Web Modelling for Web Warehouse Design

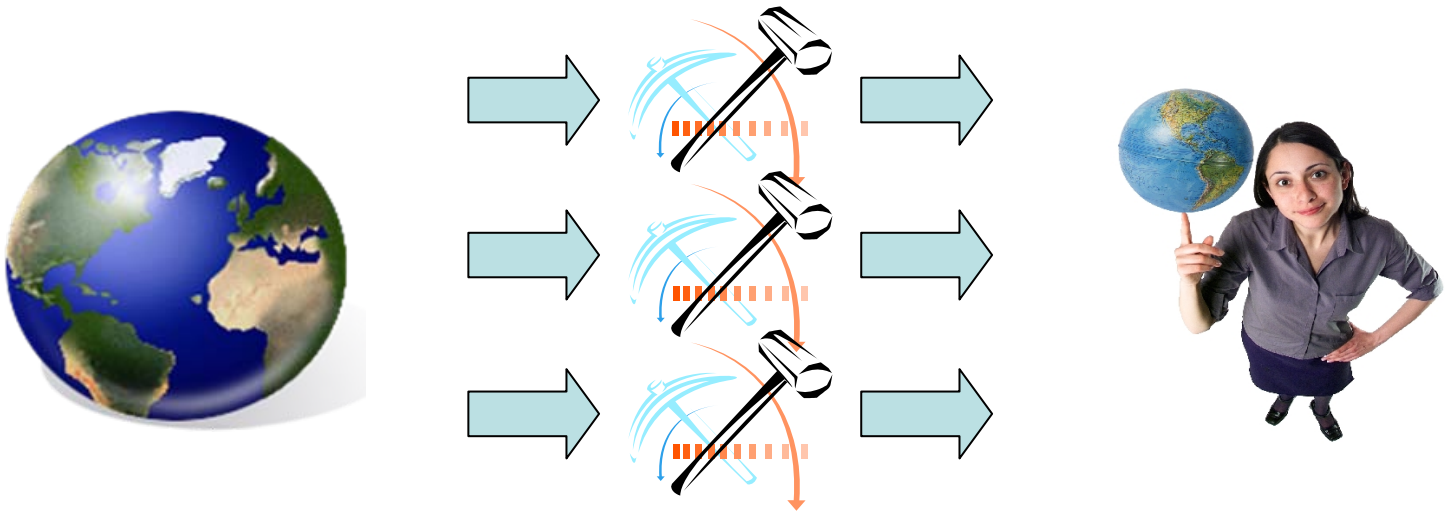
Daniel Coelho Gomes

Doutoramento em Informática

Especialidade em Engenharia Informática

19 de Março de 2007

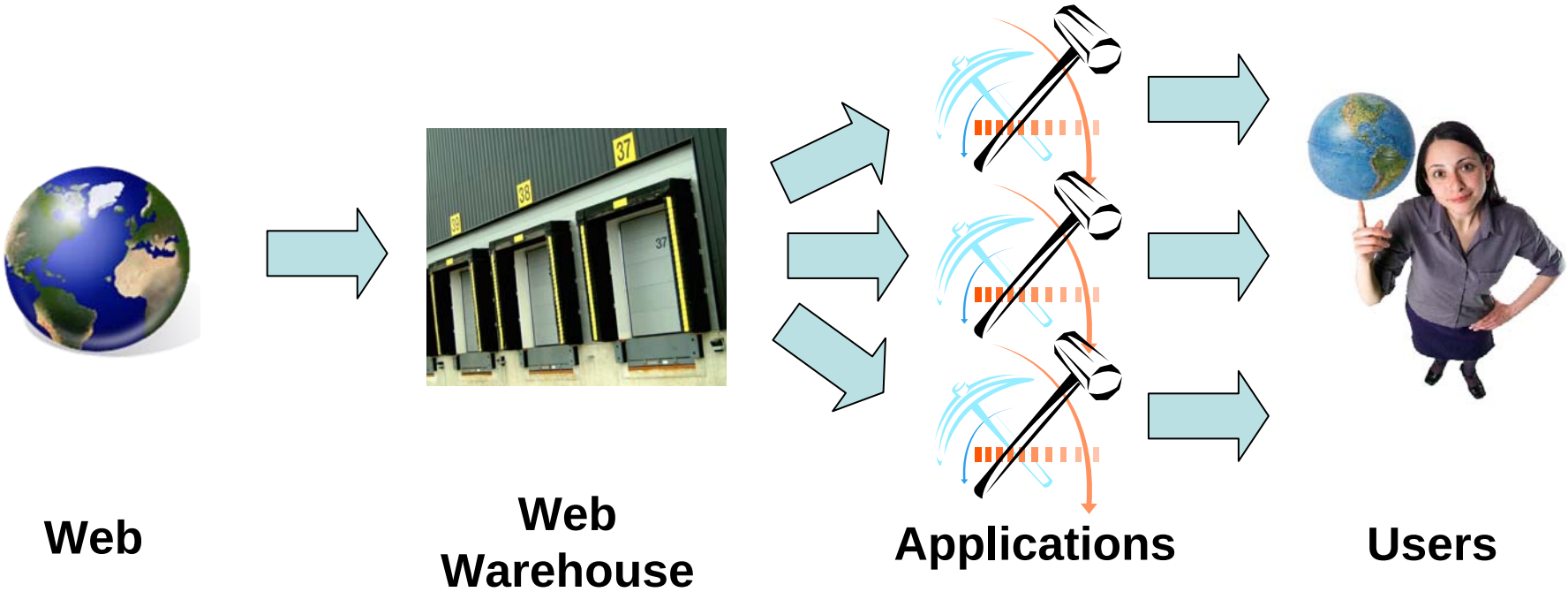
