

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

- 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 many words on the domain change?
- How do the links out of the domain change?
- How does the way the website looks changes?
- How does 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 p and 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

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

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

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

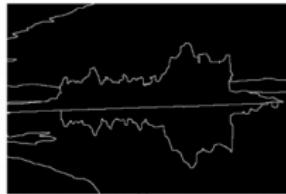
- Justification
 - Links to other websites are important to the website designer
 - If these change, the topic of the website has most likely changed as well
- Method
 - Compare vector of links on homepage at t_i and t_{i+1} as \mathbf{v}_i and \mathbf{v}_{i+1}

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

Image Comparison



(q)



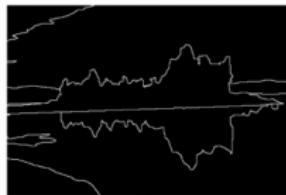
(r)



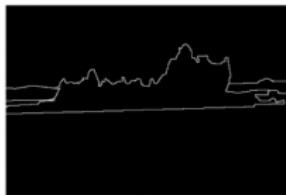
(s)



(u)



(v)



(w)

You are viewing an archived web page, collected at the request of University of Toronto using ArchiveIt. 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 bring back what's right for Canadian families.

It starts with your vote.

Find out where to vote

[Find out now](#) [Skip >](#)

You are viewing an archived web page, collected at the request of University of Toronto using ArchiveIt. 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.



Jack Layton +NDP

TODAY, VOTE FOR CHANGE

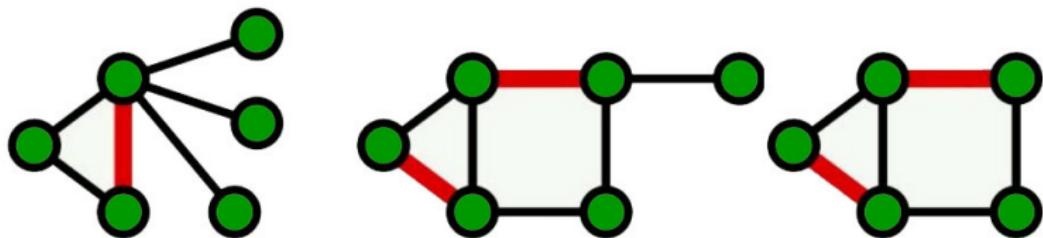
FIND OUT WHERE TO VOTE

Where to vote enter your postal code [SEARCH](#)

This time I'm voting Jack Layton

GET INVOLVED

Structure



<https://www.geeksforgeeks.org/mathematics-matching-graph-theory/>

Issues

- Data takes a long time to retrieve from servers
- Different languages
- Failed renderings
- Difficult to decipher where website fits in structure just from URL
- Internal links data showing strange behaviour

Text comparison results

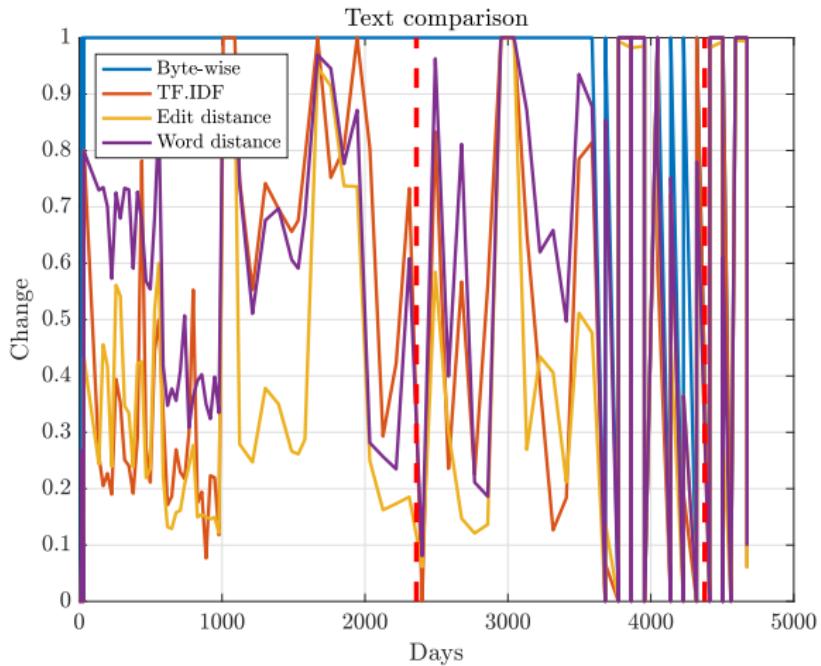


Figure 1: Red lines where new NDP leader selected.

