IPSW - Modelling Change of Website Archives

1 Problem and Aims

Web archives are past versions of domains, consisting of all the webpages with the same IP address (www.____.com), which are saved on large databases around the world to keep records on this vast but rapidly changing source of information. To keep these records, the website must be 'crawled' to obtain all the information of a website 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 o crawls. Being able to find where large changes occurred in websites so the crawls performed on these dates can be investigated further would be of significant use for historians. Furthermore, the libraries which crawl the websites and store the archives have a data 'allowance' which each crawl uses. Therefore, being able to detect when a large change has occurred before crawling the whole domain or even being able to predict when a large change will occur next is hugely valuable.

We find or derive different metrics for quantifying change between two websites. We then apply these metrics to web domains crawled from several different dates, comparing the crawls from two consecutive crawls. This allows us to produce a time series of how much the website has changed from the previous crawl against the time of the new 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 Data

3 Text

We obtain the text from the homepage at every point in time. Our goal is to compare the text from one time-point to the next and quantify how much the text has changed between these measurements. There are a variety of metrics within the literature, in particular we explore the metrics described by Kwon et al. [?].

3.1 Metrics

Byte-wise comparison metric

Compares two webpages sequentially character by character. The metric returns 0 when no change and 1 otherwise. It returns 1 for even trivial cases, for example, blank space – over-sensitive.

TF.IDF cosine distance

Comparing the number of times key words appear in the document.

Word distance

Percentage of words that have stayed the same.

Edit distance

Number of edits that are required to transform one sentence into another.

Shingling metric

Breaks up the webpage into subsequences called "shingles" that contain k words.

3.2 Issues

- Byte-wise is over-sensitive,
- TF.IDF and Word Distance cannot take into account change in word order,
- Shingling is over-sensitive to small webpages,

4 Thumbnails

A promising method to check whether meaningful changes have occurred to a domain is to compare homepage thumbnails at two different time points. This is often done manually, although automated image analysis approaches have also been investigated. This method may not always produce meaningful results – e.g., moving a single image from a homepage may modify the appearance dramatically without a change in content. Therefore, it is important to also investigate alternative metrics. A summary of different approaches is provided in [**?**]. Ian could you add in a lit review?

5 Links

There are two ways in which we use the hyperlinks of a web domain to quantify the change between the domain at 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.

The second method we use is to represent 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 [**Reference**]. 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.

6 Conclusion