IPSW - Modelling Change of Website Archives

Caoimhe Rooney[1], Ian Roper[1], Jessica Williams[1], Ian Milligan[2], and Nathan Taback[3]

[1]Mathematical Institute, University of Oxford
[2]Department of History, University of Waterloo
[3]Department of Statistical Sciences and Computer Science, University of Toronto

1 Problem and Aims

Web archives, collections of old websites dating back to the mid-1990s that have been collected by institutions like the Internet Archive and libraries around the world, are invaluable research objects. They often consist of website domains (i.e. `www.domain.com` or `www.cnn.com`) which have been "crawled" by a web archival institution, which obtains all of the information on a given domain on a particular date. The scale of all this data is overwhelming for researchers to be able to extract the relevant data from the large number of crawls. Being able to find where large changes to the domain occurred in time so that the crawls performed on these dates can be investigated further would be of significant use for historians, as for example they could find particular websites amongst hundreds of domain snapshots to examine. Furthermore, the libraries which crawl the websites and store the archives have a data 'allowance' which the cost of each crawl is taken out of. Therefore, for example, a librarian might be interested in crawling a website that is updated more frequently more often; conversely, a website that is rarely updated might be crawled less. In other words, more information would help guide crawling decisions.

In our project, we find (from existing literature, for example [1]) or create different metrics for quantifying change between websites. We then apply these metrics to a web domain crawled from several different dates, comparing two consecutive crawls. This allows us to produce a time series of how much the website has changed from the previous crawl. We then compare the different metrics against each other, and attempt to identify large changes of the organisation behind the website which we know occurred in the past.

2 Text

We obtain the plain text (i.e. what appeared on all of the HTML pages) from the domain homepage for every crawl. Our goal is to compare the text from one crawl to the next and quantify how much the text has changed between them. There are a variety of metrics within the literature, in particular we explore the metrics described by Kwon *et al.* [2]. All the metrics are normalised to have a quantity between zero and one such that a crawl of a page that was identical to the last crawl should return a score of zero, and a crawl of a page that has absolutely nothing in common with the previous crawl should return one.

2.1 Metrics

Byte-wise comparison metric: This technique compares two webpages sequentially character by character. The metric then returns one if any change has occurred and zero otherwise [3, 4, 5]. A one is returned for even very trivial changes, for example, adding a blank space. Therefore, this metric is over-sensitive and does not provide particularly meaningful insight into the change between two strings of text. However, the byte-wise metric is useful in limiting the pages of interest, namely, if the metric is zero for any crawl, we know that there have been absolutely no changes at all and hence we need not explore this crawl further.

TF.IDF cosine distance: TF.IDF is shorthand for "term frequency-inverse document frequency" and is used to quantify how important a word is to a document of text. The underlying concept is that relevant words are not necessarily the most frequent words. For example, if considering book reviews, the words "character" or "plot" might appear very frequently, but do not give valuable insight to summarise the review. This metric is calculated by finding the term frequency (TF) of a word, namely the frequency of a word in a document. We also find the inverse document frequency (IDF) of a word, which is the

measure of how significant that term is in the collection of documents. By combining these concepts, we obtain TF.IDF weighted vectors to represent the content of each document and the metric value is calculated as the cosine distance between them [6]. TF.IDF has evident success in search engine algorithms to shift the definition of word-value from frequency to relevance [7].

Word distance: The word distance metric calculates the number of words in a document that have changed [8]. This is done by counting the number of common words in each document and normalising with respect to the total number of words in the two documents. Although less sensitive than byte-wise, both TF.IDF and the word distance metric are unable to account for change in word order.

Levenshtein distance:
The Levenshtein distance between two strings is the minimum number of single character edits, (substitutions, insertions or deletions) required to transform one string into the other [9]. For example, the Levenshtein distance between "test" and "tent" is 1, due to the single substitution of "s" to "n". We calculated the Levenshtein distance between the text of two webpages and normalised this value according to the maximum Levenshtein distance: the length of the longer string.

3 Hyperlinks

There are two ways in which we use the hyperlinks of a web domain to quantify the change between the domain from different crawls. The first of which is to compare the hyperlinks from the domain to pages of external domains. For simplicity, we only consider hyperlinks on the homepage of the domain pointing to webpages of external domains and we do not consider the frequency of each hyperlink, only whether the hyperlink exists or not. Our method for quantifying the change in external hyperlinks is similar to that of the word distance method. We divide the total number of links that are present in the two different crawls of the domain and normalise this by the average number of hyperlinks on the homepage of the domain from both crawls. We believe that links to external websites would be of particular interest to the creators of a domain as they are linking to other organisations which they are passionate about. Therefore, if these links change, it would suggest a large change in the focus of the domain.

