using the Wayback Machine, a special piece of software developed by the Internet Archive. As the World Wide Web evolved, the pace at which websites changed their content and appearance accelerated dramatically. Often the Internet Archive's cache was the only record of how a website had evolved or that it had existed at all. In a few years, the practice of web archiving, as it became known, had spread beyond the Internet Archive, as organizations such as national libraries, government organizations, and universities began also to archive websites, for the purpose of preserving their digital heritage.

A high-quality archived website should be an accurate representation of the original website in content, form, and appearance. It should look and behave exactly like the original, but in practice this is rarely the case. It is common to see archived websites with no images or media or with broken links. In order to detect these quality problems, web archivists must engage in an onerous process of quality assurance (QA) where they manually inspect hundreds or even thousands of archived websites (Reyes Ayala, Phillips, and Ko, 2014). When web archiving is done by national libraries that seek to capture and preserve their national domain, quality problems grow to such a scale that human intervention is no longer enough to detect and fix them. Unless we learn how to address quality problems in a web archive, we will soon be facing an incomplete digital historical record, at a time when the record is crucial.

In the context of web archives, Reyes defines visual correspondence as "the similarity in appearance between the original website and the archived website," (2018). This paper examines how the visual correspondence of an archived website can be measured using popular image similarity measures, originally employed in Computer Science to detect differences between images. Using these measures we evaluate how visual correspondence can be used as an indication of overall archive quality. We are interested in answering the following research questions:

- How effective are different similarity measures at measuring the visual correspondence between an archived website and its live counterpart? Which measure yields the best performance?
- Can an image similarity metric successfully distinguish between high-quality archived websites and lower-quality archived websites?

2. Literature Review

In their research, McNally, Wakaruk, and Davoodi (2015) examined the extensive removal of Canadian government web content and its impact on researchers, who would no longer have access to historical Canadian government web content essential for scrutinizing government policy and activities. They stated that web archiving programs were performing a crucial role in maintaining their role as stewards of government information. The topic of quality in a web archive was first raised by Masanès (2006), who defined quality in a web archive as (1) the completeness of material (linked files) archived within a target perimeter (2) the ability to render the original form of the site, particularly regarding navigation and interaction with the user. In their paper, Denev, Mazeika, Spaniol, and Weikum (2011) introduced the Sharp Archiving of Website Captures (SHARC) framework for data quality in web archiving. examined the importance of missing elements and their impact on the quality of archived websites. Brunelle, Kelly, SalahEldeen, Weigle, and Nelson (2015) examined the importance of missing elements and their impact on the quality of archived websites. In her work, Reyes (2018) developed a multi-dimensional model of information quality in a web archive and proposed ways to quantitatively measure quality problems.

3. Methodology

We chose three different web archives in order to apply the similarity metrics, two from the University of Alberta and one from the British Library's UK Web Archives: The "Idle No More" collection (University of Alberta, 2013), the Western Canadian Arts collection (University of Alberta, 2015), and the UK Web Archives Open Access (OA) collection (Jackson, n.d.). We created a set of tools called "wa screenshot compare", currently freely available as a Github repository (Reyes Ayala, 2019). Written in Python, these tools take URLs as input and generate screenshots of the live websites. "wa screenshot compare" then generates a list of all archived versions of the live sites that are available. Screenshots are then taken of the archived websites. The tool "wa screenshot compare" then puts the two sets of screenshots through a similarity analysis based on two popular image similarity measures: Structural Similarity Index (SSIM) (Wang, Bovik., Sheikh, & Simoncelli 2004) and Mean Squared Error (MSE) (Eskicioglu, Fisher, & Chen, 1995). We chose SSIM and MSE due to their popularity in the image comparison community and their accessibility. We added a third measure, which we call "vector distance" (Rosettacode.org, 2018), that calculates the distance between the RGB values of each screenshot. The greater the distance, the greater the difference between the two images, and thus, the greater the difference between the two websites. We changed this metric slightly by subtracting every result from 100, thus giving us the percentage similarity between a pair of images. The scales for each measure are shown below.

