TodoBR: The Search Engine for the Brazilian Web
December 22, 1999
Summary

- Introduction.
- TodoBR.
- TodoBR *versus* other search engines.
- Conclusion.
Introduction

- A search engine is a system which collects and organizes Web documents, and presents a way to select documents based on certain words, phrases or patterns within documents.

- Most search engines use the whole Web.

- Some use only a limited portion of the Web.
TodoBR

- TodoBR (www.todobr.com.br) is a search engine developed by SIAM project.
- TodoBR architecture:
  - Crawler;
  - Index system;
  - Query processing;
  - Interface.
TodoBR

- How does it work?

Diagram:
- Crawler
- Documents
- Index Generator
- Web
- User Interface
- Query Processing
- Index
CoBWeb

- The CoBWeb (Collector of Brazilian Web) is a component of TodoBR search engine that systematically traverse the Web to collect documents for further processing.
- CoBWeb collects only documents from the “.br” domain.
- Up to now 7 million of pages were collected.
CoBWeb

Main features:
- many independent collecting robots operating simultaneously;
- distributed architecture;
- centralized coordination;
- observes the operational and ethical limits;
- collects documents in “.ps, .pdf, .doc, and .html” formats.
CoBWeb

- Architecture:
The index is a frequency-sorted inverted file.

Without elimination of stop-words.

New features:

- compression techniques;
- distributed index.
Query Processing

- The algorithm uses the vector model.
- Evaluation of disjunctive queries:
  - uses a pruning algorithm for filtering documents during the ranking process.
- Evaluation of conjunctive queries:
  - intersection of the inverted lists.
- Relevance feedback.
- Returns only the top 200 documents.
Interface

- TodoBR search engine allows:
  - searching for documents considering all the words;
  - searching for documents considering any words;
  - phrase searching;
  - searching by Brazilian states;
  - searching considering accents in the queries.
- Similar documents, when detected, are grouped and showed as a singular result.
Interface

Netscape: TodoBR – Todo o Brasil na Internet

PROCURA

Joana d’Arc

ESTADO

Todo o Brasil

VOÇÊ SABIA?

A Web foi inventada em 1992
Interface

1. **Joana D'Arc Home Page**
   - **URL:** [http://milener.com.br/~jaco/Joana_darc](http://milener.com.br/~jaco/Joana_darc)
   - [Mais páginas como esta](http://milener.com.br/~jaco/Joana_darc)

2–4. **Colégio Joana D’Arc**
   - **URL:** [http://www.joanadarc.com.br/frame_sala.html](http://www.joanadarc.com.br/frame_sala.html)
   - [Mais páginas como esta](http://www.joanadarc.com.br/frame_sala.html)
TodoBr versus other search engines

- We submitted five queries to TodoBR and other search engines:
  - Radix: a regional search engine;
  - Altavista and Google: generic search engines.

- We analyzed only the top ten documents retrieved.
## TodoBr versus other search engines

<table>
<thead>
<tr>
<th>Queries</th>
<th>TodoBR</th>
<th>Radix</th>
<th>Google</th>
<th>Altavista</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;roberto carlos&quot;</td>
<td>10</td>
<td>0</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>&quot;astro pelé&quot;</td>
<td>10/10</td>
<td>4/4</td>
<td>7/0</td>
<td>0/1</td>
</tr>
<tr>
<td>&quot;governo do pará&quot;</td>
<td>8/8</td>
<td>0/1</td>
<td>3/0</td>
<td>1/0</td>
</tr>
<tr>
<td>&quot;praias no sul da bahia&quot;</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>&quot;hildebrando paschoal preso&quot;</td>
<td>10</td>
<td>0</td>
<td>5 (no results)</td>
<td>0 (only)</td>
</tr>
</tbody>
</table>

Note: The numbers represent the number of relevant results found for each query.
TodoBR search engine was developed by SIAM project/DCC-UFRGS.

TodoBR search engine has the best coverage of the Brazilian Web.

TodoBR search engine focuses on the Brazilian Web and uses the geographic division to retrieve specific information of each state.

TodoBR search engine offers an opportunity to improve the knowledge about the Brazilian Web.