QUERY SUGGESTION FOR WEB ARCHIVE SEARCH

Miguel Costa, João Miranda, David Cruz and Daniel Gomes
{miguel.costa, joao.miranda, david.cruz, daniel.gomes}@fccn.pt
Foundation for National Scientific Computing — Lisbon, Portugal

MOTIVATION

• Users frequently mistyped queries and blamed the web archive for poor search results.
• Existing solutions do not work well, because they rely in predefined lexicons to detect misspellings.
• Lexicons ignore that the terminology and its use evolves throughout time.

IDEAS

• Web archives contain the new and old terms that will be searched. They also contain misspellings.
• Predefined lexicons can be used to compute suggestions and web archives to validate them.
• Common terms in web archive collections, such as names of persons, should not be corrected.
• The suggestion must occur more frequently in the web archive than the query.

METHODOLOGY

• 2 datasets created with term, misspelling pairs.
• 5 existing algorithms tested.
• 2 new algorithms tested: Aspell & Hunspell, with a set of rules automatically tuned over a 8 year web archive index.
• We evaluated only the first suggestion.

RESULTS

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Match</th>
<th>Not answered</th>
<th>Mismatch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levenshtein</td>
<td>4.6%</td>
<td>86.8%</td>
<td>8.6%</td>
</tr>
<tr>
<td>Jaro-Winkler</td>
<td>6.1%</td>
<td>70.6%</td>
<td>23.4%</td>
</tr>
<tr>
<td>N-gram</td>
<td>1.5%</td>
<td>87.8%</td>
<td>10.7%</td>
</tr>
<tr>
<td>Aspell</td>
<td>65.0%</td>
<td>10.7%</td>
<td>24.4%</td>
</tr>
<tr>
<td>Hunspell</td>
<td>73.1%</td>
<td>9.6%</td>
<td>17.3%</td>
</tr>
<tr>
<td>Aspell+Rules</td>
<td>74.1%</td>
<td>14.7%</td>
<td>11.2%</td>
</tr>
<tr>
<td>Hunspell+Rules</td>
<td>77.7%</td>
<td>12.7%</td>
<td>9.6%</td>
</tr>
</tbody>
</table>

CONCLUSIONS

• Existing solutions were improved with rules automatically tuned with an index of archived web collections.
• The query suggestion functionality had great impact on the perceived quality of the service.
• The software is publicly available as an open source project. You can test it at http://archive.pt.