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

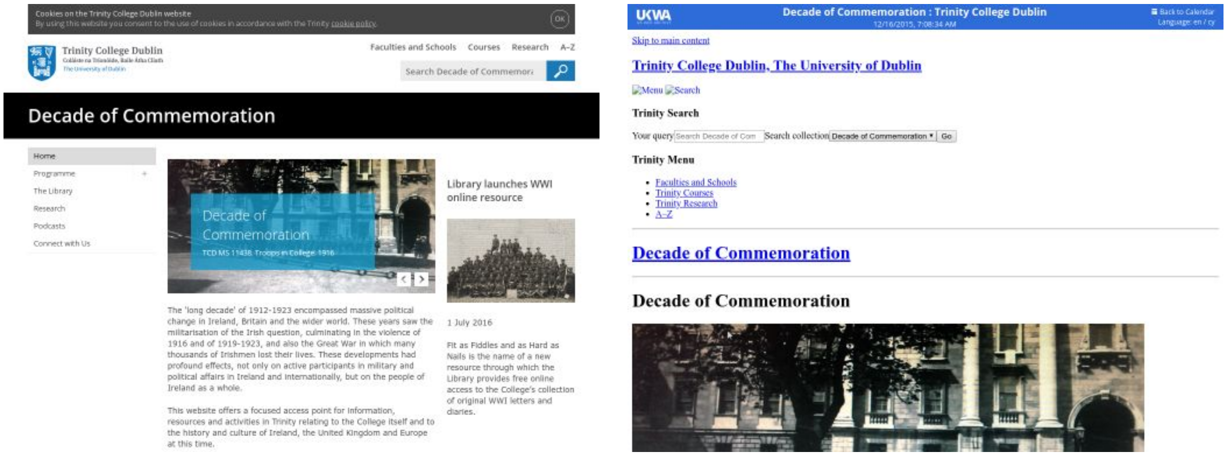
After we calculated the similarity scores for all websites in our collections, the next step was to obtain human judgements of quality and assess how closely these matched the similarity scores generated by our code. We enlisted the help of two University of Alberta librarians with previous web archiving experience and two student RAs. Quality judges met several times to discuss the process and rubric involved, and one of the researchers was responsible for checking the judgements for accuracy. Table 1 contains the rubric used to judge the visual correspondence of the archived screenshots to their current, live counterparts.

Quality Judgement	**Description**
High Quality	Intellectual content, images, and styling are preserved in both screenshots; the screenshots look almost identical. Images can be missing if not integral to webpage content (i.e. ads)
Medium Quality	Intellectual content is preserved without styling and lack of style elements does not impede readability.
Low Quality	Little or no content is preserved.
No Comparison	Script failed to take a screenshot, server connection failed, technical issues occurred, etc. OR link rot has occurred (link not

	connecting for whatever reason. For example, blog has moved, page no longer exists, etc.), or page has changed substantially so it no longer resembles archived screenshot (website redesign, etc.)

Table 1: Rubric used to assess the quality of the screenshots of an archived website.

Figures 1 and 2 illustrate how two websites are archived, with very different results. The archived website for Trinity College in Dublin, seen in Figure 1, was classified as one of "medium quality." The intellectual content of the website has been preserved, but its styling has been lost, which is reflected in its similarity scores. Figure 2 presents an example of an archived website that is of low quality; the archived version is simply a blank page and all content has been lost. This is reflected in the very low SSIM and vector distance scores.



