Characterizing Search Behavior in Web Archives

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

TWAW2011, Hyderabad, India

MCTES
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

UMIC
Agência para a Sociedade do Conhecimento

POS_CONHECIMENTO
Programa Operacional Sociedade do Conhecimento

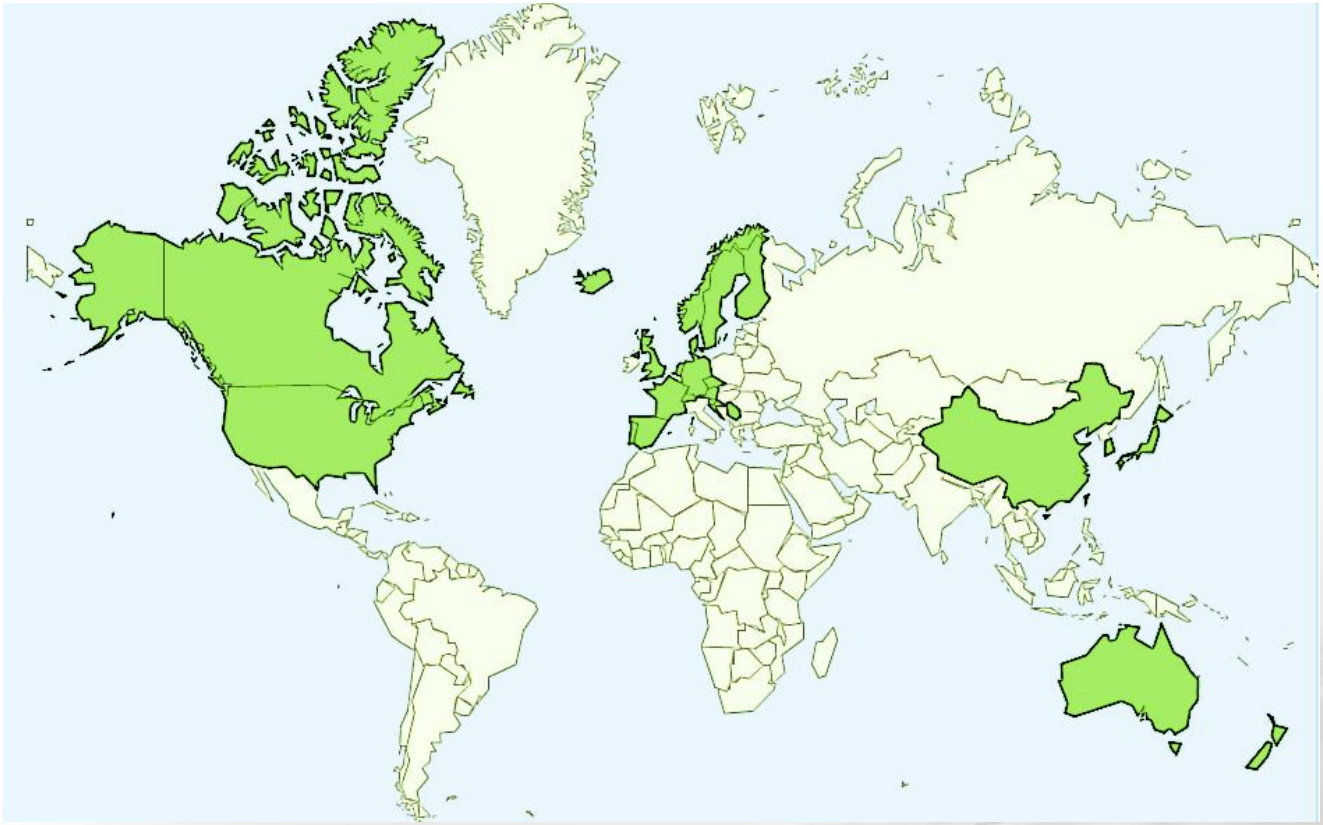
União Europeia
FEDER

Ephemeral Web

- The web contains unique and valuable information
  - news, interviews, opinions, feelings
- 80% of the web documents are unavailable after 1 year.

