Understanding the Information Needs of Web Archive Users

Miguel Costa, Mário J. Silva
LaSIGE @ Faculty of Sciences, University of Lisbon
Foundation for National Scientific Computing

IWAW2010, Vienna, Austria

MCTES
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

UMIC
Agência para a Sociedade do Conhecimento

POS_CONHECIMENTO
Programa Operacional Sociedade do Conhecimento

União Europeia
FEDER

Problem

What do web archive users' need?

Applications

- Design technology that satisfies the users.
  - provide what the users want

- Tailor full-text search results toward the users’ goal.
  - return more effective results

Outline

- Introduction
- Methodology
- Results
- Conclusions

Methodology

Portuguese Web Archive

- ~150M documents:
  - searchable by full-text and URL.
  - range between 1996 and 2009.

  (currently, 182M of 779M archived)
- Available since 2010.

http://archive.pt

1st: Search Logs

```
213.22.91.10 [03/Feb/2010:21:15:27] QUERY mining
213.22.91.10 [03/Feb/2010:21:15:35] QUERY data mining
213.22.91.10 [03/Feb/2010:21:15:42] CLICK   data mining RANK=2
```

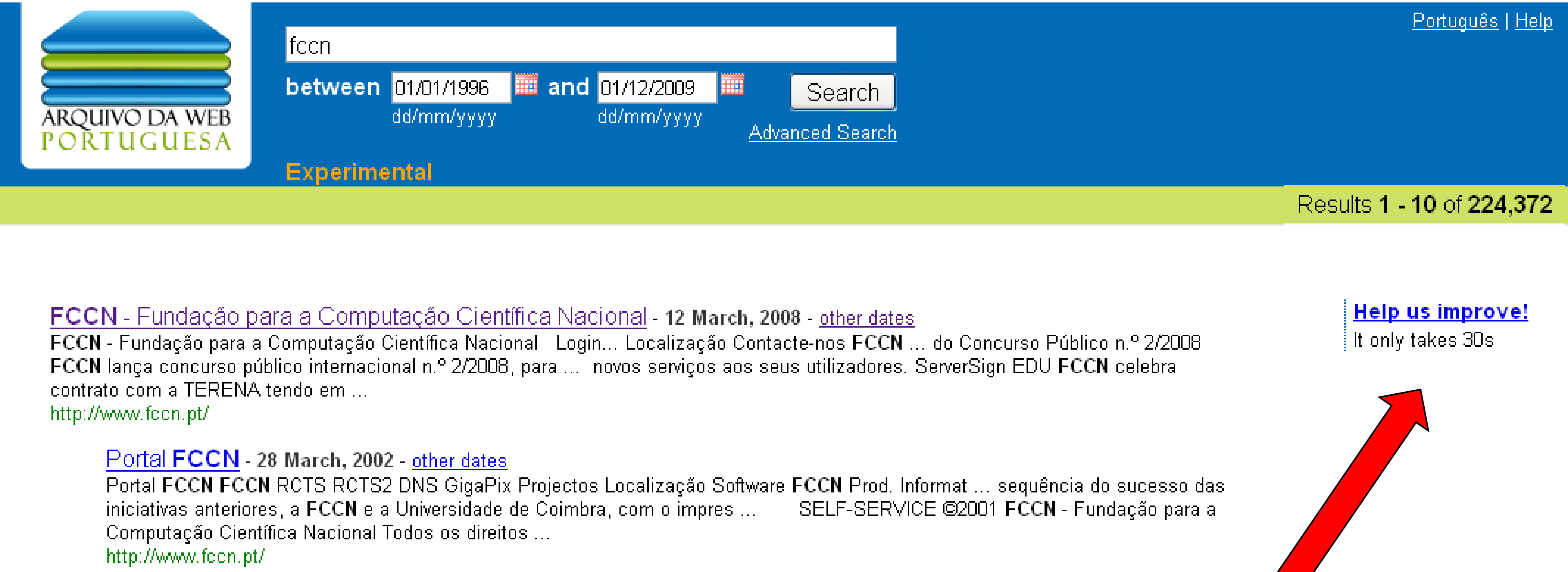
```
234.67.61.32 [03/Feb/2010:21:16:11] QUERY ford focus
234.67.61.32 [03/Feb/2010:21:16:19] CLICK   ford focus RANK=1
```

```
234.67.61.32 [03/Feb/2010:22:01:32] QUERY fccn
234.67.61.32 [03/Feb/2010:22:01:40] CLICK   fccn RANK=3
```

1st: Search Logs

- Procedure
  - cleansing
  - session delimitation
  - session classification (400 random sessions)
    - 2 evaluators
    - taxonomies: information needs & topics
- Users
  - anonymous
  - 81% of IP addresses are assigned to Portugal
  - 94% of interactions through the PT language interface

