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
Web Archiving Initiatives

- 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
  - don’t understand the end users
1\textsuperscript{st} : Understanding Users

- **Why** do users search? (information needs)
- **What** do users search for? (topics)
- **How** do users search? (search behavior)
  - this study: 1\textsuperscript{st} 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

Results Page
Interface: URL search

Did you want to find results containing the text: "http://www.fccn.pt"?
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
# full-text queries per session

85%
Query Refinement

# full-text terms changed

- 71% of queries had 1 term added in 42% of queries.
- Terms removed in 25% of queries.

Number of terms changed:

- ≤5 terms changed: 0%
- -4 terms changed: 5%
- -3 terms changed: 10%
- -2 terms changed: 15%
- -1 terms changed: 20%
- 0 terms changed: 25%
- 1 term changed: 30%
- 2 terms changed: 35%
- 3 terms changed: 40%
- 4 terms changed: 45%
- ≥5 terms changed: 50%
Exploring Popularity

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

- Many queries → 50% query volume
- Few terms → 50% query volume
- Popular pages seen → 26% all pages seen
- 66% clicks → 1st result page
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 little 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%

- **full-text**
- **URL**
Oldest Versions are more Searched

- **URL queries**
- **full-text queries**

% queries restricted by date

- Years: 1996 to 2009

- 1996: peak
- 2002: significant drop for full-text queries
- 2009: plateau for URL queries
Oldest Versions are more Clicked

<table>
<thead>
<tr>
<th>Year</th>
<th># pages</th>
<th>Date</th>
<th># clicks/# times displayed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>0 pages</td>
<td>2 December</td>
<td>55</td>
</tr>
<tr>
<td>1997</td>
<td>0 pages</td>
<td>12 December</td>
<td>5</td>
</tr>
<tr>
<td>1998</td>
<td>3 pages</td>
<td>25 January</td>
<td>16</td>
</tr>
<tr>
<td>1999</td>
<td>7 pages</td>
<td>25 January</td>
<td>29</td>
</tr>
<tr>
<td>2000</td>
<td>32 pages</td>
<td>29 February</td>
<td>12</td>
</tr>
<tr>
<td>2001</td>
<td>10 pages</td>
<td>1 March</td>
<td>6</td>
</tr>
<tr>
<td>2002</td>
<td>1 page</td>
<td>9 March</td>
<td>7</td>
</tr>
<tr>
<td>2003</td>
<td>0 pages</td>
<td>5 April</td>
<td>4</td>
</tr>
<tr>
<td>2004</td>
<td>0 pages</td>
<td>13 April</td>
<td>3</td>
</tr>
<tr>
<td>2005</td>
<td>73 pages</td>
<td>14 September</td>
<td>2</td>
</tr>
</tbody>
</table>
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
  – larger datasets, other sources, throughout time

• Use results to improve:
  – ranking considering time
  – throughput and response speed
  – user interface
Thank you.

http://archive.pt