Harnessing the Web



Web **Applications** **Users**

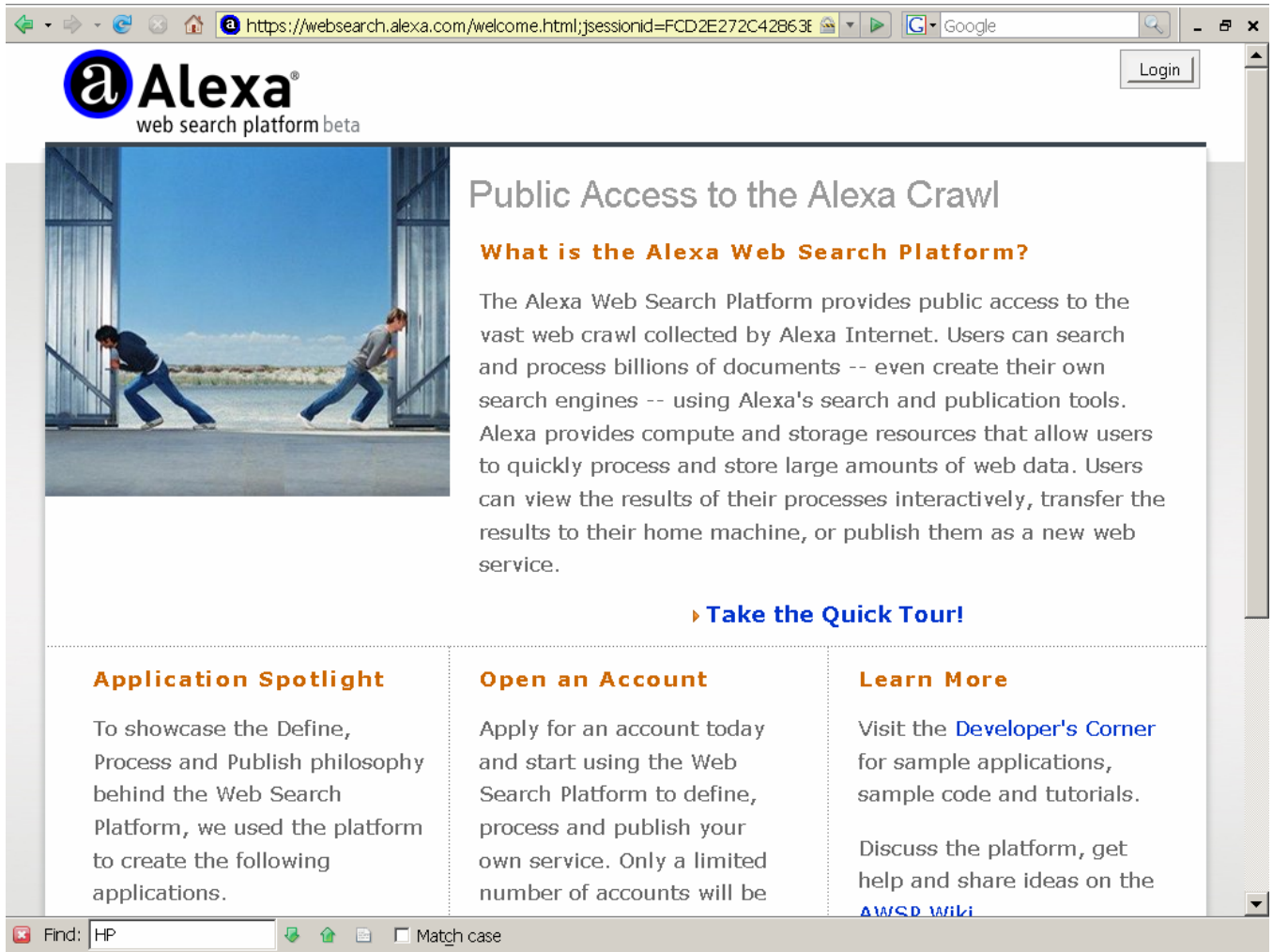
- The web is the largest source of information
- Users need applications to extract knowledge from web data
- Each application has to manage its own data