Link comparison results

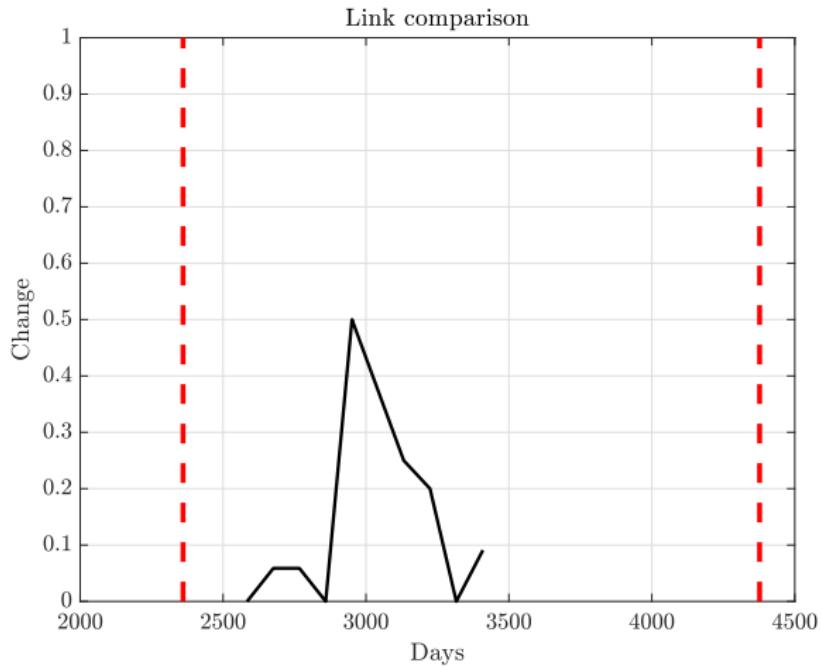


Figure 2: Red lines where new NDP leader selected.

Image comparison results

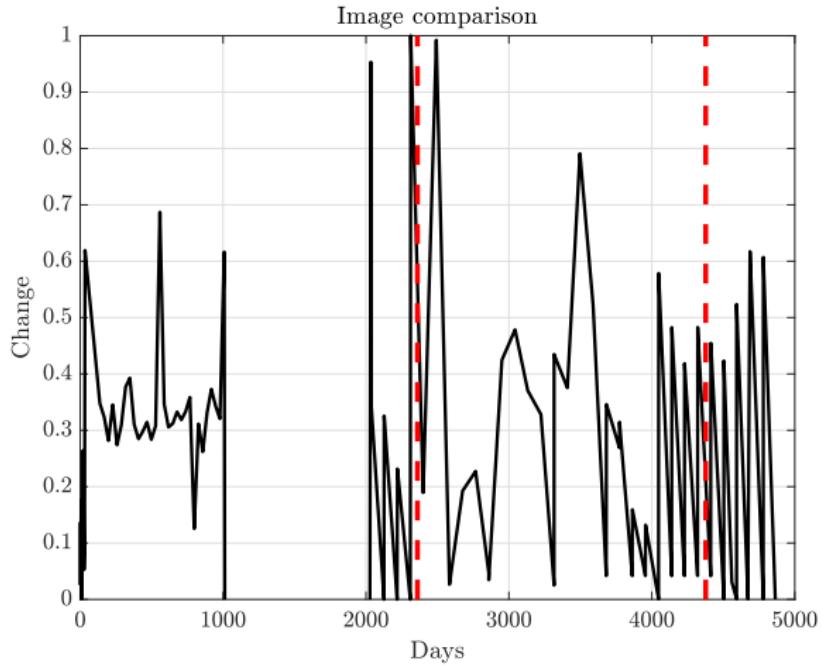


Figure 3: Red lines where new NDP leader selected.

Conclusions

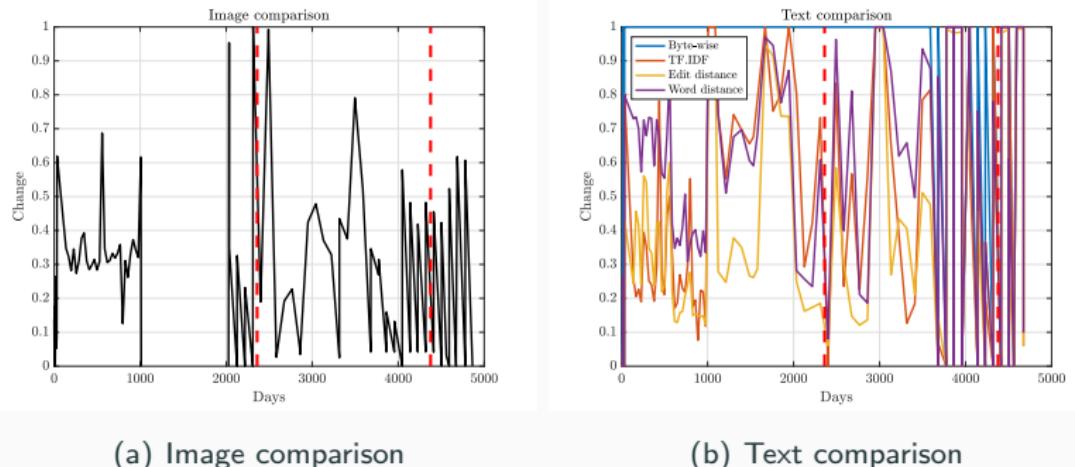


Figure 4: Locations of high change correspond to new NDP leader selected.

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 and screenshot metrics align for certain substantial changes.

Moving Forward

- Use more datasets to find which metrics are useful for which kind of websites
- Use more sophisticated language detection techniques to convey context (machine learning...)
- Computer science knowledge needed to understand greater indicators of substantial changes
- Full internal link data could be used to see website structure changes
- Distribution of the change could be used to predict when next crawl should be made (next study group?!)

QUESTIONS?

