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

WHAT if we made thumbnails of everything?

<table>
<thead>
<tr>
<th>Time Required to Generate Thumbnails (5sec/thumb)</th>
<th>Space Required</th>
<th>thumbnail size</th>
</tr>
</thead>
<tbody>
<tr>
<td>~38,000 years</td>
<td>335 TB</td>
<td>56x56 240 pixels</td>
</tr>
<tr>
<td></td>
<td>14.5 PB</td>
<td>64x64 300 pixels</td>
</tr>
</tbody>
</table>

Many Near-Duplicates!

Alternate Views!
- Coverflow
- Timeline

Summarization Technique
- Fetch All TimeMaps for URI
- Fetch All HTML pages
- Summarize!
- Generate Thumbnails with Most Distinct 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

WayBackAddOn
http://ws-all.cs.odu.edu