How Are Web Characteristics Evolving?

João Miranda, Daniel Gomes, {joao.miranda,daniel.gomes}@fccn.pt
Foundation for National Scientific Computing, Lisbon, Portugal
www.arquivo.pt

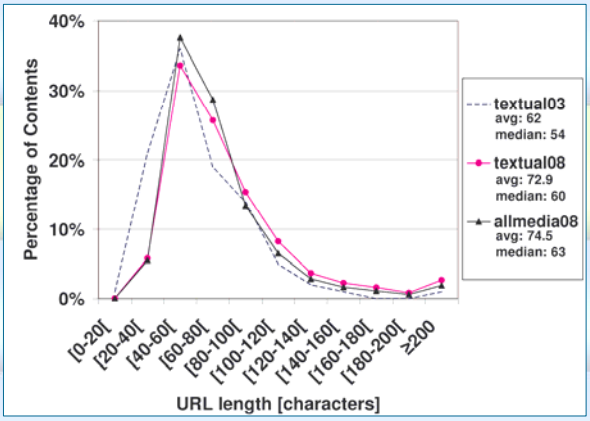
Motivation

- Web formats and tools are permanently evolving
- Web characteristics affect software design, such as browsers, search engines or Web archives
- The Portuguese Web Archive periodically harvests and stores Web data
- This study contributes with quantitative measurements about the evolution of Web characteristics

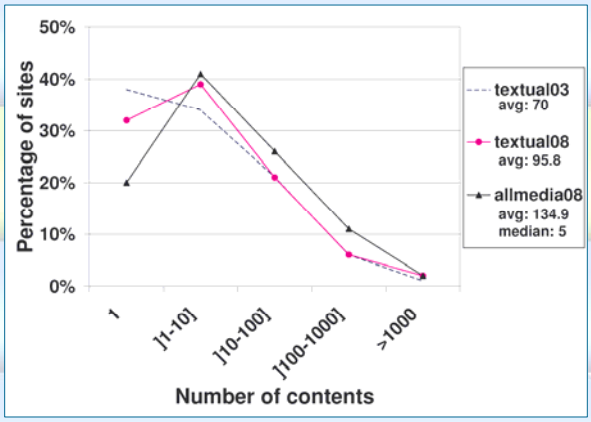
Methodology

- Performed comparisons between the data sets to derive evolution trends
- Portuguese Web data:
  - 2003: 3 million textual contents (textual03)
  - 2005: 4 million all media contents (allmedia05)
  - 2008: 50 million all media contents (allmedia08)

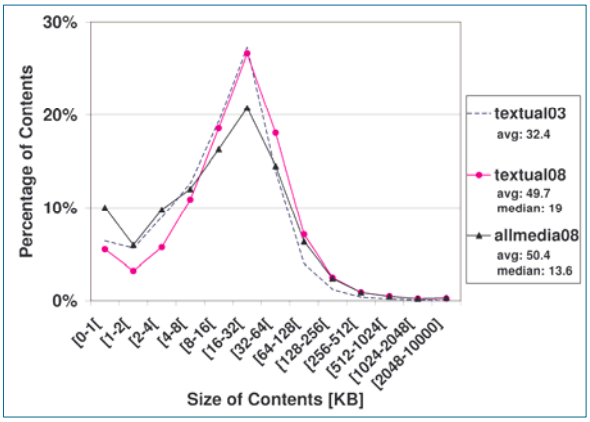
Results



• URL length has increased.



• Sites continue to be typically small but their number of contents grew.



• The average content size increased but it maintained a similar distribution, 73% of the contents have sizes between 4 KB and 64 KB.

Media type	% contents textual03	% contents textual08	Trend
text/html	95.9702%	93.9178%	-2.1%
app'n/pdf	1.9208%	3.0274%	+57.6%
text/plain	1.0229%	1.6207%	+58.5%
app'n /x-shockwave-flash	0.5440%	1.1737%	+115.8%
app'n/msword	0.4332%	0.1803%	-58.4%
powerpoint	0.0644%	0.0299%	-53.6%
excel	0.0283%	0.0438%	+55.0%
text/rtf	0.0069%	0.0010%	-85.2%
app'n/rtf	0.0060%	0.0024%	-59.5%
app'n/x-tex	0.0020%	0.0021%	+2.5%
text/tab-separated-values	0.0013%	0.0007%	-45.3%
text/richtext	0.0001%	0.0000%	-40.7%

• HTML is the dominant textual format. PDF and Flash formats tend to gain popularity.

Conclusions

- Web publishing behaviors changed radically in the past years but macro-scale Web characteristics present small changes
- There are hundreds of formats available on the Web but only a very small set has a significant presence
- Content characteristics vary significantly according to media type
- The number of contents hosted per site tends to increase and most servers host a single site