PORTUGUESE WEB ARCHIVE

Trends in the Evolution of the Web

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

FCCN
Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

UMIC
Agência para a Sociedade do Conhecimento

POS_CONHECIMENTO
Programa Operacional Sociedade do Conhecimento

União Europeia
FEDER

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)

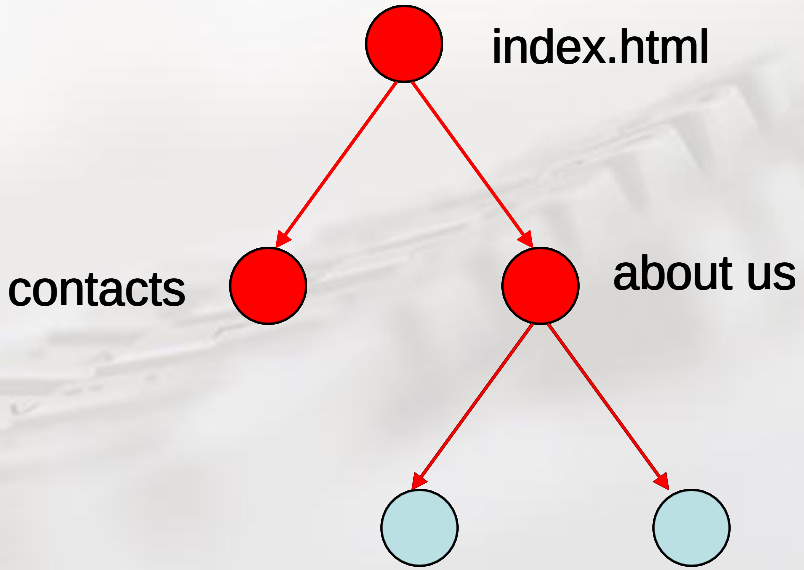
Web Archives

- Altavista across time

2004

Crawler

- 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

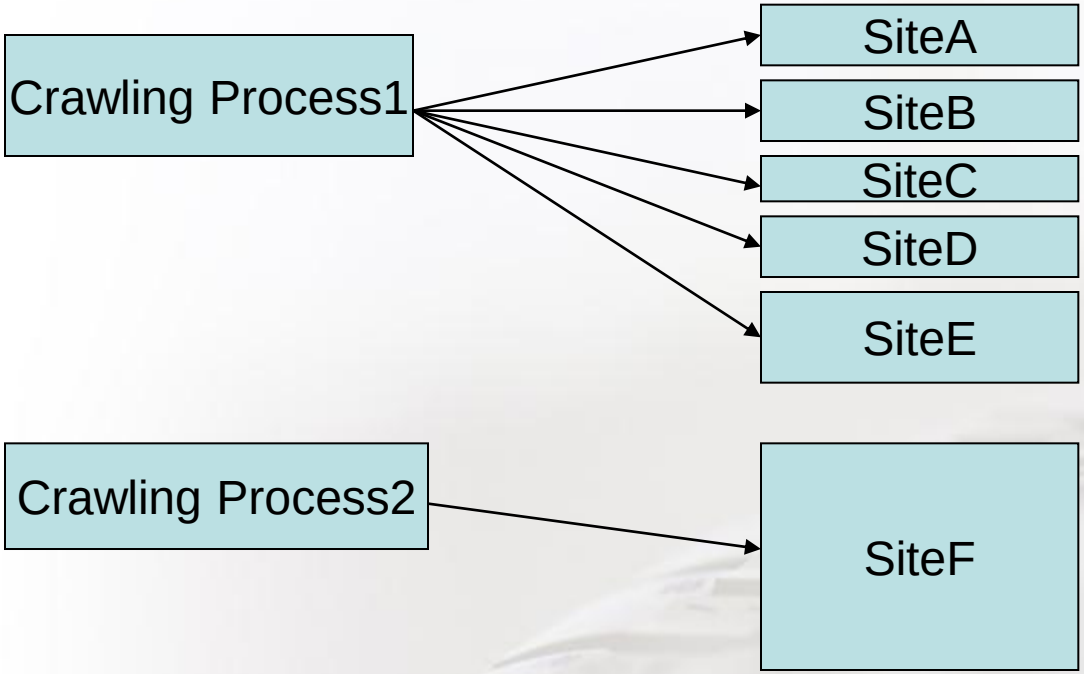
Trends

- 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?