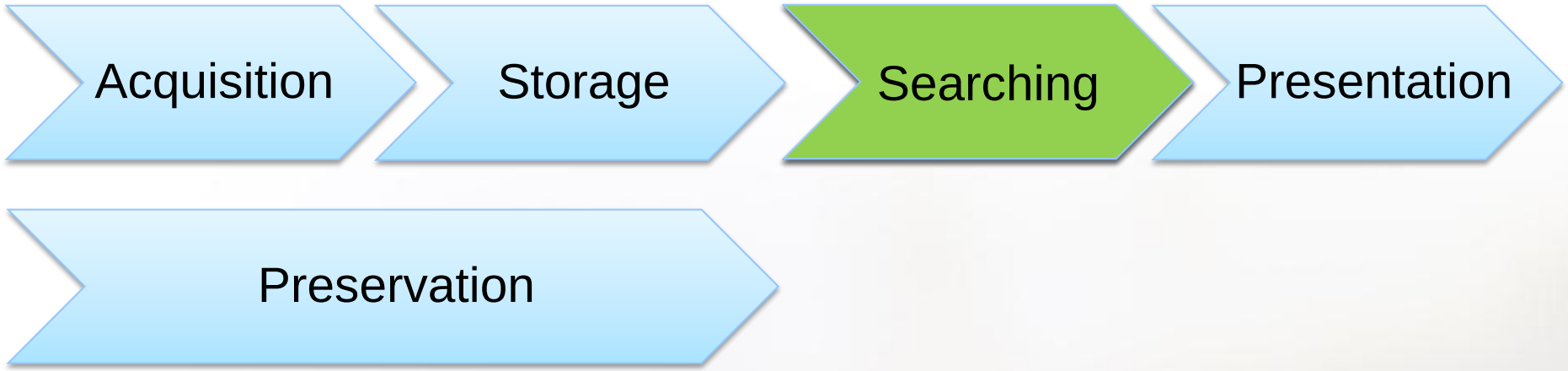
- Knowledge gap for future generations

Web Archiving Initiatives



- 42 web archiving initiatives in 26 countries.
- +180 billion documents archived since 1996.

Web Archiving Workflow



- Search technology based on web search engines
  - ignores the temporal dimension
  - doesn’t understand the end users

1st : Understanding Users

- **Why** do users search? (information needs)
- **What** do users search for? (topics)

- **How** do users search? (search behavior)
  - this study: 1st characterization

Predicting users' behavior can improve

- Response time
  - e.g. cache, special indexes
- Quality of results
  - e.g. better ranking, suggest queries
- Web design
  - e.g. make most used functionalities stand out

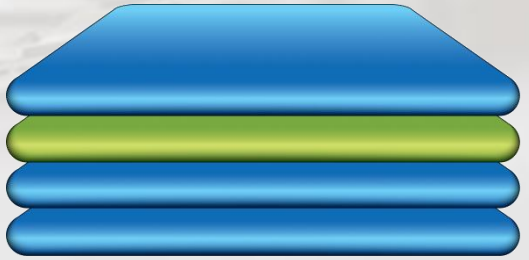
Portuguese Web Archive

- Archives the Portuguese Web ≈ .PT domain

- ≈ 182M documents:
  - searchable by full-text and URL.
  - range between 1996 and 2009.

- Search available since 2010.

http://archive.pt



Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

Interface: full-text search

Result Page

Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

Interface: URL search

same text box

430 Results

Did you want to find results containing the text: "http://www.fccn.pt" ?

Search Results between **1 January, 1996** and **5 February, 2011**

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
1 page	1 page	3 pages	7 pages	25 pages	12 pages	8 pages	7 pages	57 pages	116 pages	89 pages	102 pages	2 pages	0 pages	0 pages	---
13 October	10 December	15 February	16 January	1 March	18 January	28 March	10 February	21 January	6 January	1 January	1 January	12 March			Available soon
		3 December	25 January	2 March	2 February	3 June	6 June	15 April	7 January	6 January	2 January	12 March			
		12 December	28 January	10 May	7 February	20 July	12 June	9 May	12 January	15 January	11 January				
			22 February	10 May	24 February	2 August	9 August	26 May	16 January	18 January	16 January				
			17 April	20 May	1 March	27 September	18 October	6 June	20 January	18 January	21 January				
			23 April	20 May	2 March	29 September	23 October	11 June	22 January	27 January	26 January				
			28 April	28 May	1 April	2 October	24 November	12 June	29 January	2 February	27 January				