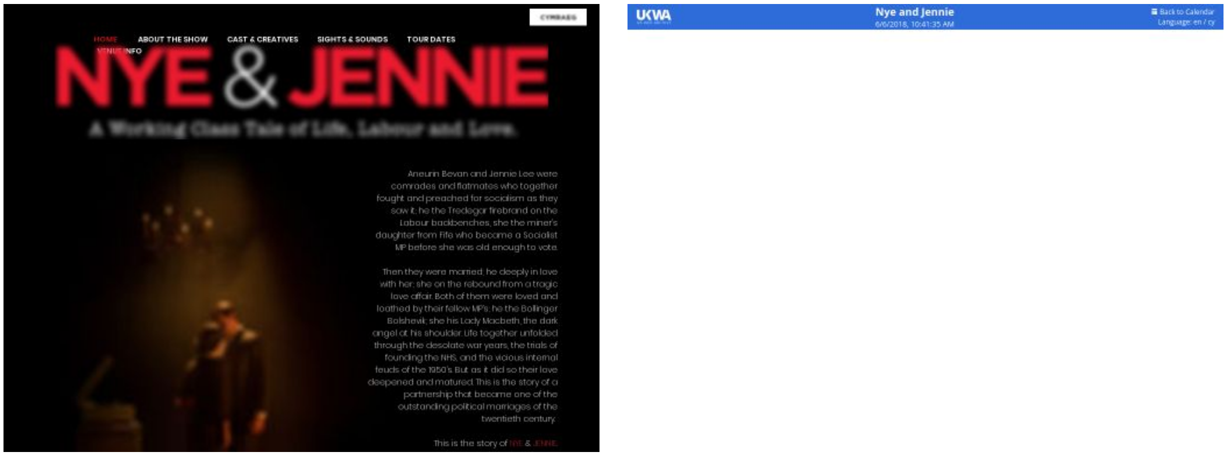
By using this website you consent to the use of cookies in accordance with the Trinity cookie policy.

The 'long decade' of 1912-1923 encompassed massive political change in Ireland, Britain and the wider world. These years saw the militarisation of the Irish question, culminating in the violence of 1916 and of 1919-1923, and also the Great War in which many thousands of Irishmen lost their lives. These developments had profound effects, not only on active participants in military and political affairs in Ireland and internationally, but on the people of Ireland as a whole.

This website offers a focused access point for information, resources and activities in Trinity relating to the College itself and to the history and culture of Ireland, the United Kingdom and Europe at this time.

Fit as Fiddles and as Hard as Nails is the name of a new resource through which the Library provides free online access to the College's collection of original WWI letters and diaries.

Figure 1: Comparison of images for the website "Trinity College Dublin: Decade of Commemoration". SSIM = 0.51, MSE = 61536.53, Vector Distance = 59.87



Aneurin Bevan and Jennie Lee were comrades and flatmates who together fought and preached for socialism as they saw it; he the Tredegar firebrand on the Labour backbenches, she the miner's daughter from Fife who became a Socialist MP before she was old enough to vote.

Then they were married, he deeply in love with her; she on the rebound from a tragic love affair. Both of them were loved and loathed by their fellow MP's; he the Bollinger Bolshevik; she his Lady Macbeth, the dark angel at his shoulder. Life together unfolded through the desolate war years, the trials of founding the NHS, and the vicious internal feuds of the 1950's. But as it did so their love deepened and matured. This is the story of a partnership that became one of the outstanding political marriages of the twentieth century.

This is the story of NYE & JENNIE.

Figure 2: Comparison of images for the website of the play "Nye & Jennie". SSIM = 0.28, MSE = 169603.88, Vector Distance = 8.83

4. Results and Discussion

Throughout our analyses, the MSE measure proved to be the most difficult to interpret, as it has no proper upper bound. For this reason, we discarded MSE from our image similarity metrics. Qualitative inspection of the scores also revealed that the vector distance would sometimes yield inaccurate results, as some screenshots that were completely blank had scores in the 0.50-0.60 range. For these reasons, we decided on SSIM as the best image similarity measure for the task.

After the quality judgements were finished, we removed the screenshots labeled “No Comparison” and examined the remaining similarity scores. We found that neither SSIM, nor vector similarity were able to fully distinguish between low and medium-quality archived websites, therefore we merged these two categories into a single category labeled “Low/Medium Quality.” Afterwards we created a balanced random sample of 100 screenshots, with 50 images deemed to be “High Quality” and 50 images that were “Low/Medium Quality.”