2nd: Interactive Questionnaire



fccn

between 01/01/1996 and 01/12/2009
dd/mm/yyyy dd/mm/yyyy

Results **1 - 10** of **224,372**

FCCN - Fundação para a Computação Científica Nacional - 12 March, 2008 - other dates
FCCN - Fundação para a Computação Científica Nacional Login... Localização Contacte-nos **FCCN** ... do Concurso Público n.º 2/2008 **FCCN** lança concurso público internacional n.º 2/2008, para ... novos serviços aos seus utilizadores. ServerSign EDU **FCCN** celebra contrato com a TERENA tendo em ...
http://www.fccn.pt/

Portal **FCCN** - 28 March, 2002 - other dates
Portal **FCCN** **FCCN** RCTS RCTS2 DNS GigaPix Projectos Localização Software **FCCN** Prod. Informat ... sequência do sucesso das iniciativas anteriores, a **FCCN** e a Universidade de Coimbra, com o impres ... SELF-SERVICE ©2001 **FCCN** - Fundação para a Computação Científica Nacional Todos os direitos ...
http://www.fccn.pt/

Help us improve!
It only takes 30s

Help us improve!
It only takes 30s

Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

2nd: Interactive Questionnaire

- Which of the following phrases describe best what you were doing?
  - Seeing how a web page or site, that I know, was in the past.
  - Collecting information about a subject written in the past.
  - Downloading an old file.
  - Recovering a web page or site that disappeared.
  - Seeing the evolution over time of a web page or site.
  - Seeing the evolution over time of the popularity of a subject.
- Were you searching between specific dates? Yes, No
- What other functionalities would you like our service to offer?
- Give examples of how our service could help in your profession or activities.
- Suggestion and critics:

2nd: Interactive Questionnaire

- Procedure
  - 2 pre-evaluation studies
  - 1st approach: URL in the results page
    - 0 responses! Design failed in captivating users.
  - 2nd approach: request via Facebook, Twitter, Email
    - 21 responses

- Participants
  - anonymous
  - 19 (2 excluded from 21) – 75 opened the URL

3rd: Laboratory Study

- Procedure
  - conducted individually
  - chose their own tasks based on their real needs
  - encouraged to follow a normal behavior
  - two researchers observed
  - interactions logged and recorded on video
  - post-questionnaire: same 5 questions

- Participants
  - 21 – 8 male, 13 female
  - different ages, professions, interests, academic degrees

Results

Information Needs Taxonomy

- **Navigational** – to reach a web page or site in mind
  (e.g. archive.pt)
- **Informational** – to collect information about a topic, usually from multiple pages without a specific one in mind
  (e.g. Iraq war)
- **Transactional** – to perform a web-mediated activity
  (e.g. downloading a file)

Broder. A taxonomy of Web Search. 2002

- Needs of web archive users focus on the **past**

Information Needs

Which of the phrases describe best what you were doing?	Taxonomy	Exp. #1 (logs)	Exp. #2 (quest.)	Exp. #3 (lab)
Seeing how a web page or site, that I know, was in the past	Navigational			
Seeing the evolution over time of a web page or site				
Collecting information about a subject written in the past	Informational			
Downloading an old file	Transactional			
Were you searching between specific dates?				
Yes				
No				

