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?

Aims for this week

- Find and explore ways to quantify change in a website
- Compare these quantifications
- See if we can identify big changes in an organisation from our research

Big events in the NDP

- 28 November 2005: election called.
- 23 January 2006: federal election.
- 14 October 2008: federal election.
- 2 May 2011: federal election.
- July 2011: NDP leader announces leave of absence; replaced by interim.
- 22 August 2011: NDP leader dies.
- 24 March 2012: New NDP leader selected.
- 19 October 2015: federal election.
- 10 April 2016: NDP leader loses vote of confidence.
- 1 October 2017: New NDP leader selected.

Attempted Approaches

- How much words on the domain change
- How the links out of the domain change
- How the way the domain looks changes
- How the structure of the websites within the domain change

Four Metrics for Text

- Byte-wise comparison:
  - If any change in characters has occurred, $= 1$
  - If text is *exactly* the same, $= 0$
- TF·IDF
  - Calculates cosine distance between two different vectors of characters $\boldsymbol{p}$ and $\boldsymbol{p}'$
- Word distance
  - How many words have changed
- Edit distance
  - "Edit distance" δ is the amount of insertion/deletion/substitution needed to turn one sequence into the other

$$1 - \frac{\boldsymbol{p} \cdot \boldsymbol{p}'}{||\boldsymbol{p}||_2 ||\boldsymbol{p}'||_2}$$

$$1 - \frac{2|common\ words|}{m+n}$$

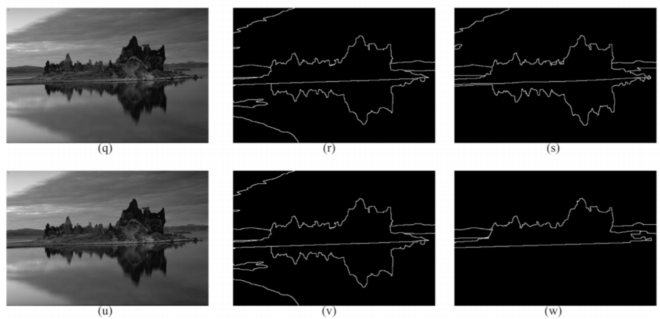
$$\frac{\delta}{m+n}$$

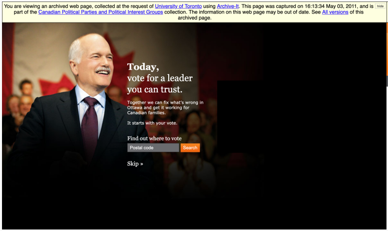
External Links

- Justification
  - Links to other websites are important to the website designer
  - If these change, the topic of the website has most likely changed as well
- Method
  - Compare vector of links on homepage at t_i and t_{i+1} as $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$

$$1 - \frac{2|common\ links|}{|\mathbf{v}_i| + |\mathbf{v}_{i+1}|}$$

Screenshot Comparison



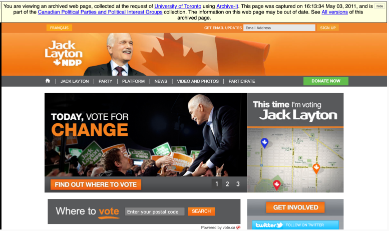

You are viewing an archived web page, collected at the request of University of Toronto using Archive-It. This page was captured on 16:13:34 May 03, 2011, and is part of the Canadian Political Parties and Political Interest Groups collection. The information on this web page may be out of date. See All versions of this archived page.

Today,
vote for a leader
you can trust.

Together we can fix what's wrong in Ottawa and get it working for Canadian families.

It starts with your vote.

Find out where to vote


You are viewing an archived web page, collected at the request of University of Toronto using Archive-It. This page was captured on 16:13:34 May 03, 2011, and is part of the Canadian Political Parties and Political Interest Groups collection. The information on this web page may be out of date. See All versions of this archived page.

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.

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

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

Current methods

- 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.