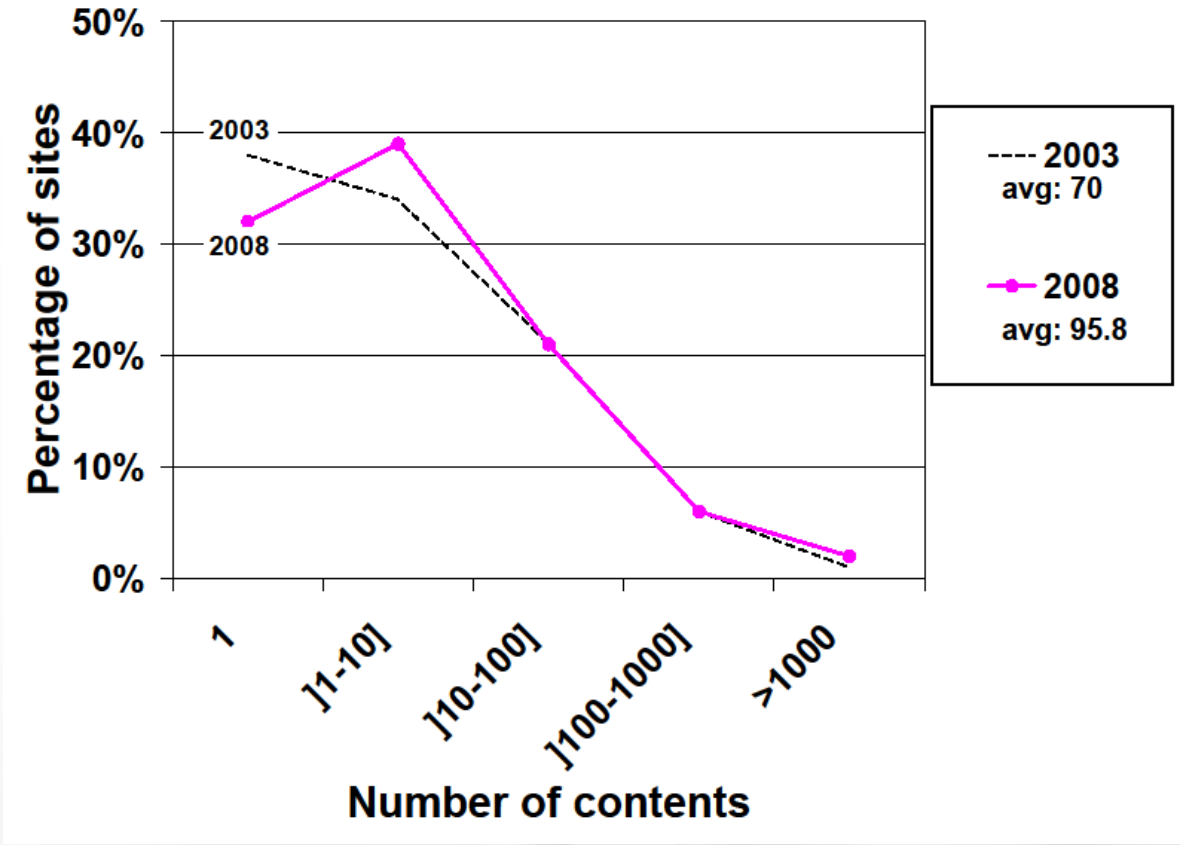
- Why is it useful to analyse trends?
  - Improve the processment of web data

Site size – Why?

- Efficiently partition a large data set of URLs across several crawling processes



Site size - Results



- Average number of contents per site increased from 70 to 96

Media types – Why?

- Browsers or document viewers for cellphones

All Media types - Results

95.2% 90.1%

	Media type	% contents 2005	% contents 2008	Trend $\left(\frac{\%\text{ contents 2005} - \%\text{ contents 2008}}{\%\text{ contents 2005}}\right)\times 100$
1	Text/html	61.2%	57.8%	-5.5%
2	Image/jpeg	22.6%	22.8%	+1.2%
3	Image/gif	11.4%	9.4%	-17.4%
4	Text/pdf	1.6%	1.9%	+18.5%
-	Other	3.2%	8.1%	-

- Decrease in *text/html and image/gif*
- Increase in *image/jpeg* and *text/pdf*
- Media type prevalence is more spread

Textual Media types - Results

	Media type	% contents 2003	% contents 2008	Trend $\left(\frac{\%\text{ contents 2003} - \%\text{ contents 2008}}{\%\text{ contents 2003}}\right)\times 100$
1	Text/html	96.0%	93.9%	-2.1%
2	App'n/pdf	1.9%	3.0%	+57.6%
3	Text/plain	1.0%	1.6%	+58.5%
4	App'n/x-shockwave-flash	0.5%	1.2%	+115.8%
-	Other	0.7%	0.3%	-

- *Text/html* lost presence to other formats
- Increase in *app'n/pdf*, *text/plain*, *app'n/shockwave-flash*

Content size – Why?