The second method we use is to represent all of the webpages in a domain as nodes in a network which are connected by directed edges representing hyperlinks from one webpage to another. This network representing the structure of a domain may change from one crawl to the next, when webpages are added and removed from a domain between two crawls or when hyperlinks within a domain change. Several global metrics to quantify the change between these networks have been proposed, including compactness and stratum, which are explained in detail by Botafogo *et al.* in [15]. Depending on the structural changes that are to be detected or expected from the domain, these different metrics can be more or less useful. Unfortunately, at the time of writing, we do not have sufficient webpage data to perform this analysis on the data from the NDP domain data, however, we hope to be able to continue this work once the data is able to be obtained.

4 Thumbnails

In addition to changes in website text and structure, meaningful change in domain is often reflected by a change in the visual structure of the page. We can generate a website snapshot of one of the webpage crawls using the Wayback Machine (`www.archive.org/web`) which holds one of the web archives. Hence, a promising approach to quantify the change in a web domain, is by applying image analysis techniques to detect the similarity between webpage thumbnails [10]. This has previously been considered for pairs of images, although plotting similarity over time has yet to be considered. Several methods for comparing thumbnails have been proposed and implemented previously [11, 12, 13], and an accessible summary is provided in [10].

Some image comparison techniques may not always produce meaningful results – e.g., images on a homepage may change frequently, with no change in website content. Therefore, we propose the use of the structural similarity index (SSIM) [14], which measures the similarity between two images by comparing average pixel intensity in various sub-windows of the page. The SSIM value is between -1 and 1 so in order to compare this metric with the text and hyperlink metrics, we scale the SSIM value to lie between zero and one, and call this metric, d.

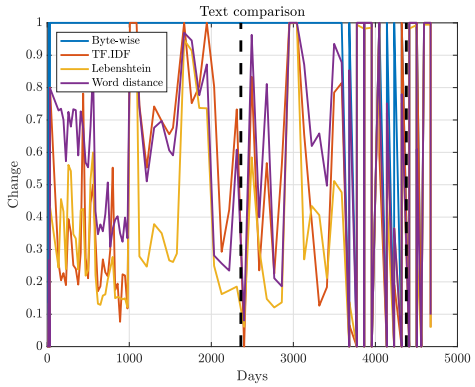
Our `python` code automatically generates thumbnails of the homepage of all crawls of a chosen domain recorded in the Wayback Machine, and uses this to obtain a library of 108 thumbnails from `www.ndp.ca`

Figure 1: Three image comparisons from `www.ndp.ca`.

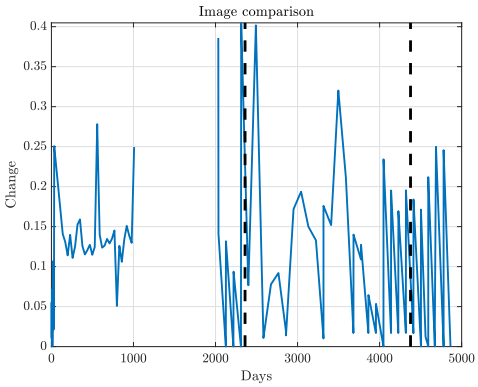
from 2005-2019. There were times when the Wayback Archive had only saved a page that had failed to render, which presented itself as a primarily white webpage. We detected these 'fails' and removed them from the data set by imposing a maximum percentage of white pixels (80%). In Figure 1, we display an example of a failed render, as well as two different timesteps which demonstrate visually the value of the SSIM metric.

5 Results

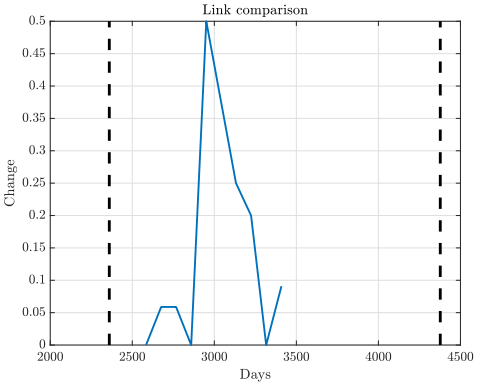
Results for all metrics considered in this report are plotted in Figure 2 for data obtained from `www.ndp.ca`. Figure 2(a) displays the four text metrics described in Section 2, and we see that all metrics follow a similar pattern of peaks and troughs with the exception of the Byte-wise metric. This is due to it only taking values of zero or one as explained in Section 2. The image comparison, using the SSIM metric described in Section 4 and scaled appropriately is plotted in Figure 2(b). The section of missing data corresponds to a time interval when the wayback archive failed to generate a representative thumbnail (see Section Figure 1). The link comparison is shown in Figure 2(c). We see a single significant peak around 3000 days, but we unfortunately only had data for a small time period, so it remains to explore this metric in more detail once a larger data set has been obtained. Finally, in Figure 2(d), we plot the three different approaches to quantifying domain change, using word distance as a representative measure for change in text. We see that, although the scale of the produced metric varies, the general pattern appears to be consistent. This is reassuring that all metrics appear to qualitatively agree on locations of significant change. It is worthwhile to note, that the data we used was not collected at fixed time intervals. Therefore, the high frequency of peaks observed between 4000 and 5000 days in Figures 2(a) and 2(b) can be attributed to the larger frequency in data collection. Additionally, due to the non-uniform time steps between crawls, peak magnitude is not necessarily a reliable metric of change magnitude. Assuming a change to a domain occurs gradually, for data collected at smaller intervals, we expect peak magnitudes to be smaller but more frequent, summing to a total larger change over the full interval.