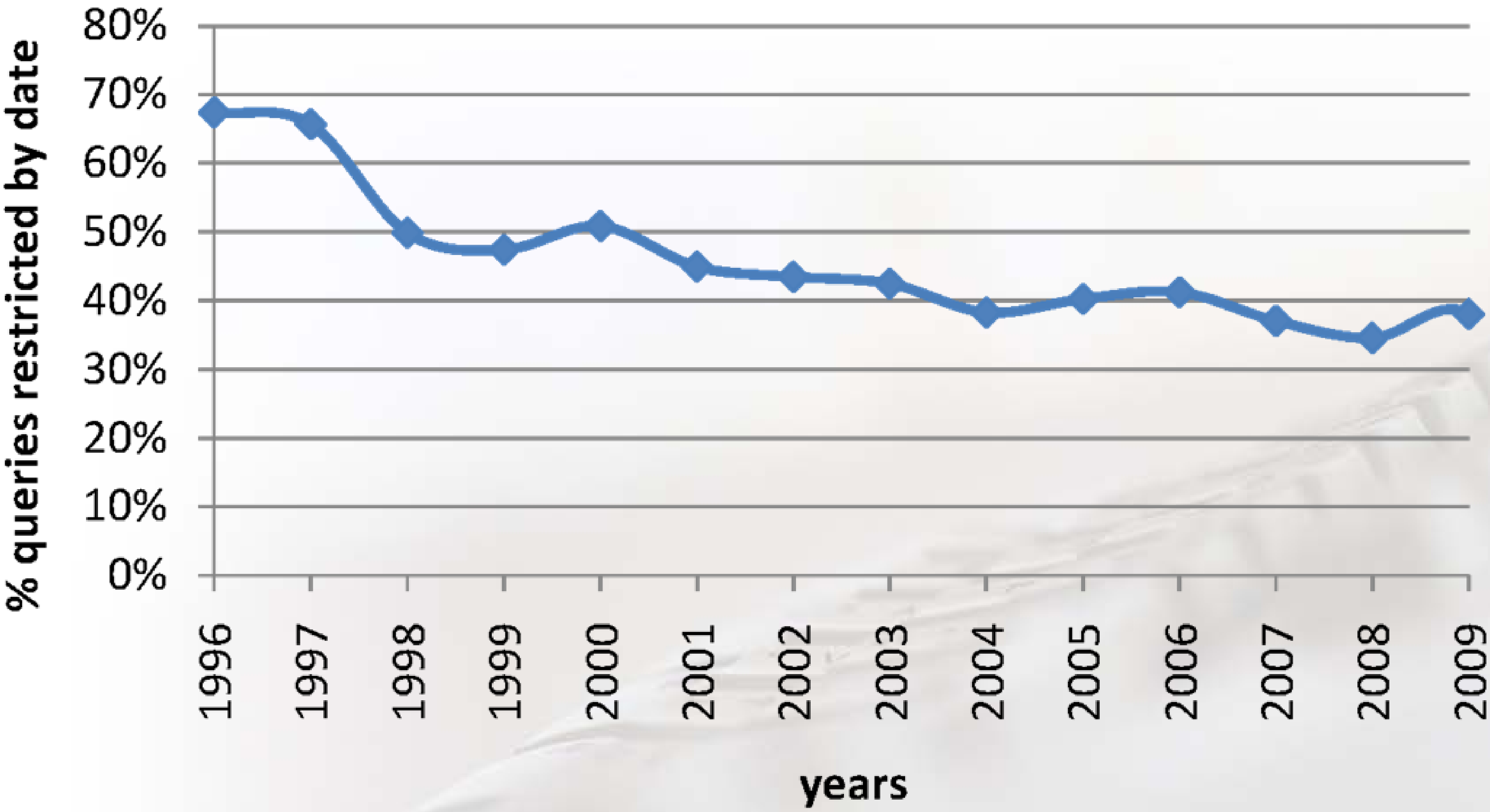
Information Needs

21% of the queries were URLs

Which of the phrases describe best what you were doing?	Taxonomy	Exp. #1 (logs)	Exp. #2 (quest.)	Exp. #3 (lab)
Seeing how a web page or site, that I know, was in the past	1st Navigational	47.7%	31.6%	47.6%
Seeing the evolution over time of a web page or site		9.2%	21.1%	33.3%
Collecting information about a subject written in the past	2nd Informational	37.8%	31.6%	14.3%
Downloading an old file	3rd Transactional	5.3%	10.5%	4.8%
Were you searching between specific dates?				
Yes		15.8%	47.4%	9.5%
No		84.2%	52.6%	90.5%

Search Logs Results

Users prefer the oldest documents:



Search Logs Results

Distribution of topics per navigational needs

Topic	%
Commerce	28.3
Computers or Internet	14.5
Education	14.5
Government	8.4
Entertainment	7.2
Sciences	6.0
Society	5.4
Things	3.0
Health	2.4
Sports	1.8
…	

Distribution of topics per informational needs

Topic	%
People	36.5
Health	14.8
Entertainment	9.6
Things	7.0
Sports	6.1
Places	4.4
Sciences	4.4
Education	3.5
Travel	2.6
Economy	2.6
…	

Exploratory Questions

- What other functionalities would you like our service to offer?
  - searching for images or photos
  - seeing fast the evolution of a web page or site
- Give examples of how our service could help in your profession or daily activity:
  - research old information – e.g. political events
  - satisfy curiosities – e.g. old photos

Conclusions

Conclusions

- Most users have navigational needs.

Conclusions

- Most users have navigational needs.
- Most users do not restrict searches by date.

Conclusions

- Most users have navigational needs.
- Most users do not restrict searches by date.
- Users prefer full-text search, but 21% of queries are URLs.

Conclusions

- Most users have navigational needs.
- Most users do not restrict searches by date.
- Users prefer full-text search, but 21% of queries are URLs.
- Users prefer the oldest documents.

Conclusions

- Most users have navigational needs.
- Most users do not restrict searches by date.
- Users prefer full-text search, but 21% of queries are URLs.
- Users prefer the oldest documents.
- Users search significantly about names (e.g. people).

Future Work

- **Validate** results with larger datasets.
- **Validate** results with other sources.
- **Validate** results throughout time.
- Use results to **improve** full-text search.
- Use results to **improve** interface.

Questions

Thank you.