Trends in Web Characteristics

João Miranda, Daniel Gomes
{joao.miranda,daniel.gomes} @ fccn.pt
http://arquivo.pt
Summary

- Introduction
- Methodology
- Trends
- Conclusions
The Web

• The Web is a huge source of information
  – Information published exclusively on the Web
  – Information disappears
    • Only 20% of the URLs still reference a valid content after 1 year (Ntoulas, 2004)

• Preservation started by the Web Archives
  – Access for future generations
  – Enables research on finding trends (temporal dimension)
Altavista across time
• How does a crawler work?
  – Collects contents from the Web, starting from an initial set of addresses
  – Iteratively downloads contents and extracts links to find new ones
Methodology
Methodology

• Previous works
  – 2003 - textual media types
  – 2005 - all media types

• Ours
  – 2008 - all media types, textual media types
Trends
• Reflect the changes in the characteristics of contents and sites:
  – Are sites larger?
  – Which media types are prevalent?
  – Has the average size of contents grown?

• Why is it useful to analyse trends?
  – Improve the processment of web data
Site size – Why?

- Efficiently partition a large data set of URLs across several crawling processes
• Average number of contents per site increased from 70 to 96
Media types – Why?

• Browsers or document viewers for cellphones
### All Media types - Results

<table>
<thead>
<tr>
<th>Media type</th>
<th>% contents 2005</th>
<th>% contents 2008</th>
<th>Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Text/html</td>
<td>61.2%</td>
<td>57.8%</td>
<td>-5.5%</td>
</tr>
<tr>
<td>2 Image/jpeg</td>
<td>22.6%</td>
<td>22.8%</td>
<td>+1.2%</td>
</tr>
<tr>
<td>3 Image/gif</td>
<td>11.4%</td>
<td>9.4%</td>
<td>-17.4%</td>
</tr>
<tr>
<td>4 Text/pdf</td>
<td>1.6%</td>
<td>1.9%</td>
<td>+18.5%</td>
</tr>
<tr>
<td>- Other</td>
<td>3.2%</td>
<td>8.1%</td>
<td>-</td>
</tr>
</tbody>
</table>

- Decrease in `text/html` and `image/gif`
- Increase in `image/jpeg` and `text/pdf`
- Media type prevalence is more spread
### Textual Media types - Results

<table>
<thead>
<tr>
<th>Media type</th>
<th>% contents 2003</th>
<th>% contents 2008</th>
<th>Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Text/html</td>
<td>96.0%</td>
<td>93.9%</td>
<td>-2.1%</td>
</tr>
<tr>
<td>2 App’n/pdf</td>
<td>1.9%</td>
<td>3.0%</td>
<td>+57.6%</td>
</tr>
<tr>
<td>3 Text/plain</td>
<td>1.0%</td>
<td>1.6%</td>
<td>+58.5%</td>
</tr>
<tr>
<td>4 App’n/x-shockwave-flash</td>
<td>0.5%</td>
<td>1.2%</td>
<td>+115.8%</td>
</tr>
<tr>
<td>- Other</td>
<td>0.7%</td>
<td>0.3%</td>
<td>-</td>
</tr>
</tbody>
</table>

- *Text/html* lost presence to other formats
- Increase in *app’n/pdf, text/plain, app’n/shockwave-flash*
• Estimate the storage resources required to create Web data repositories
### Content size - Results

| Media type               | Avg Size 2003 | Avg Size 2008 | Trend  
\[
\left( \frac{\text{Avg size 2003} - \text{Avg size 2008}}{\text{Avg size 2003}} \right) \times 100
\] |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Text/html</td>
<td>21 KB</td>
<td>30 KB</td>
<td>+45.9%</td>
</tr>
<tr>
<td>App’n/pdf</td>
<td>207 KB</td>
<td>252 KB</td>
<td>+21.6%</td>
</tr>
<tr>
<td>Text/plain</td>
<td>11 KB</td>
<td>44 KB</td>
<td>+58.5%</td>
</tr>
<tr>
<td>App’n/x-shockwave-flash</td>
<td>44 KB</td>
<td>90 KB</td>
<td>+115.8%</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>powerpoint</td>
<td>1055 KB</td>
<td>500 KB</td>
<td>-52.6%</td>
</tr>
<tr>
<td>Text/rtf</td>
<td>476 KB</td>
<td>143 KB</td>
<td>-70.0%</td>
</tr>
</tbody>
</table>

- Content size grew but for some types decreased
Dynamically generated contents – Why?

- Identify technological trends in Web publishing
Dynamically generated contents - Results

- The percentage of URLs containing parameters increased

- 2003: 47.2%
- 2008: 63.3%
URL length – Why?

- Influences interaction design
- Determine adequate length for input boxes that receive URLs

- How many characters should be presented on a search engine results page
- Average URL length increased from 62 to 73 after 5 years
Conclusions
Conclusions I

• After 5 years what were the changes?
  – URL length increased slightly but the average content size increased significantly
  – Sizes did not grow for all media types
• After 5 years what were the changes?
  – Dynamically generated contents became widely used
  – Number of contents per site increased
  – HTML, GIF or JPEG became prevalent
Future work

• Analyze evolution of metrics related to Web quality

• Measure the evolution of the Web usability and accessibility across time
  – New technologies do not imply significantly better user experience
Collaborate with us

• Testing of developed systems
• Research and development projects
  – Image search
• Lend disk space to keep backup copies
  – Just need to install rARC
• Crawl logs available for research purposes

http://arquivo.pt
Thank you.

http://arquivo.pt