- Estimate the storage resources required to create Web data repositories

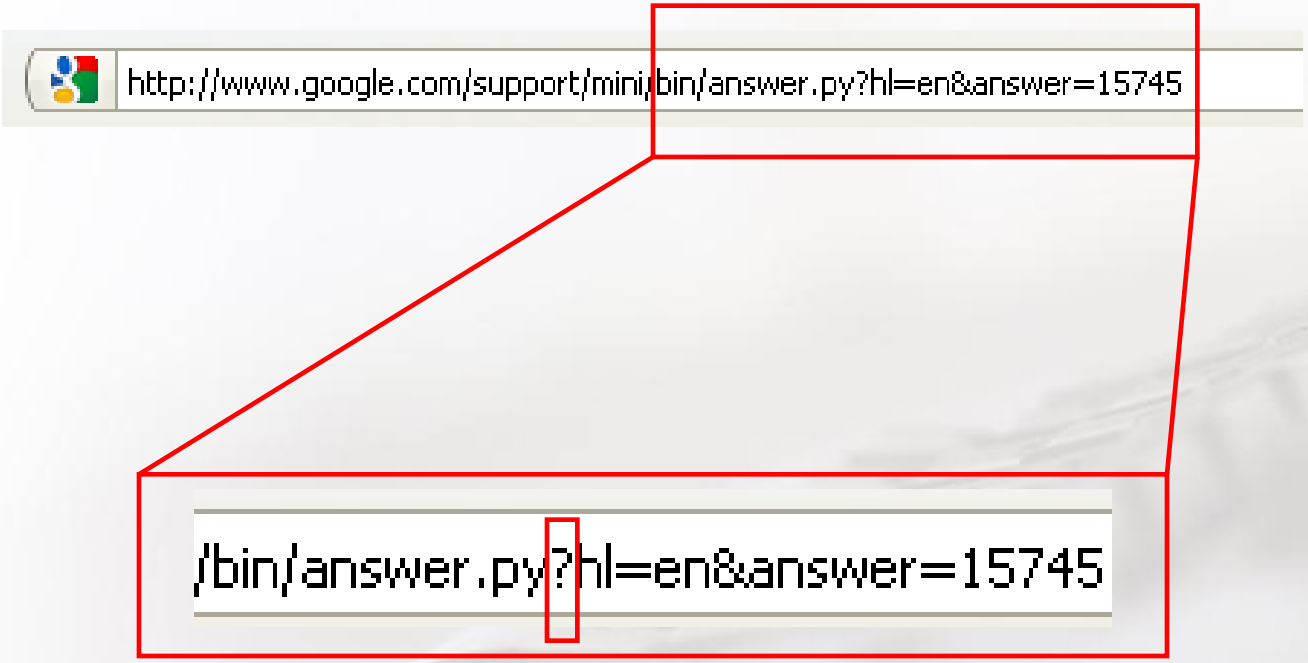
Content size - Results

Media type	Avg Size 2003	Avg Size 2008	Trend $\left(\frac{\text{Avg size 2003} - \text{Avg size 2008}}{\text{Avg size 2003}}\right) \times 100$
Text/html	21 KB	30 KB	+45.9%
App'n/pdf	207 KB	252 KB	+21.6%
Text/plain	11 KB	44 KB	+58.5%
App'n/x-shockwave-flash	44 KB	90 KB	+115.8%
…			
powerpoint	1055 KB	500 KB	-52.6%
Text/rtf	476 KB	143 KB	-70.0%

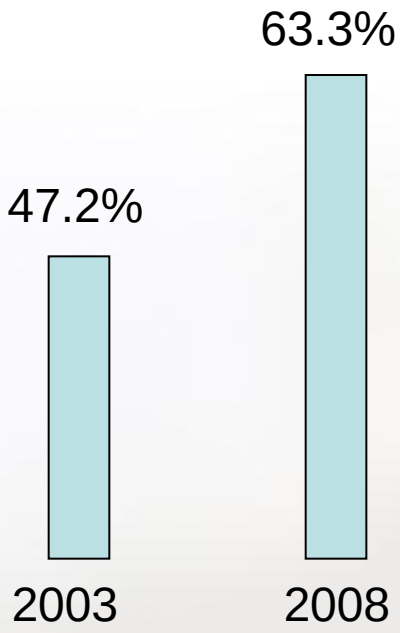
- Content size grew but for some types decreased

Dynamically generated contents – Why?

