Introducing the Portuguese web archive initiative

Daniel Gomes, André Nogueira, João Miranda, Miguel Costa
wg-tomba@fccn.pt
Fundação para a Computação Científica Nacional
(National Foundation for Scientific Computing)

Objectives

- Research & development project
  - Officially begun in January, 2008
- Begin preserving Portuguese web contents
- Public access to archived data
- Resources for research
  - Web data sets, web characterizations

Challenges

- “Good” coverage of the Portuguese web
  - Not to broad, not to narrow
- Provide *efficient* search mechanisms
  - Web search engines are not web archives
  - Temporal web information retrieval is in its infancy
  - Search mechanisms may be language dependent
    - Most are evaluated over English texts

Outline

- Introduction
- Previous work
- Work in progress
  - The Portuguese Web Archive system
  - New contributions
- Future work
- Conclusions

Digital Deposit (2001)

- Digital Deposit (2001)
  - University of Lisbon and National Library of Portugal
  - Preservation of selected online publications
- Configure periodic crawls and verify quality.
- 800 000 contents
  - lost ☹

Tomba: 1st Portuguese web archive prototype (2006)

- Research scope
  - University of Lisbon
  - 2 researchers in the PWA
- Web archiving tools and selection criteria
- URL search interface
  - Internally developed SW
- Textual contents (2002-2006)
  - 54 M; 1,5 TB
  - Conversion to ARC and integration in the PWA

Work in progress

The Portuguese web archive system

Adopted Archive-access tools

- “In-house” technology was abandoned
  - Previous work facilitated learning
- Heritrix 1.12
  - enough for the PT web, took 1 week with 1 machine
- ARC format
- NutchWAX 0.11 and Wayback 1.2.1
  - Hadoop, Lucene

Other software

- PostgreSQL: extract characterizations based on crawl.log
  - Table with ~63M performs well
- Research on temporal search
  - Weka: collection of machine learning algorithms for data mining
  - SVMlight: Support Vector Machines library
- Plone: Multi-lingual Content Management System

Hardware

- Blade system
  - 7 blades: 2 x Quad-core, 8 GB
  - Room for 16 blades
- Storage
  - SAN: 25.6 TB in RAID 5
  - FATA and Fibre Channel disks
  - Tape library: 12 TB

Hardware choice: pros and cons

- Pros
  - Management tools
  - Saves on physical space
  - Full support
- Cons
  - Expensive
  - Hidden licensing costs
  - Centralized

Selection criteria for Portuguese web

- Sites under .PT and embedded contents from other domains
  - Loosing half of the Portuguese web
- A new crawl every 3 months
- Future: all contents written in the Portuguese language
  - Search algorithms can be tuned for Portuguese
  - Crawl and identify language in pages spread across the web

1st characterization of the PT web

Metric	Volume
URLs visited	72 million
Sites visited	455 thousand
Contents crawled	56 million
Downloaded data	2.8 TB
Archived data in compressed format	2 TB

Table 1: Visited resources and volume of harvested information from the Portuguese web.

HTTP code	Nr. URLs	%	Description
200	56 046 288	85.2%	OK
302	4 305 265	6.5%	Temporary redirection
404	3 669 855	5.6%	Not found
301	789 133	1.2%	Permanent redirection
500	325 225	0.5%	Internal server error
400	266 318	0.4%	Bad request
403	164 241	0.2%	Access forbidden
303	124 385	0.2%	Redirection to other resource
401	48 334	0.1%	Unauthorized
Others	36 136	0.1%	–