The need for Web Warehouses

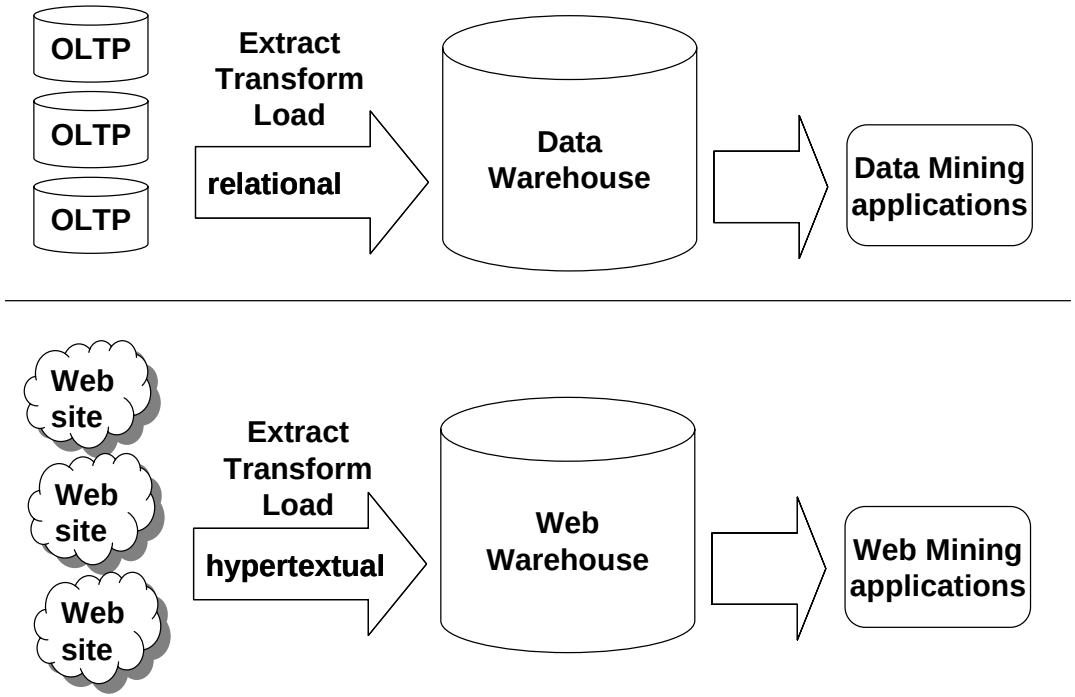


- A WWh releases applications from data management
  - Applications focus on their purposes
- Enables web data reuse

Web Warehousing supports mining applications



Web vs. Data Warehousing



- Must know data to design a warehouse
- The Web does not follow a relational model
- Web data models are required

What is a web model?

- A Web model describes the characteristics of a web portion
  - Distribution of sites per Top-Level Domain
  - Content media types
  - Incoming links per URL

What is a web portion?

