An updated Portrait of the Portuguese Web

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

- Introduction
- Methodology
- Metrics
- Conclusions
Introduction
• The Web is a huge source of information
  – Information published exclusively on the Web
  – Information disappears

• Preservation started by the Web Archives
  – Access for future generations

• First initiative: Internet Archive
Altavista across time
The Portuguese Web Archive

- The Portuguese Web Archive

![Diagram of web archive components: Web, Crawler, Content Storage, Indexer, User Interface]
• What is a crawler?
  – Collects contents from the Web
  – Starts from an initial set of addresses

• How does it work?
  – Iteratively downloads contents
  – Extracts links to find new ones
Methodology
Methodology

• Crawl of the Portuguese Web (March-May, 2008)
  – .PT domain
  – Heritrix crawler
  – 180 000 initial addresses
  – 48 million contents
  – 2.5 TB

• No content analysis, only log analysis
Metrics
Sites hosted per IP address – Why?

- Politeness policies for crawling
Sites hosted per IP address - Results

- 75% of the IP addresses host 1 site
Successful responses – Why?

• Quality indicator

• Large % of broken links mines trust of users
Successful responses - Results

- 18% of the sites returned 100% OK responses
Media types – Why?

- Browsers or document viewers for cellphones
- Parsing and indexing for search engines
Media types - Results

- 90% of the number of contents are html, jpeg, gif

<table>
<thead>
<tr>
<th>Media type</th>
<th>% contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Text/html</td>
<td>57.8%</td>
</tr>
<tr>
<td>2 Image/jpeg</td>
<td>22.8%</td>
</tr>
<tr>
<td>3 Image/gif</td>
<td>9.4%</td>
</tr>
<tr>
<td>4 Text/xml</td>
<td>1.9%</td>
</tr>
<tr>
<td>- Other</td>
<td>8.1%</td>
</tr>
</tbody>
</table>

- 69% of the amount of data are html, pdf, jpeg

<table>
<thead>
<tr>
<th>Media type</th>
<th>% amount data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Text/html</td>
<td>35.4%</td>
</tr>
<tr>
<td>2 App’n/pdf</td>
<td>17.9%</td>
</tr>
<tr>
<td>3 Image/jpeg</td>
<td>16.1%</td>
</tr>
<tr>
<td>4 Text/plain</td>
<td>4.2%</td>
</tr>
<tr>
<td>- Other</td>
<td>26.4%</td>
</tr>
</tbody>
</table>
Content size – Why?

• Estimate the storage resources required to create Web data repositories
Content size - Results

- 96% lower than 128 KB
Dynamically generated contents – Why?

- Identify technological trends in Web publishing
• At least 46.3% of the contents were dynamically generated
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
84% lower than 100 characters
Conclusions
Conclusions I

• Long URL addresses

• Half of the contents are dynamically generated (mainly PHP)

• 90% of the contents are HTML, JPEG and GIF

• 69% of the amount of data are HTML, PDF and JPEG
Conclusions II

- 96% of the contents are smaller than 128 KB
- Half of the sites present a successful response rate below 80%
- Most IP addresses host a single site
“Future” work

- Study trends in the evolution of web characteristics
  - João Miranda, Daniel Gomes, Trends in Web characteristics, 7th Latin American Web Congress

- Analyze metrics extracted from content and link analysis
Contribute to preserve the Web

• Anyone can contribute to preserve the Web
• Lend disk space to keep backup copies
  – Just need to install rARC
  – [http://arquivo.pt/rarc](http://arquivo.pt/rarc)
• Help required to test beta version
Thank you.

Logs used in this study are available for research purposes. Please contact us.

http://arquivo.pt