Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

PORTUGUESE WEB ARCHIVE

Proposal for a collaborative project with the Portuguese Web Archive

Creating a benchmark for temporal search results

FCCN is currently engaged in the Portuguese Web Archive (PWA) project and seeks to cooperate with Research and Development organisations who are interested in participating in innovative activities. This document presents a proposal for a project with an estimated duration of 1 year, which could form part of a master's thesis or introduction to research.

The PWA periodically compiles and archives the Portuguese web. However, a system is needed with mechanisms which make the information accessible to the public.

PWA launched a service that provides historical search of the archived information. This makes it possible to search and access archived pages going back a number of years - pages which are no longer available on the Web. Users can search for pages containing certain terms using a search interface which is similar to that provided by current Web search engines such as Google.

The PWA needs to rank the results so that the most relevant results appear in the first positions. However, existing algorithms have been developed to sort results taken from today's Web and not from historical data.

Instead, PWA needs to rank the results considering also a temporal perspective, searching various Web compilations made over time. Therefore, new algorithms are being developed for the PWA to suit these needs.

However, it is essential to have rigorous mechanisms which assess the general impact of the algorithms developed. It is often the case that an attempt to solve one problem which is affecting certain searches, has a negative impact on the results of other searches.

There are joint initiatives for the evaluation of search systems on the Web, where a data set is given for processing, along with a set of search results/responses which are considered correct. (e.g. TREC - **Text REtrieval Conference**). Although these data sets are a good basis for evaluation, they may not be used directly for evaluating the results returned by the PWA search system because:

- They do not consider the temporal dimension of the information. The test sets comprise a single crawl of the web, with just one version per URL;
- The tasks defined may not reflect the actual needs of diverse PWA's users;
- The ranking mechanisms are adjusted depending on the language, and the collections used mainly consist of texts written in English, whereas in the case of PWA, the text are mostly written in Portuguese.

The main objective of the proposed work is to create a benchmark for the search results which will make it possible to evaluate and compare the performance of various algorithms in ranking the results, taking into account the specifications of the PWA and its users.

The benchmark would consist of a list of tasks and relevant results, similar to those created for TREC. For example, for the task "Find information about the campaign of a particular party during the 2009 elections", the relevant results could comprise the website of the party and the blogs of its candidates, as compiled by PWA in May 2009. Only those results existent in the PWA are considered relevant for a task.

The creation of a quality benchmark involves efforts of a significant number of people to compile a set of tests which are representative of the various search types. To achieve this goal using a limited set of resources, tools can be used such as social networks, remote surveys or collaborative work systems.

Bibliography

- David Hawking e Nick Craswell, Very Large Scale Retrieval and Web Search, The TREC Book, 2004.
- Miguel Costa and Mário J. Silva, Towards Information Retrieval Evaluation over Web Archives, SIGIR 2009 Workshop on The Future of IR Evaluation, 2009;
- Omar Alonso and Stefano Mizzaro, Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment, SIGIR 2009 Workshop on The Future of IR Evaluation, 2009.