Version Page

Methodology

Search Log Analysis

- Pros
  - Large and varied
  - Less bias
  - Cheaper
  - Non-intrusive
  - Real information needs
- Cons
  - Lack of context
  - Lack of control

Dataset of Search Logs

- ≈ 10K sessions - 7 months of 2010

- Procedure
  - cleansing
    - normalized and excluded invalid sessions & queries
  - session delimitation
    - used IP, user session and a 30 minute gap

- Users
  - 72% of IP addresses → Portugal
  - 89% of interactions → PT language interface

How do users search?

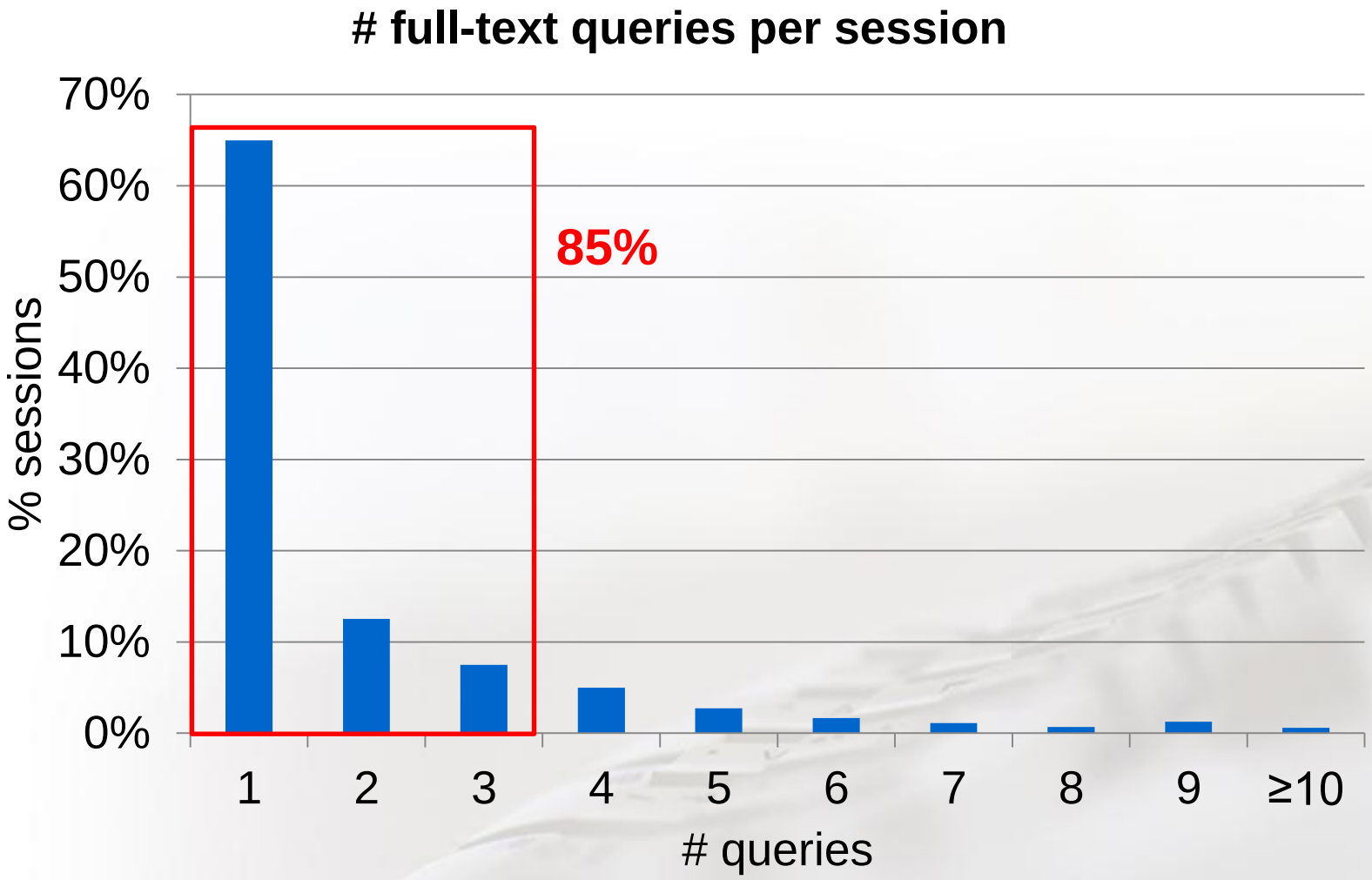
General Statistics

- Full-text sessions + URL sessions ≈ 90%
- Full-text sessions / URL sessions ≈ 2:1

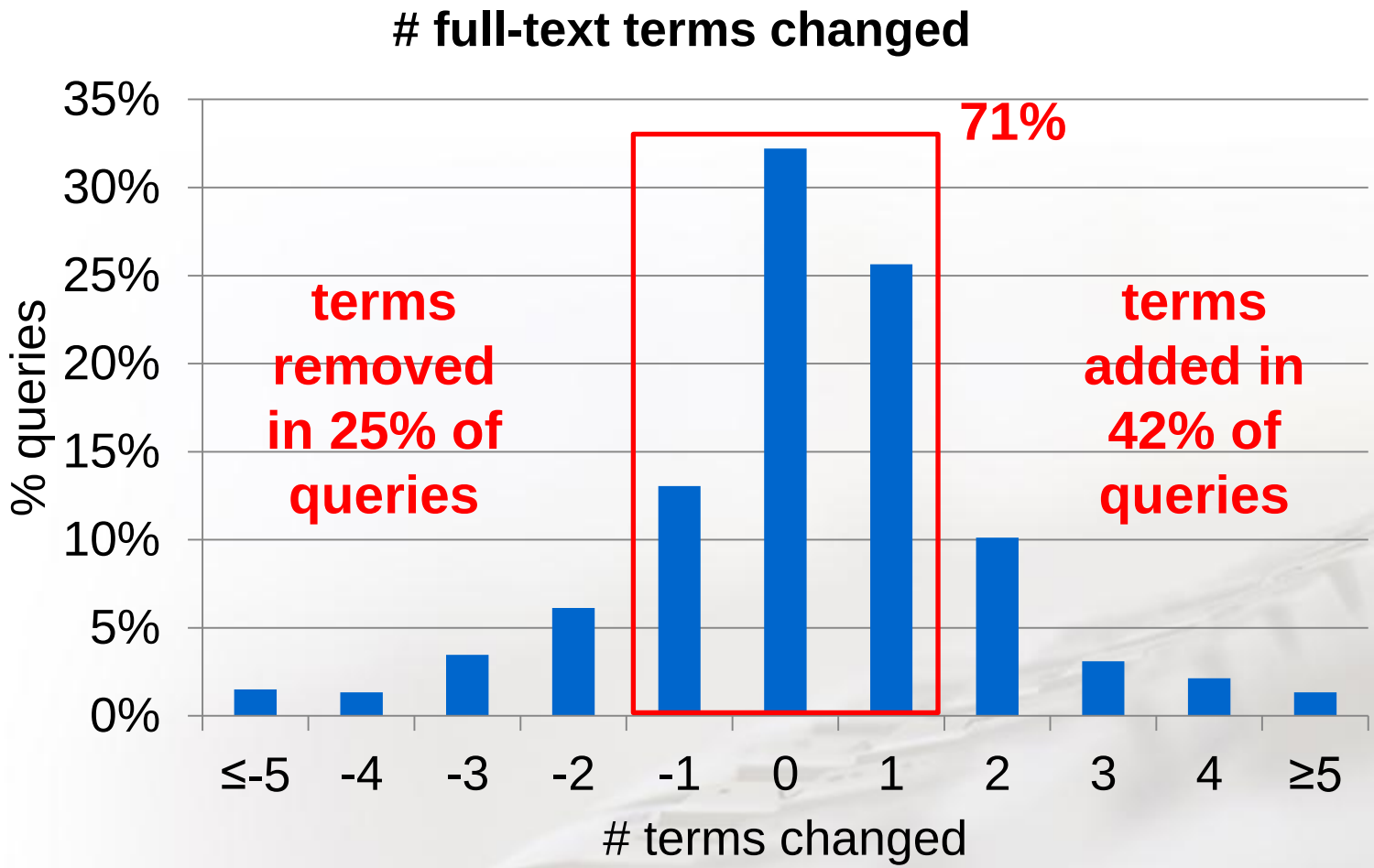
- A typical full-text session:
  - 1 or 2 queries
    - 1 to 3 terms per query
    - 1 or 2 result pages seen per query
    - 1 click per query