- A WWh must be populated with contents relevant to its users
- A web portion is the set of relevant web contents selected to be warehoused
- The Portuguese web
  - Empirical definition: contents relevant to the Portuguese community
  - Formal definition:
    - Contents under the .PT domain
    - Contents outside .PT in Portuguese and linked from .PT

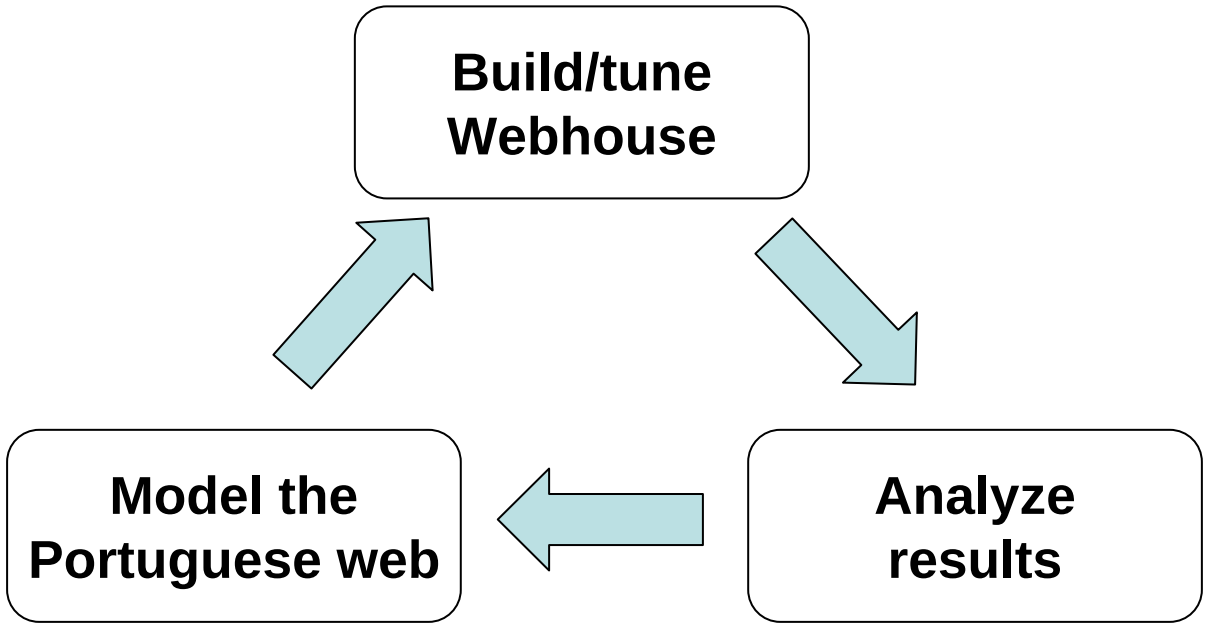
Outline

- Motivation
- Objectives and methodology
- Contributions
- Conclusions
- Future Work

Research questions

1. Which features should be considered in a web model?
2. How can the boundaries of a web portion be defined?
3. What can bias a web model?
4. How persistent is information on the web?
5. How do web characteristics influence Web Warehouse design?

