

IPSW - Modelling Change of Website Archives

Group 4

Ian Milligan, Ian Roper, Caoimhe Rooney,
Nathan Taback, Jessica Williams, Nich Worby

May 9, 2019

The Problem

The Problem

- Decades of web archives with lots of vital historic information
- Gets overridden from history when the website is updated
- Ever increasing amount of data
- Scale overwhelms the search for the meaningful information

The Problem

- Decades of web archives with lots of vital historic information
- Gets overridden from history when the website is updated
- Ever increasing amount of data
- Scale overwhelms the search for the meaningful information

Can we

- find out when large changes have occurred?
- predict when a big change is going to occur?

The goal

- Construct and compare different metrics to quantify domain changes over time,
- Determine a single quantitative measure to describe magnitude of the change in the domain since the previous time-step.

$$\sigma(t) = (\text{change in links})w_1 + (\text{change in text})w_2 + (\text{change in content management server})w_3, \quad (1)$$

where t is time, and w_1 , w_2 , and w_3 weight the relevant contributions of URL changes, text changes, and CMS changes.

- Meaningful changes determined by comparing thumbnails manually.
- Could automate this by using image analysis to quantify the difference between website thumbnails at two time points.

Game plan

- Run code to compare text.
- Do image analysis on thumbnails.
- Take link data and compare lists at different times:
 - Internal vs. external links.
 - Obtain a , b , and c .
 - What is the best timestep?
- Determine whether the content management server (CMS) has changed.
- Look at different weightings - how best to choose these? We don't want to double-count changes.
- Run test cases.
- Look at the variability in change over time. What is the distribution?
- Compare measures for looking at the difference between URLs and text.

- Trying to quantify change using text, thumbnails and links.
- Lots of metrics about the how the text differs and some of these are similar.
 - There is one that is overly sensitive but there is still one timestamp that says there is absolutely no change and so could still be useful.
- Thumbnails obtained using the wayback archive which renders the homepage and takes a screenshot.
 - We've used a metric that looks at structural similarity instead of just pixel to pixel which is good.
 - We've had a problem with the website not always rendering and giving us just a white page which obviously causes a huge change. This needs to be accounted for tomorrow.

- The link data has been analysed
- - Unfortunately the dates for these data is shorter than the text data frame so it is difficult to get a good comparison.
 - We have Ian, the history professor on this task.
- Graphs of links
 - One last thing we were thinking of doing is getting the internal links within a whole domain instead of just the homepages, seeing how the structure of the graph of links between them changes.
 - This is more of a structural change than a content change, which could be useful for rapidly updating websites such as news websites and blogs whose words change rapidly but fairly meaninglessly.