- A typical URL session:
  - 1 or 2 queries
    - 1 or 2 clicks per query

Query Distribution

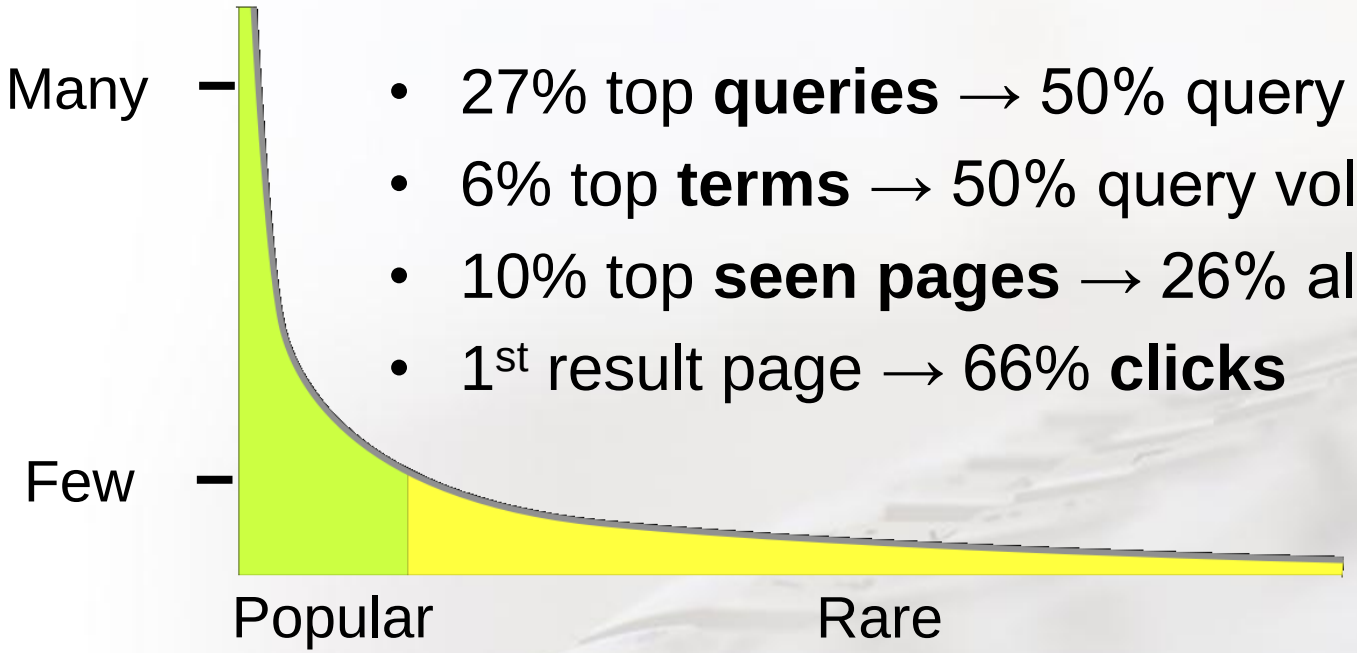


Query Refinement



Exploring Popularity

- Queries, terms, clicks and archived pages seen
  - follow a **power law distribution**



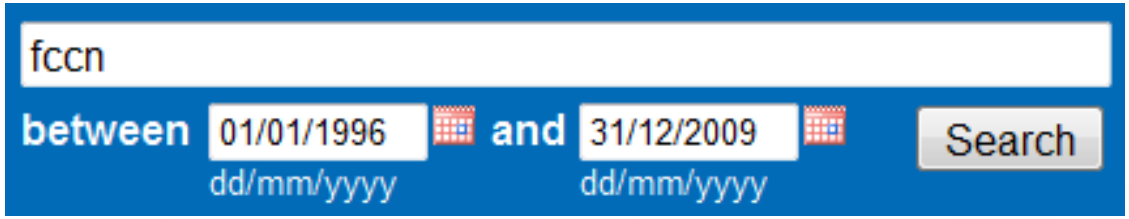
- 27% top **queries** → 50% query volume
- 6% top **terms** → 50% query volume
- 10% top **seen pages** → 26% all seen pages
- 1st result page → 66% **clicks**

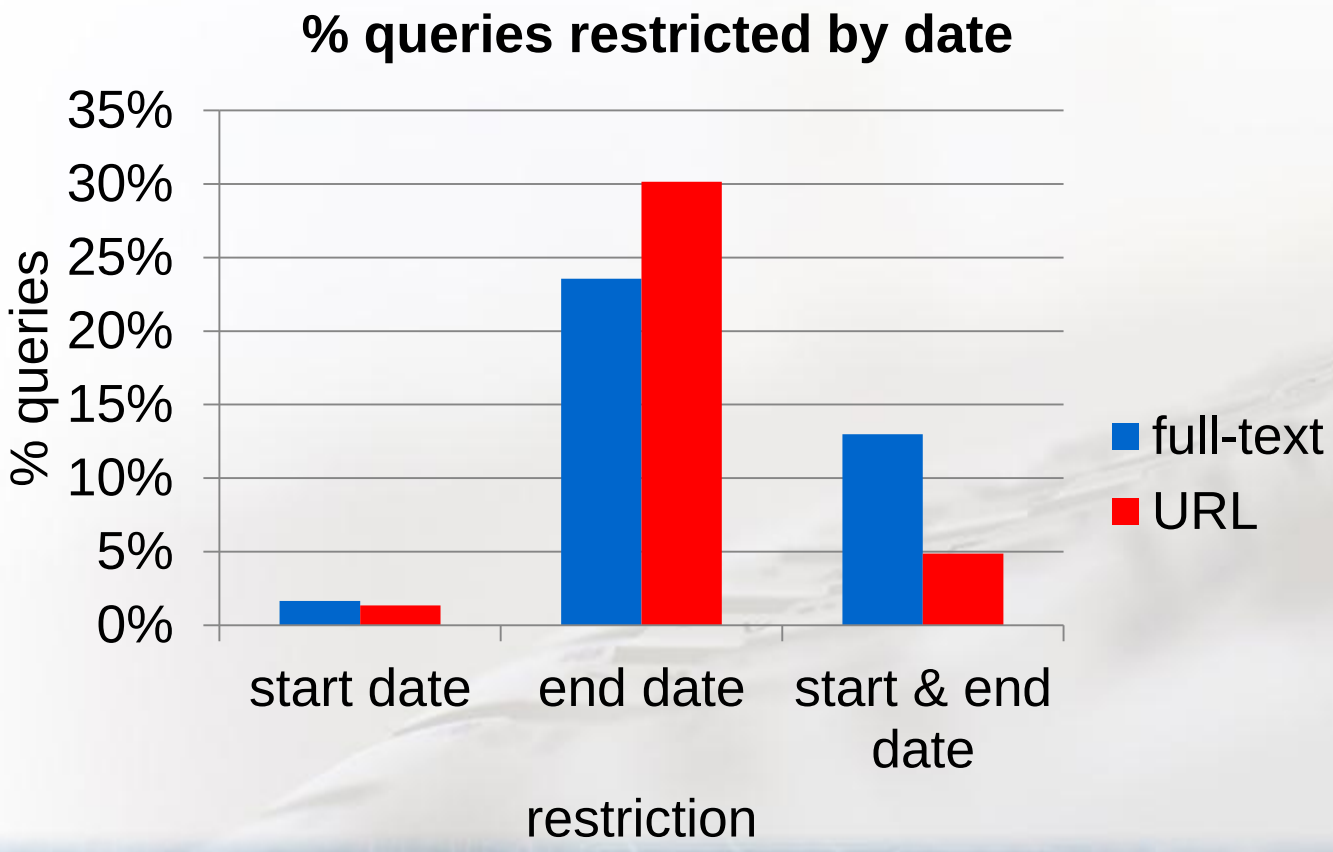
How do users search?

- Spend **little time and effort** on individual searches
- Search and explore following **power law** distributions
- **Search in web archives as in web search engines**
  - Excite (U.S.), Fast (Europe), Tumba! (Portugal)
  - A little less queries, but a bit longer

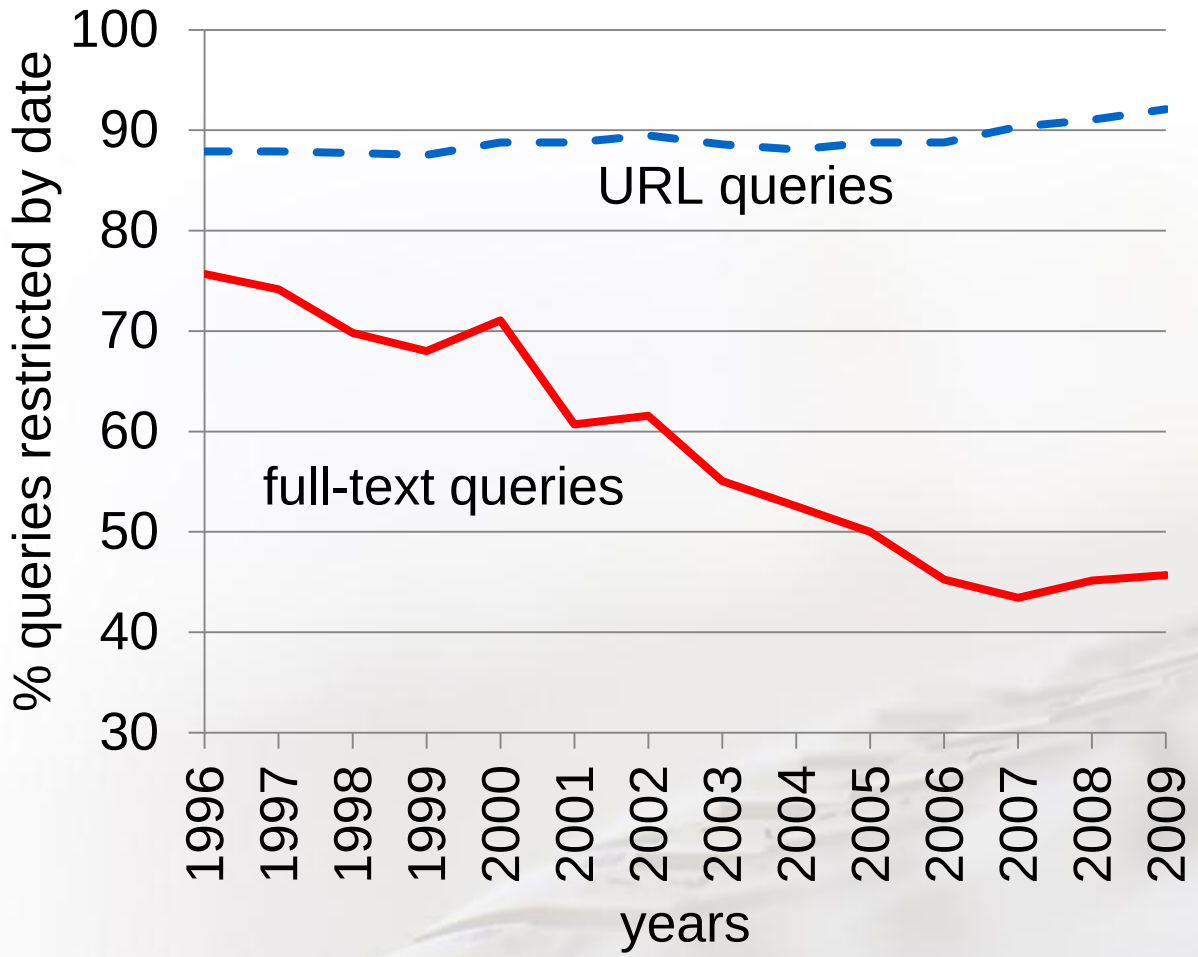
But what about time?