Experimental methodology

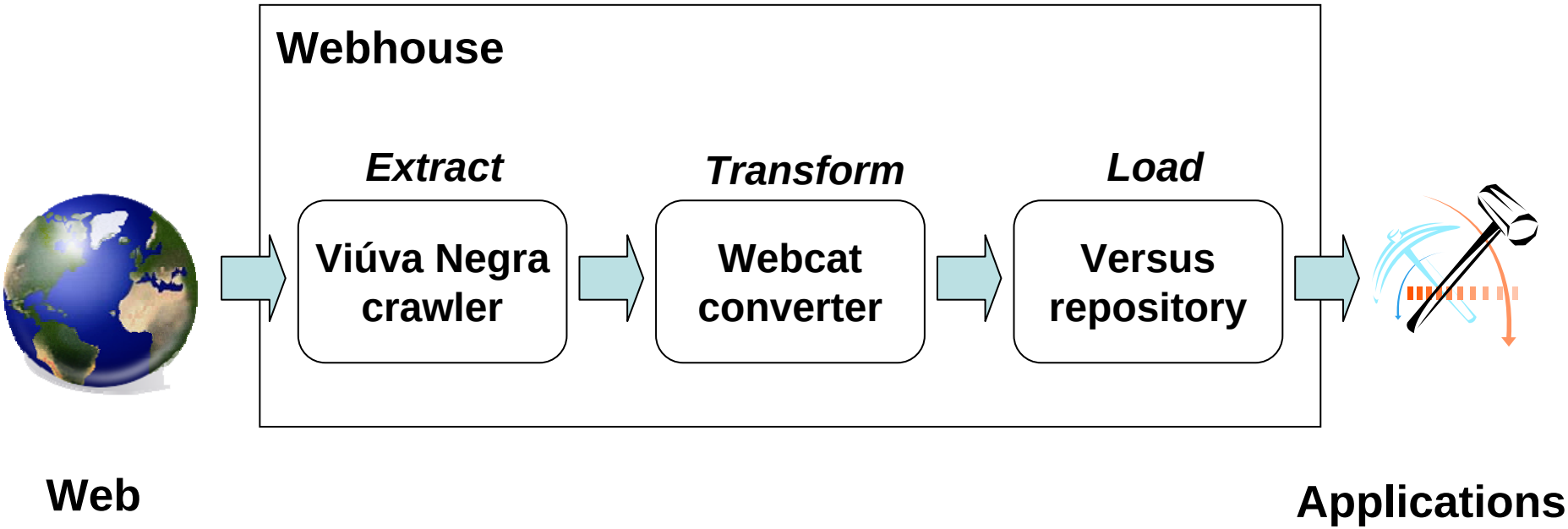


- Successive versions of Webhouse enabled the identification of the influence of web characteristics in its design

Why the Portuguese Web?

- General models of the Web may not be representative of the data to be warehoused
  - The Portuguese Web can be exhaustively harvested and accurately modelled
  - Still provides a general model of web data because it contains several publication genres
  - The Portuguese Web is relevant to a significant amount of users (10M)

Webhouse architecture



Outline

- Motivation
- Objectives and methodology

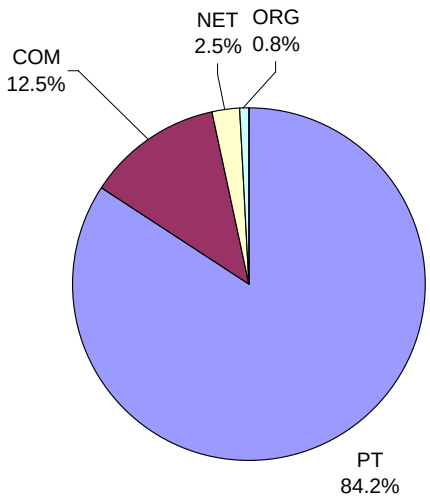
- Contributions
- Conclusions
- Future Work

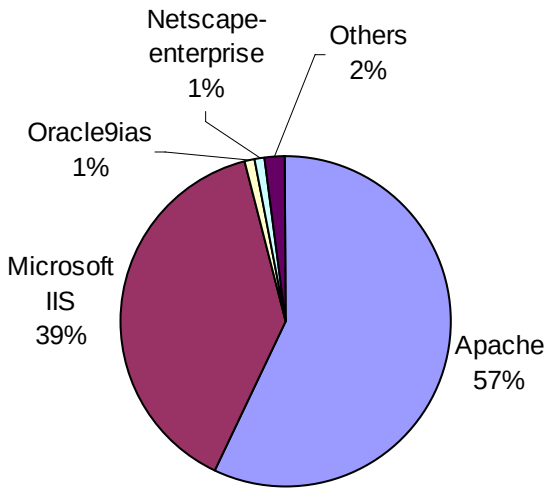
Innovation of this research

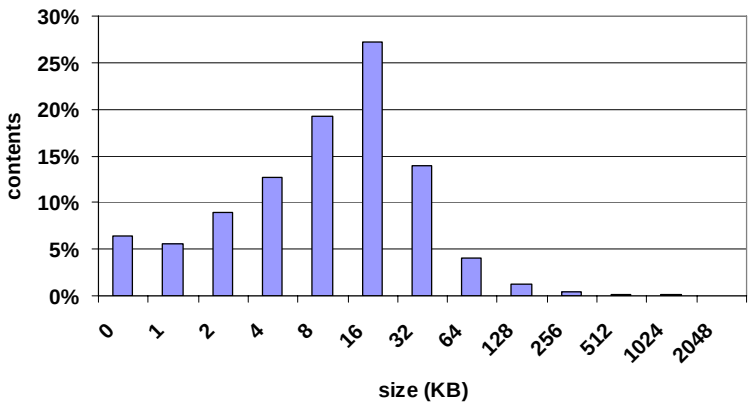
- Includes web modelling in the web data integration process
  - Web Warehousing has been done assuming that the data sources were well known
- Studies the influence of web characteristics in the several stages of web data integration
  - From extraction to access
- Combines knowledge from different research domains
  - Web Characterization: monitors and models the web
  - Web Crawling: automatic extraction of web data
  - Web Warehousing: web data integration