Image comparison results

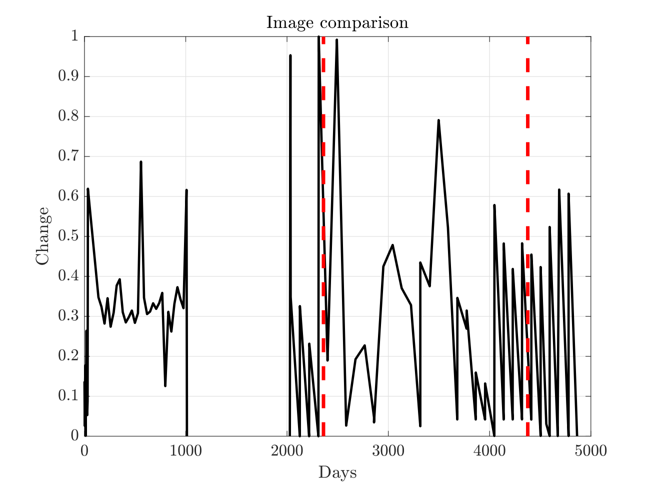


Figure 1: Red lines where new NDP leader selected.

Link comparison results

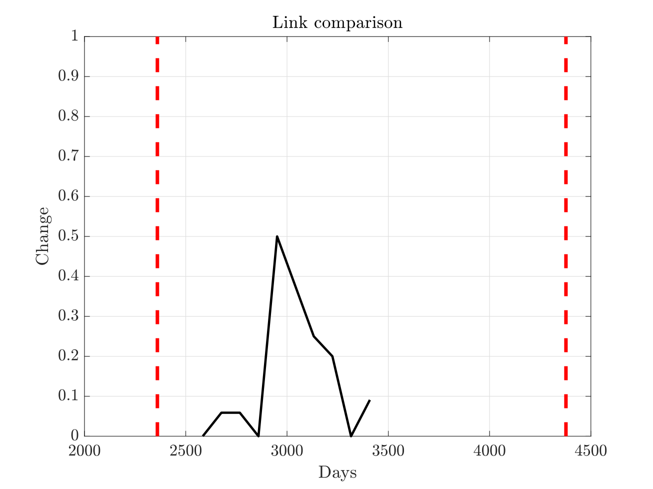


Figure 2: Red lines where new NDP leader selected.

Text comparison results

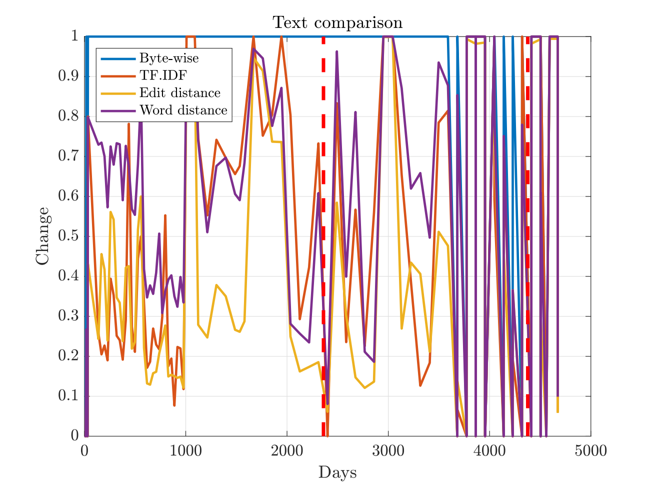
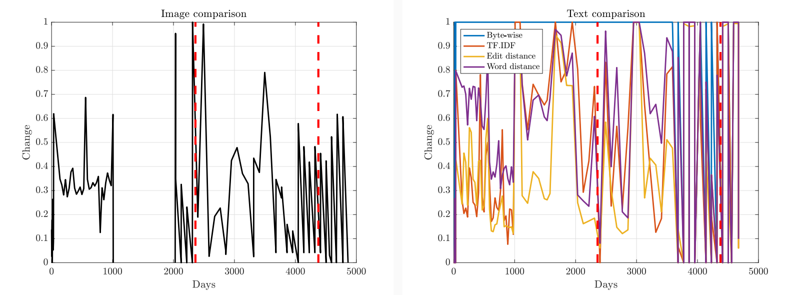


Figure 3: Red lines where new NDP leader selected.

Conclusions



(a) Image comparison

(b) Text comparison

Figure 4: Locations of high change correspond to new NDP leader selected.

Conclusions

- Different metrics are useful for different websites or for different types of change,
  - e.g. news websites update content every day but this might not indicate significant change – structural metric more informative than text metric,
  - e.g. governmental websites depend sensitively on text and content – text metric most informative,
  - e.g. job registers will link to new advertisements – link data most informative.
- We see that the text, link and thumbnail metrics align for certain substantial changes.