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
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
• 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

Experimental

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/

Help us improve!
It only takes 30s
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

<table>
<thead>
<tr>
<th>Which of the phrases describe best what you were doing?</th>
<th>Taxonomy</th>
<th>Exp. #1 (logs)</th>
<th>Exp. #2 (quest.)</th>
<th>Exp. #3 (lab)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seeing how a web page or site, that I know, was in the past</td>
<td>Navigational</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seeing the evolution over time of a web page or site</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collecting information about a subject written in the past</td>
<td>Informational</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Downloading an old file</td>
<td>Transactional</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Were you searching between specific dates?</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Information Needs

#### 21% of the queries were URLs

<table>
<thead>
<tr>
<th>Which of the phrases describe best what you were doing?</th>
<th>Taxonomy</th>
<th>Exp. #1 (logs)</th>
<th>Exp. #2 (quest.)</th>
<th>Exp. #3 (lab)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seeing how a web page or site, that I know, was in the past</td>
<td>Navigational 1st</td>
<td>47.7%</td>
<td>31.6%</td>
<td>47.6%</td>
</tr>
<tr>
<td>Seeing the evolution over time of a web page or site</td>
<td>Navigational 2nd</td>
<td>9.2%</td>
<td>21.1%</td>
<td>33.3%</td>
</tr>
<tr>
<td>Collecting information about a subject written in the past</td>
<td>Informational 3rd</td>
<td>37.8%</td>
<td>31.6%</td>
<td>14.3%</td>
</tr>
<tr>
<td>Downloading an old file</td>
<td>Transactional</td>
<td>5.3%</td>
<td>10.5%</td>
<td>4.8%</td>
</tr>
</tbody>
</table>

#### Were you searching between specific dates?

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>15.8%</td>
<td>47.4%</td>
<td>9.5%</td>
</tr>
<tr>
<td>No</td>
<td>84.2%</td>
<td>52.6%</td>
<td>90.5%</td>
</tr>
</tbody>
</table>
Users prefer the oldest documents:
### Distribution of topics per navigational needs

<table>
<thead>
<tr>
<th>Topic</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commerce</td>
<td>28.3</td>
</tr>
<tr>
<td>Computers or Internet</td>
<td>14.5</td>
</tr>
<tr>
<td>Education</td>
<td>14.5</td>
</tr>
<tr>
<td>Government</td>
<td>8.4</td>
</tr>
<tr>
<td>Entertainment</td>
<td>7.2</td>
</tr>
<tr>
<td>Sciences</td>
<td>6.0</td>
</tr>
<tr>
<td>Society</td>
<td>5.4</td>
</tr>
<tr>
<td>Things</td>
<td>3.0</td>
</tr>
<tr>
<td>Health</td>
<td>2.4</td>
</tr>
<tr>
<td>Sports</td>
<td>1.8</td>
</tr>
</tbody>
</table>

### Distribution of topics per informational needs

<table>
<thead>
<tr>
<th>Topic</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>People</td>
<td>36.5</td>
</tr>
<tr>
<td>Health</td>
<td>14.8</td>
</tr>
<tr>
<td>Entertainment</td>
<td>9.6</td>
</tr>
<tr>
<td>Things</td>
<td>7.0</td>
</tr>
<tr>
<td>Sports</td>
<td>6.1</td>
</tr>
<tr>
<td>Places</td>
<td>4.4</td>
</tr>
<tr>
<td>Sciences</td>
<td>4.4</td>
</tr>
<tr>
<td>Education</td>
<td>3.5</td>
</tr>
<tr>
<td>Travel</td>
<td>2.6</td>
</tr>
<tr>
<td>Economy</td>
<td>2.6</td>
</tr>
</tbody>
</table>

...
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.
• 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.
Thank you.