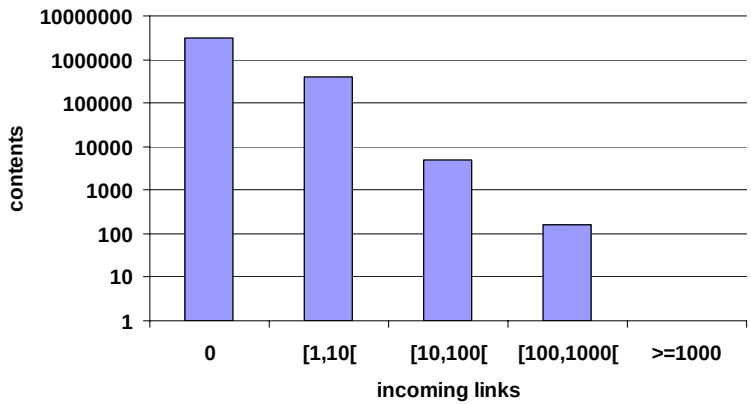
Web Characterization

A model of the Portuguese web

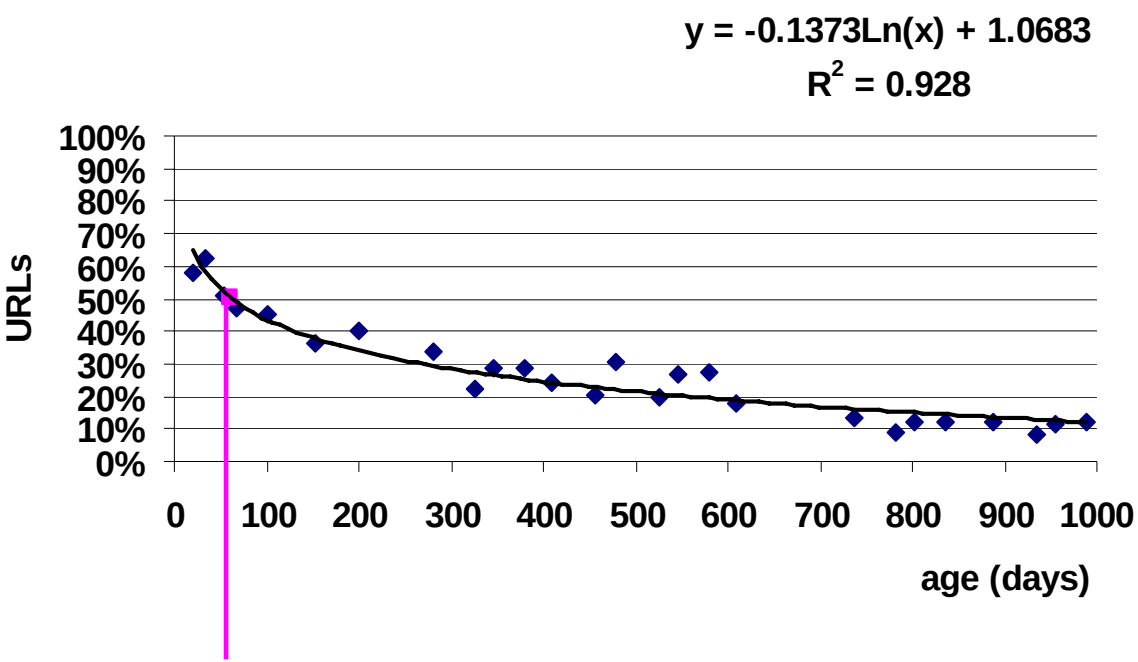








Models for estimating web data persistence



half-life=61 days

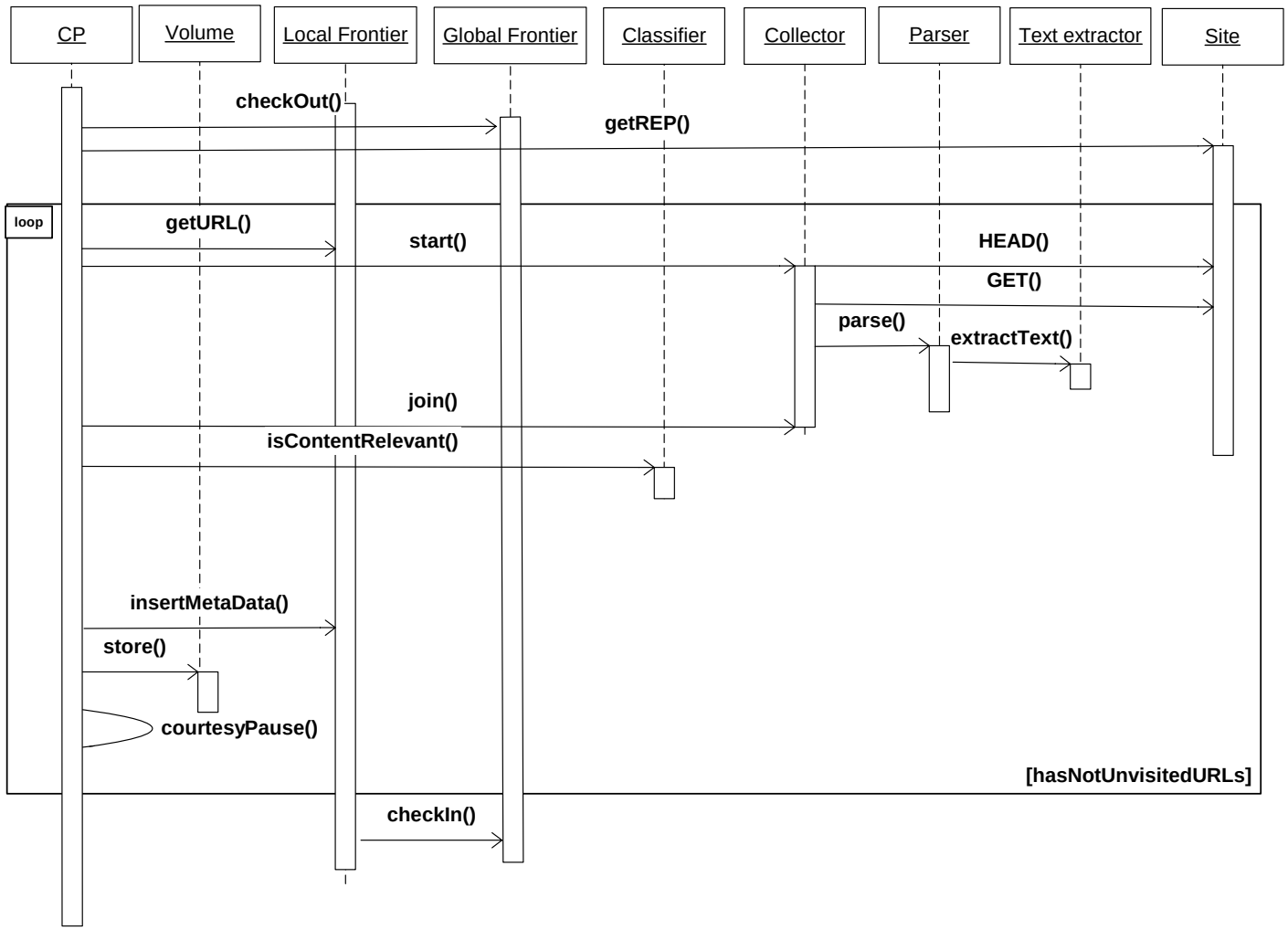
- URL transience is much more problematic in WWh than in "book marking"
- In 2 months 50% of the URLs in a data set are no longer valid

Comparison with other studies on URL persistence

Study	Results	My estimation	Comparison
Koehler (2002)	**50%**	**17%**	**>**
Cho (2000)	**70%**	**60%**	**>**
Fetterly (2003)	**88%**	**47%**	**>**
Ntoulas (2004)	**20%**	**26%**	**~**