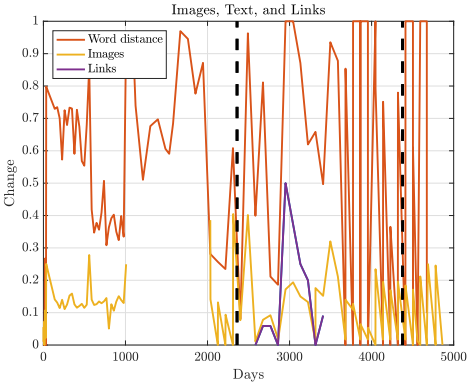
(a) Results of the four text comparison metrics.



(b) Results of the image comparison (the missing data corresponds to wayback rendering fails).



(c) Results of the link comparison (plotted for the period where we had data available).



(d) Results from text comparison, link comparison, and image comparison on the same axis. Text metric used in this plot is word distance.

Figure 2: Preliminary results using data generated from `www.ndp.ca`. The dashed black lines correspond to the two days during this period when the NDP leadership changed.

6 Conclusion

This report outlines some preliminary findings at the Industrial Problem Solving Workshop 2019 at the Fields Institute, Toronto. We explore the quantification of website domain change over time through the analysis of text, link, and thumbnail change. We believe that these different metrics have different uses depending on the domain being analysed and the type of change that is trying to be detected. For example, news websites update their text content every day, so any metric reliant solely on text may be over-sensitive. On the other hand, governmental websites may only update their text at times of relevant policy or leadership change. The results in Figure 2(d) demonstrate, for the data analysed in this report (`www.ndp.ca`), that the metrics agree qualitatively on the locations of large and small change. This implies some consistency between the metrics and suggests that with further work and added sophistication to the metrics, these techniques could be used to successfully indicate significant changes in web domains. An extension to this work could include combining the three different metrics in a weighted sum, where the weightings for each metric could be chosen based on the content being analysed, and by using these results to predict when a domain is next going to have a large change.

References

[1] D. Dhyani, W. K. Ng, and S. S. Bhowmick, "A survey of web metrics," *ACM Computing Surveys (CSUR)*, vol. 34, no. 4, pp. 469–503, 2002.

[2] S. Y. Kwon, S. H. Lee, and S. J. Kim, "A precise metric for measuring how much web pages change," in *International Conference on Database Systems for Advanced Applications*, pp. 557–571, Springer, 2006.

[3] B. E. Brewington and G. Cybenko, "How dynamic is the web?," *Computer Networks*, vol. 33, no. 1-6, pp. 257–276, 2000.

[4] J. Cho and H. Garcia-Molina, "The evolution of the web and implications for an incremental crawler," tech. rep., Stanford, 1999.

[5] S. J. Kim and S. H. Lee, "An empirical study on the change of web pages," in *Asia-Pacific Web Conference*, pp. 632–642, Springer, 2005.

[6] G. Salton and M. J. McGill, "Introduction to modern information retrieval," 1986.

[7] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "paper recommender systems: a literature survey," *International Journal on Digital Libraries*, vol. 17, no. 4, pp. 305–338, 2016.

[8] A. Ntoulas, J. Cho, and C. Olston, "What's new on the web?: the evolution of the web from a search engine perspective," in *Proceedings of the 13th international conference on World Wide Web*, pp. 1–12, ACM, 2004.

[9] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet physics doklady*, vol. 10, 1966.

[10] A. AlSum and M. L. Nelson, "Thumbnail summarization techniques for web archives," in *European Conference on Information Retrieval*, pp. 299–310, Springer, 2014.

[11] Henzinger and Monika, "Finding near-duplicate web pages: A large-scale evaluation of algorithms," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 284–291, 2006.

[12] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, "Syntactic clustering of the web," *Computer Networks and ISDN Systems*, vol. 29, no. 8, pp. 1157 – 1166, 1997.

[13] G. S. Manku, A. Jain, and A. Das Sarma, "Detecting near-duplicates for web crawling," in *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, (New York, NY, USA), pp. 141–150, ACM, 2007.

[14] D. Brunet, E. R. Vrscay, and Z. Wang, "On the mathematical properties of the structural similarity index," *IEEE Transactions on Image Processing*, vol. 21, pp. 1488–1499, April 2012.

[15] R. A. Botafogo, E. Rivlin, and B. Shneiderman, "Structural analysis of hypertexts: Identifying hierarchies and useful metrics," *ACM Transactions on Information Systems (TOIS)*, vol. 10, no. 2, pp. 142–180, 1992.