A Search Log Analysis of a Portuguese Web Search Engine

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

INFORUM 2010, Braga, Portugal
Problem

Do Portuguese users search in the same way as other users?

Does search behavior influence web search engine design?
Applications

- **Speed**
  - e.g. special indexes, cache

- **Quality of results**
  - e.g. better ranking

- **Web design**
  - e.g. stand out most used functionalities
Summary

• Introduction
• Methodology & Dataset
• Results
• Conclusions
Search Log Analysis

PROS:
• Large and varied
• Less bias
• Cheap
• Non-intrusive

CONS:
• Lack of context
• Lack of control
• Tumba – [http://www.tumba.pt](http://www.tumba.pt)
• 2 full years – 2003 & 2004
  – several studies from the same period
  – baseline for future works

• 90% of the IP addresses → Portugal
• 98% of the interactions → Portuguese interface
How do users search?
Fast and short sessions

- Fast
- Few queries
- Few terms
- Few result pages
- Few clicks
How do users search

Queries per Session

Terms per Query

# Terms Changed

Result Page Viewed
Evolution from 2003 to 2004

- $\frac{1}{2}$ term of query length
- +10% of sessions with less than 1 minute
- +9% of sessions with only one query
- +8% of sessions where only the first result page was viewed
Evolution from 2003 to 2004

- $\frac{1}{2}$ term of query length
- +10% of sessions with less than 1 minute
- +9% of sessions with only one query
- +8% of sessions where only the first result page was viewed

Less data submitted, less results seen
What do users search for?
Top Search Queries

- sexo (2%)
- others (98%)

[Image of a word cloud with 'sexo' and 'others' highlighted]
<table>
<thead>
<tr>
<th>Categories</th>
<th>2003 % queries</th>
<th>2004 % queries</th>
<th>Δ%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Commerce, Travel, Employment or Economy</td>
<td>22.4</td>
<td>20.3</td>
<td>- 2.1</td>
</tr>
<tr>
<td>2 People, Places or Things</td>
<td>14.8</td>
<td>17.7</td>
<td>2.9</td>
</tr>
<tr>
<td>3 Health or Sciences</td>
<td>10.5</td>
<td>11.8</td>
<td>1.3</td>
</tr>
<tr>
<td>4 Education or Humanities</td>
<td>7.2</td>
<td>10.5</td>
<td>3.3</td>
</tr>
<tr>
<td>5 Society, Culture, Ethnicity or Religion</td>
<td>5.6</td>
<td>6.1</td>
<td>0.5</td>
</tr>
<tr>
<td>6 Computers or Internet</td>
<td>6.4</td>
<td>5.9</td>
<td>- 0.5</td>
</tr>
<tr>
<td>7 Sex or Pornography</td>
<td>4.9</td>
<td>5.8</td>
<td>0.9</td>
</tr>
<tr>
<td>8 Entertainment or Recreation</td>
<td>8.7</td>
<td>5.1</td>
<td>- 3.6</td>
</tr>
<tr>
<td>9 Government</td>
<td>7.0</td>
<td>4.2</td>
<td>- 2.8</td>
</tr>
<tr>
<td>10 Performing or Fine arts</td>
<td>1.6</td>
<td>1.6</td>
<td>0.0</td>
</tr>
<tr>
<td>11 Unknown or Other</td>
<td>11.2</td>
<td>11.3</td>
<td>0.1</td>
</tr>
<tr>
<td>world region search engine</td>
<td>U.S. Excite</td>
<td>Europe FAST</td>
<td>Portugal Tumba!</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------------</td>
<td>-------------</td>
<td>----------------</td>
</tr>
<tr>
<td>single term queries</td>
<td>20% -30%</td>
<td>25% -35%</td>
<td>40%</td>
</tr>
<tr>
<td>terms per query</td>
<td>2.6</td>
<td>2.3</td>
<td>2.2</td>
</tr>
<tr>
<td>result pages viewed</td>
<td>1.7</td>
<td>2.2</td>
<td>1.4</td>
</tr>
<tr>
<td>queries per session</td>
<td>2.3</td>
<td>2.9</td>
<td>2.49 -2.94</td>
</tr>
<tr>
<td>topic most seen</td>
<td>Commerce, Travel</td>
<td>People, Places</td>
<td>Commerce, Travel</td>
</tr>
</tbody>
</table>

Less data submitted, less results seen
Conclusions
Conclusions

- Portuguese users
  - spend little time and effort on individual searches
  - tend to submit less data and see less results
  - search differently than other users
  - specificities can be used to tune web search engines
Future Work

- Updated characterization of Portuguese users
- Characterization of Portuguese users from web archives
80% of the web documents are unavailable after 1 year

http://archive.pt
Thank you.