Web Crawling

Crawling algorithms and techniques

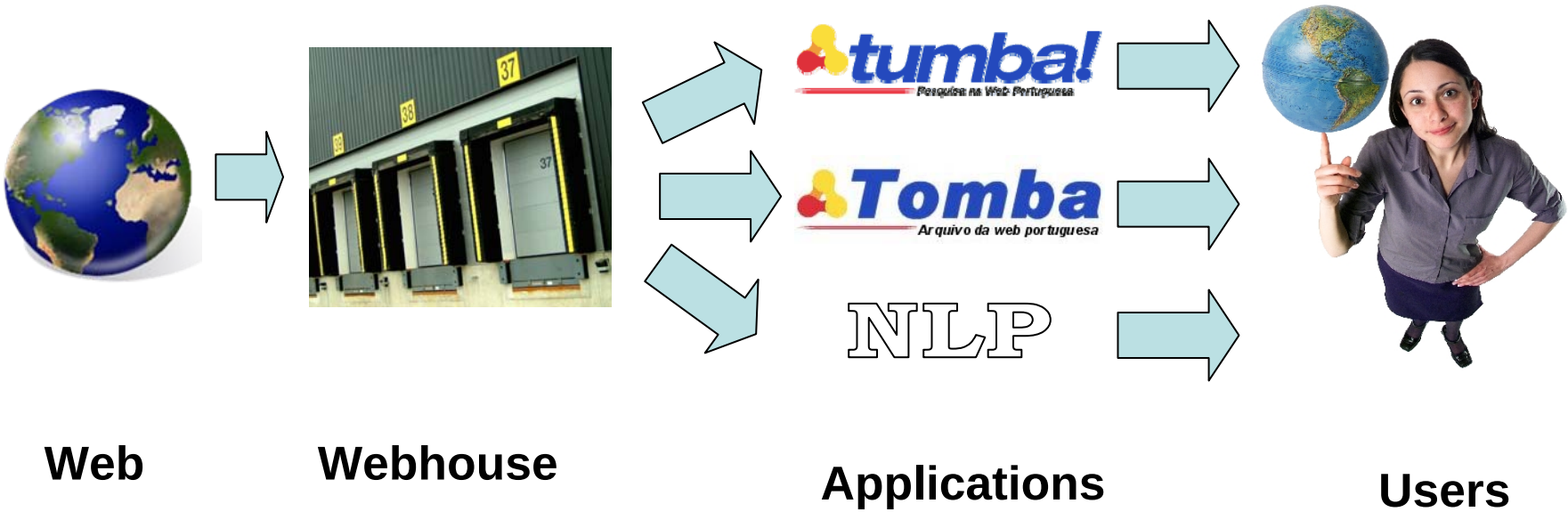


Coping with hazardous situations

- Documentation and solutions to address hazardous situations to crawling
- Spider traps
  - Infinite sites
- Duphosts
  - Sites with different names that provide the same content
  - tucows.com, www.tucows.com, tucows.ip.pt
  - Waste of WWh resources

Web Warehousing

Applications of Webhouse



Answers

1. Which features should be considered in a web model?
   - Vary according to application requirements
   - Site, content, link structure and data persistence
2. How can the boundaries of a web portion be defined?
   - Automatic harvesting policy
   - Domain restrictions and content classification

Answers

3. What can bias a web model?
   - Hazardous situations
   - Sampling methodology must emulate extraction stage
4. How persistent is information on the web?
   - The web is getting more transient but there is also persistent data
5. How do web characteristics influence Web Warehouse design?
   - Extraction stage
   - Storage requirements
   - Schedule maintenance operations

Future work

- Is a model of the Portuguese web representative of other web portions?
  - Differences due to sampling methods and dates?
  - Crawl different portions in parallel and compare models
- Web warehousing research is crucial to deploy large-scale web archiving
  - How to search among historical web collections?

Main publications

- Journals
  - Daniel Gomes and Mário J. Silva, *The Viúva Negra crawler: an experience report*, Software: Practice and Experience, Wiley InterScience (accepted for publication);
  - Daniel Gomes and Mário J. Silva, *Characterizing a national community web*, Transactions on Internet Technology, ACM, 2005.
- Conferences
  - Daniel Gomes, Sérgio Freitas, Mário J. Silva, *Design and Selection Criteria for a National Web*, ECDL'06 (best paper by young researcher);
  - Daniel Gomes, Mário J. Silva, *Modelling information persistence on the web*, ICWE'06 (best paper candidate);
  - Daniel Gomes, André Santos, Mário J. Silva, *Managing duplicates in a web archive*, SAC'06.

Thank you for your attention