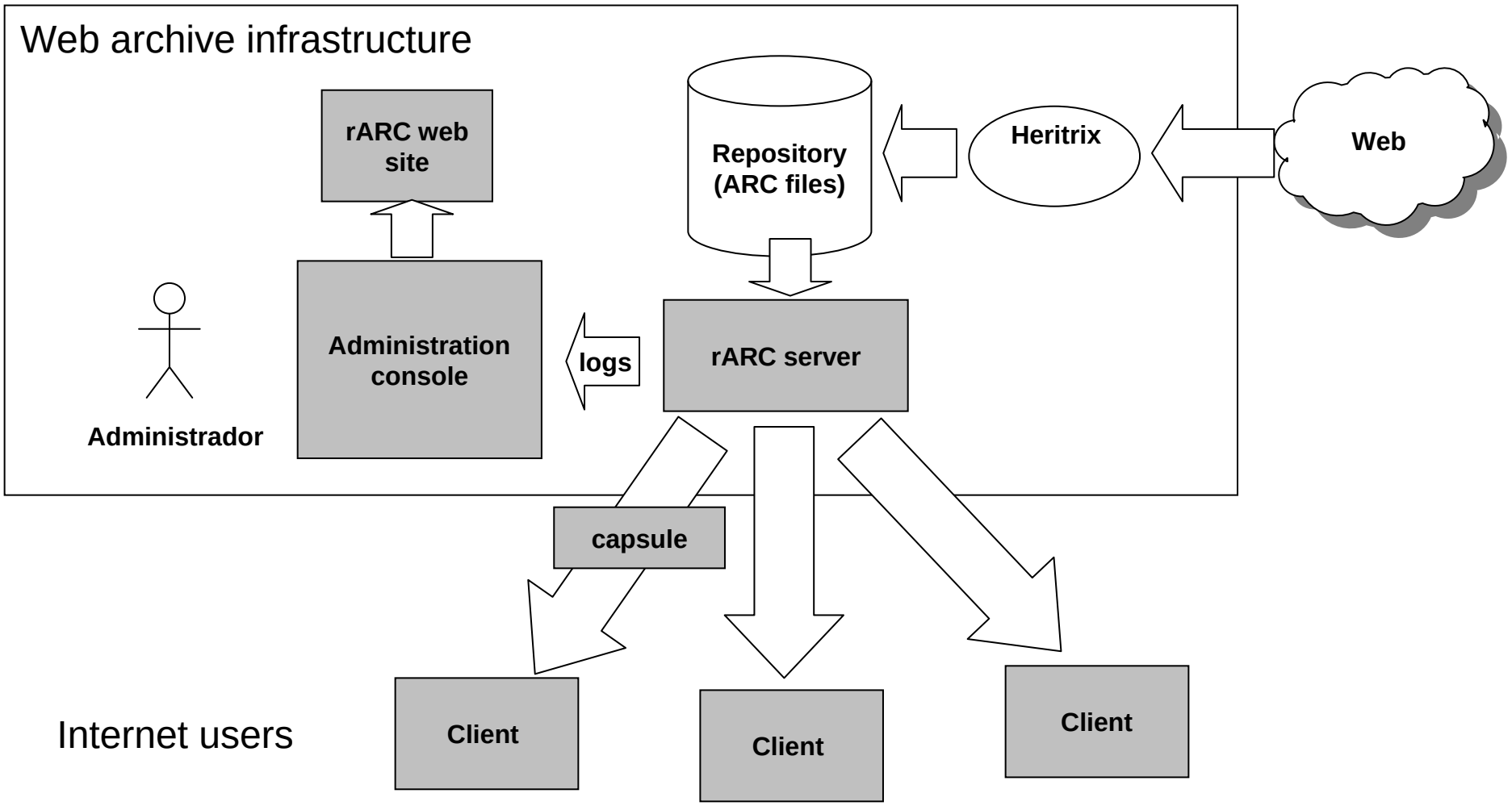
Table 2: HTTP response codes.

