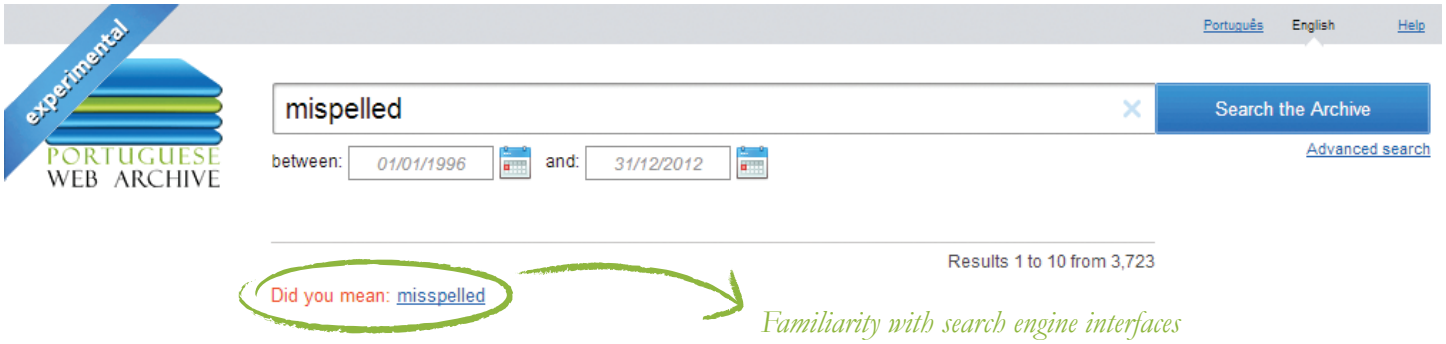
QUERY SUGGESTION FOR WEB ARCHIVE SEARCH

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

	Match	Not answered	Mismatch
Levenstein	4.6%	86.8%	8.6%
Jaro-Winkler	6.1%	70.6%	23.4%
N-gram	1.5%	87.8%	10.7%
Aspell	65.0%	10.7%	24.4%
Hunspell	73.1%	9.6%	17.3%
Aspell+Rules	74.1%	14.7%	11.2%
Hunspell+Rules	**77.7%**	12.7%	**9.6%**

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.

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA
FCCN Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing
GOVERNO DE PORTUGAL
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA