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
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
• 42 web archiving initiatives in 26 countries.
• +180 billion documents archived since 1996.
Web Archiving Workflow

- **Acquisition**
- **Storage**
- **Searching**
- **Presentation**

Preservation

- 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
Interface: full-text search

Result 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

# full-text queries per session

% sessions

# queries

1 2 3 4 5 6 7 8 9 ≥10

1 2 3 4

85%
Query Refinement

# full-text terms changed

- Terms removed in 25% of queries
- Terms added in 42% of queries

71%
Exploring Popularity

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

Many

• 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

Few

Popular Rare
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

% queries restricted by date

- Start date: 0%
- End date: 25%
- Start & end date: 10%

Restriction types:
- Full-text
- URL
Oldest Versions are more Searched

% queries restricted by date

URL queries

full-text queries

years

Oldest Versions are more Clicked
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
Thank you.

http://archive.pt