Efficient Thumbnail Generation for Web Archives

Mat Kelly, Michael L. Nelson, Michele C. Weigle
{mkelly, mln, mweigle}@cs.odu.edu

WHAT is this work about?

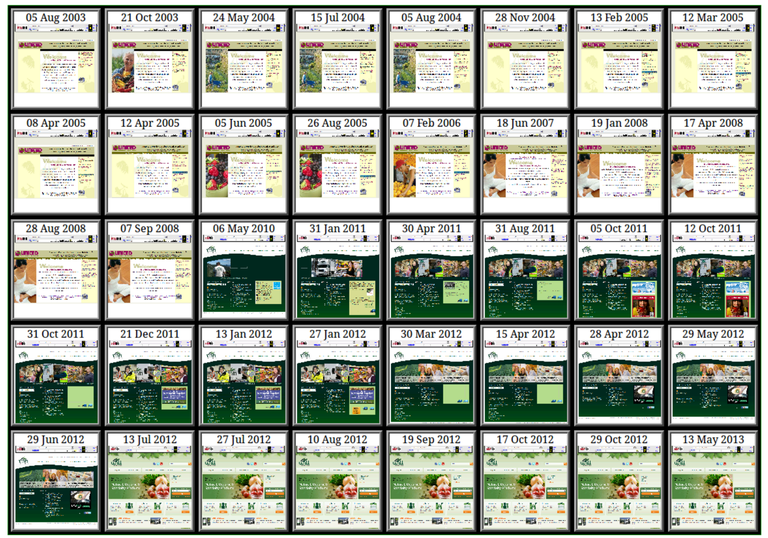
- thumbnails are useful for archival summary of a site
- thumbnails are expensive to generate, store, and view
- we analyze HTML and create thumbnails only for dissimilar pages

All Thumbnails

All Thumbnails (grouped by similarity)

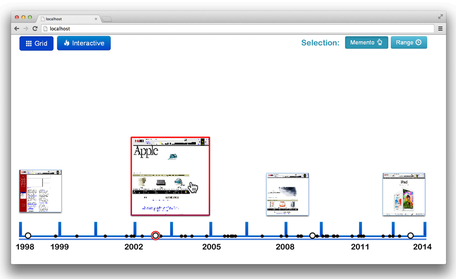
WHAT if we made thumbnails of everything?

Time Required to Generate Thumbnails (5sec/thumb)	Space Required	multiplied by **240 Billion** mementos	Thumbnail size *pixels, space/ea*	
~38,000 years	335 TB		64x64	3kB
	14.5 PB		600x600	133kB



Summarized Thumbnails

Alternate Views!

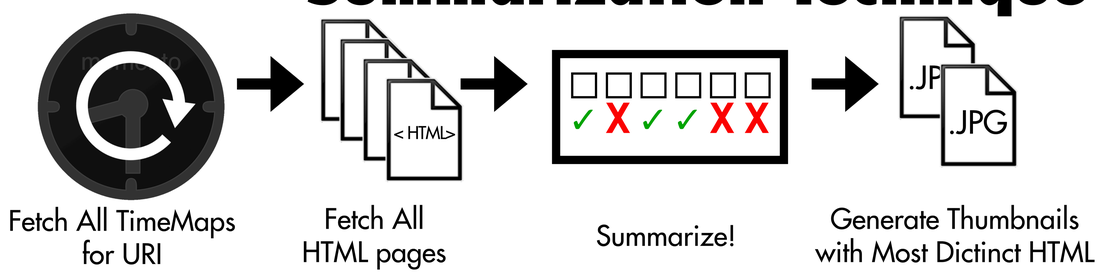


Coverflow

Timeline

WayBackAddOn

Summarization Technique



Fetch All TimeMaps for URI

Fetch All HTML pages

Summarize!

Generate Thumbnails with Most Dictinct HTML

Acknowledgements and References

This work supported in part by NSF IIS 1009392, the Andrew Mellon Foundation, and the National Endowment for the Humanities (NEH) Digital Humanities Start-Up Grant, HD-51670-13

A. AlSum, and M. L. Nelson. "Thumbnail Summarization Techniques for Web Archives." In Proceedings of the 36TH European Conference on Information Retrieval, ECIR 2014, 2014