IPSW - Modelling Change of Website Archives

Group 4

1 Introduction

One 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 [1]. Ian could you add in a lit review?

Our goal is to investigate a novel metric for domain change, focussing on changes in *links*, pointing either to internal webpages or to external domains. We propose that the magnitude of the change in the domain at time-step n from time-step $n-1$ is

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

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

2 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 number of links added, links removed, and links changed.
  - 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.

References

[1] Shin Young Kwon, Sang Ho Lee, Sung Jin Kim. *A Precise Metric for Measuring How Much Web Pages Change*. School of Computing, Soongsil University.