A Mann-Whitney U test (chosen because the distribution was not a normal one) was run to determine if there were differences in similarity scores between high-quality archived websites and low-quality archived websites. Median SSIM scores were statistically significantly higher for high-quality websites (0.95) than for websites judged to be medium or low-quality (0.65), $U = 279$, $z = -6.69$, $p < .001$. Therefore we conclude that the SSIM scores for high-quality archived websites are statistically significantly higher than those of low-quality websites.

Our results indicated that a) the Structural Similarity Index metric was most able to successfully measure visual correspondence between an archived website and its live counterpart and b) that it was able to successfully distinguish between website captures of poor quality and those of higher quality. If applied to the QA process of an institution, this similarity metric could help web archivists quickly detect quality problems in their web archives, and be able to more successfully focus their efforts. This paper is connected to the CAIS themes of “diverging methodologies in information science” and “contested grounds in data collection, data interpretation and study findings”, as it presents an approach to a relatively new research dataset: web archives, and examines them using both quantitative and qualitative methods. Attendees will learn about the issues of quality and completeness in web archives, and how these can be addressed. This grant-funded research is only the first step in developing a comprehensive toolkit for automated or semi-automated quality assurance processes in web archives, which will in turn help web archivists create better web archives in the future.

Reference List

Brunelle, J., Kelly, M., SalahEldeen, H., Weigle, M. C., & Nelson, M. L. (2015). Not all mementos are created equal: measuring the impact of missing resources. *International Journal on Digital Libraries*, 1-19. doi: 10.1007/s00799-015-0150-6

Denev, D., Mazeika, A., Spaniol, M., & Weikum, G. (2011, March). The SHARC framework for data quality in web archiving. *The VLDB Journal 20*(2), 183–207. doi: 10.1007/s00778-011-0219-9

Eskicioglu, A. M., Fisher, P. S., & Chen, S. (1995). Image quality measures and their performance. *IEEE Transactions on Communications, 43*(12), 2959 - 2965.

Hinkle, D.E., Wiersma, W., & Jurs, S.G. (2003). Other nonparametric tests. In *Applied statistics for the behavioral sciences* (5th ed.) (pp. 572-586). Boston, MA: Houghton Mifflin Company.

Internet Archive. (n.d.). Archive-It: Web Archiving Services for Libraries and Archives. https://archive-it.org

Hinkle, D.E., Wiersma, W., & Jurs, S.G. (2003). Other nonparametric tests. In *Applied statistics for the behavioral sciences* (5th ed.) (pp. 572-586). Boston, MA: Houghton Mifflin Company.

Internet Archive. (n.d.). Archive-It: Web Archiving Services for Libraries and Archives. https://archive-it.org

Jackson, A. (n.d.). UKWA Manual QA Dataset. https://github.com/iipc/qa2019/tree/master/ukwa-manual-qa-dataset

Masanès, J. (2006). *Web archiving*. Berlin, Germany: Springer.

McNally, M.B, Wakaruk, A., & Davoodi, D. (2015). Rotten by design: Shortened expiry dates for government of Canada web content. *Proceedings of the Annual Conference of CAIS, Canada*. doi: https://doi.org/10.29173/cais909

Reyes Ayala, B. (2019). Wa screenshot compare. [Computer software]. Retrieved from https://github.com/reyesayala/wa_screenshot_compare

Reyes Ayala, B., Phillips, M. E., & Ko, L. (2014). Current quality assurance practices in web archiving (Research Report). Retrieved from http://digital.library.unt.edu/ark:/67531/metadc333026/

Reyes, B. (2018). *A grounded theory of information quality in web archives* (Order No. 11005435). Available from ProQuest Dissertations & Theses Global. (2130612579). Retrieved from https://login.ezproxy.library.ualberta.ca/login?url=https://search.proquest.com/docview/2130612579?accountid=14474

Rosettacode.org. (2018). Percentage difference between images. Retrieved from https://rosettacode.org/wiki/Percentage_difference_between_images#Python

University of Alberta. (2013). Idle No More Collection. https://archive-it.org/collections/3490

University of Alberta. (2015). Western Canadian Arts Collection. https://archive-it.org/collections/6296

Wang, Z, Bovik. A. C., Sheikh, H. R., Simoncelli, E. P., et al. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600–612.