1/3 Queries are Restricted by Date

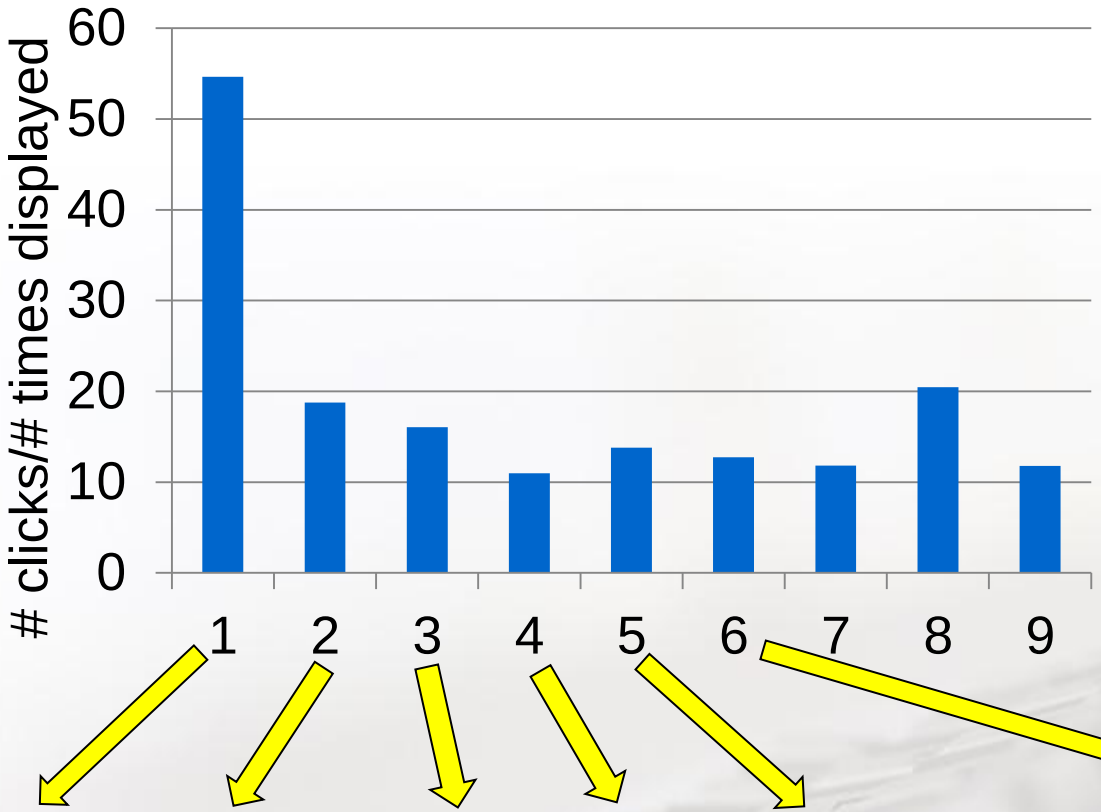




Oldest Versions are more Searched



Oldest Versions are more Clicked



1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
0 pages	0 pages	3 pages	7 pages	32 pages	10 pages	1 page	0 pages	0 pages	73 pages
		2 December 12 December 12 December	25 January 25 January 8 February	29 February 29 February 1 March	9 March 5 April 13 April	14 September			13 October 13 October 16 October

Conclusions

Conclusions

- Web archive users:
  - search as in web search engines
  - prefer full-text search over URL search
  - prefer the oldest documents over the newest

Future Work

- Validate results:
  - with larger datasets
  - with other sources
  - throughout time
- Use results to improve:
  - ranking
  - throughput and response speed
  - user interface

Questions

Thank you.

http://archive.pt