- Identify technological trends in Web publishing



Dynamically generated contents - Results

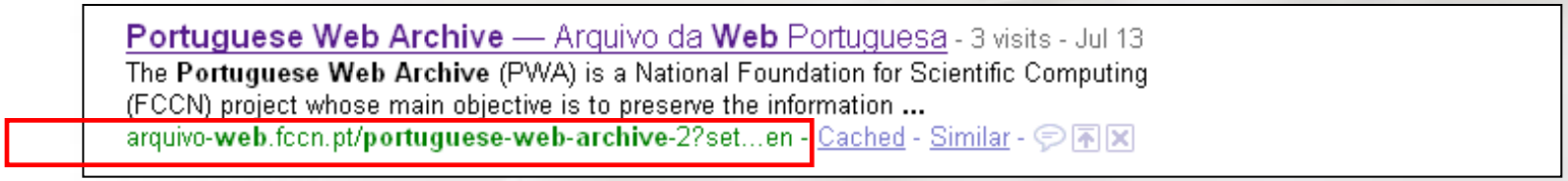


- The percentage of URLs containing parameters increased

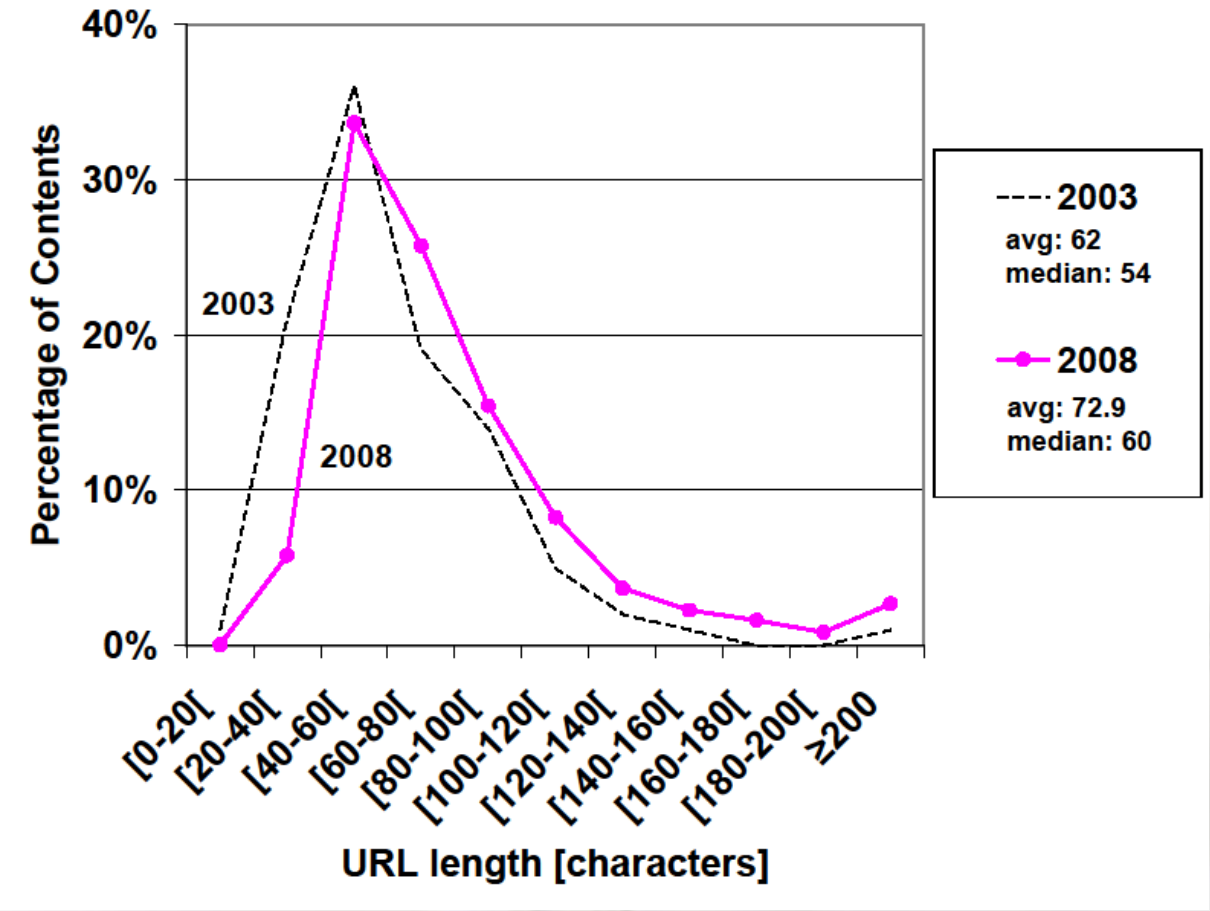
URL length – Why?

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

- How many characters should be presented on a search engine results page



URL length - Results



- 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

Conclusions II

- 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

Collaborate with us

- Testing of developed systems
- Research and development projects
- Crawl logs available for research purposes

http://archive.pt

Thank you.

http://archive.pt