- A detailed study is in progress
- Analyze evolution across time

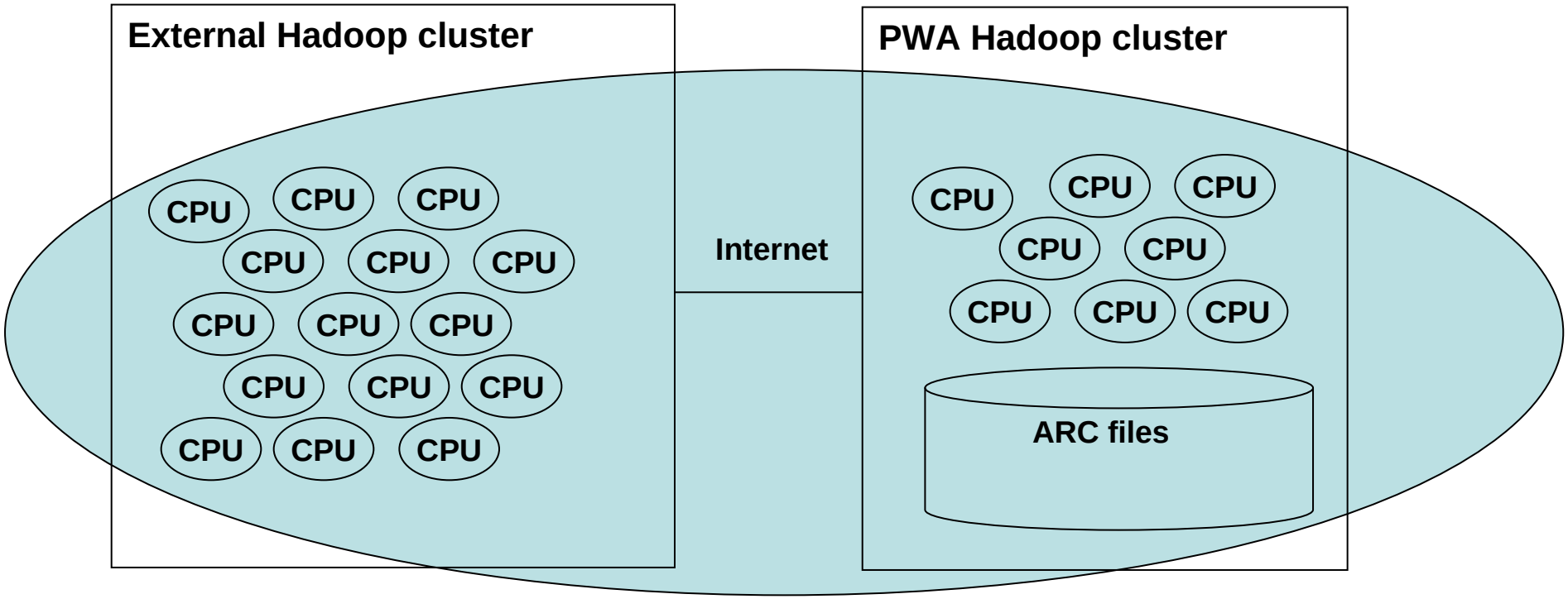
Contributions

- New tools for web archiving
  - rARC: ARC replicator
  - GAppA: Grid Appliance for the archive
  - WebClass: content classifier
- Research to improve temporal search

rARC: ARC Replicator



GAppA: Grid Appliance for Web Archives



- Access to archived contents
- Share computing resources

WebClass: automatic classification of archived contents

- Newspaper sections
  - Easy to gather training data sets
  - Users understand class labels
- Alternative search method: see all Sports news on a given day
- Help to disambiguate term searches
  - Figo query: Football player, fruit or International Federation of Gynecology and Obstetrics?

Improve temporal search

- Relevant results but fast responses
  - Optimize index structures
  - Select data to keep on memory
- Lack of temporal ranking algorithms
  - NutchWax OPIC didn’t yield good results

The strategy

- Get the most relevant results from each crawl
  - Try existing ranking algorithms
  - Evaluation through TREC test collections: web data set + query/results (GOV2)
- And merge them *efficiently*
  - Try new temporal ranking algorithms
  - How to evaluate? No temporal test collection!
  - Create test collection, analyze logs, user surveys

Future work

- Integrate and index Internet Archive collections gathered from .PT
  - 130M URLS
  - ~ 4TB of data
- Image search
  - NutchWax Image Indexer
- Migrate to NutchWax 0.14
  - Reproduce bug fixes and customizations
- Extend web characterizations
  - Content analysis: accessibility, format conformity levels

Conclusions

- Portuguese web archive
  - We are here
- We count on your help
- You can count on us
  - We want to contribute to the community efforts

ARQUIVO DA WEB PORTUGUESA

Thank you for your attention contact us:

daniel.gomes@fccn.pt

wg-tomba@fccn.pt

http://arquivo-web.fccn.pt (also in English)