Proposta de projecto de colaboração com o Arquivo da Web Portuguesa

Criação de uma bancada de teste para resultados de pesquisas temporais

A FCCN tem em curso o projecto de Arquivo da Web Portuguesa (AWP) e procura colaborar com entidades de Investigação e Desenvolvimento, que tenham interesse em participar na realização de actividades inovadoras. Este documento apresenta uma proposta de um projecto com a duração estimada de 1 ano, que poderia fazer parte de um trabalho de mestrado ou de iniciação à investigação.

O AWP periodicamente recolhe e armazena a web portuguesa. No entanto, é necessário dotar o sistema de mecanismos que permitam tornar a informação acessível ao público.

O AWP prevê lançar em breve um serviço de pesquisa histórica sobre a informação arquivada. Será assim possível a pesquisa e acesso a páginas arquivadas ao longo dos anos, que já não estão disponíveis na Web. Os utilizadores poderão pesquisar por páginas que contenham determinados termos, através de uma interface de pesquisa semelhante à disponibilizada pelos motores de busca sobre a Web actual, como por exemplo, o Google.

Tal como nos motores de busca actuais, o AWP necessita de ordenar os resultados, para que os mais relevantes surjam nas primeiras posições. Porém, os algoritmos existentes foram desenvolvidos para ordenar resultados relativos à Web actual.

Por sua vez, o AWP necessita de ordenar resultados considerando uma perspectiva temporal, pesquisando sobre várias recolhas da Web realizadas ao longo do tempo. Assim sendo, o AWP está a desenvolver novos algoritmos de ordenação adequados às suas necessidades.

No entanto, é imprescindível que existam mecanismos rigorosos de avaliação do impacto geral dos algoritmos desenvolvidos. É frequente, que ao experimentar resolver um determinado problema que prejudica certas pesquisas, se prejudique o resultado de outras.

Existem iniciativas conjuntas para a avaliação de sistemas de pesquisa sobre a Web, em que é fornecido um conjunto de dados a processar e um conjunto de pares pesquisa/resposta considerados

correctos (ex. TREC - **Text REtrieval Conference**). Embora estes conjuntos de dados sejam um boa base de avaliação, não poderão ser utilizados directamente para avaliar os resultados devolvidos pelo sistema de pesquisa sobre o AWP porque:

- As tarefas definidas poderão não reflectir as necessidades reais dos diversos utilizadores do AWP;
- Não consideram a dimensão temporal da informação. Os conjuntos de teste são constituídos apenas por uma recolha da Web, existindo apenas uma versão de cada conteúdo;
- A afinação dos mecanismos de ordenação é dependente da língua e as colecções usadas são maioritariamente compostas por textos em inglês, ao passo que no AWP, a maioria dos textos estão escritos em português.

O principal objectivo do trabalho proposto é a criação de uma bancada de teste para os resultados das pesquisas, que permita avaliar e comparar o desempenho de vários algoritmos de ordenação de resultados, tendo em consideração a especificidade do AWP e dos seus utilizadores.

A bancada de teste consistiria numa lista de tarefas e resultados relevantes, à semelhança das criadas para o TREC. Por exemplo, para a tarefa "Encontre informação acerca da campanha de um determinado partido durante as eleições de 2009", os resultados relevantes poderiam ser o sítio web do partido e os blogs dos seus candidatos, recolhidos pelo AWP em Maio de 2009. Só poderão ser considerados como resultados relevantes para uma tarefa, aqueles que existam no AWP.

A criação de uma bancada de teste de qualidade implica a mobilização de um número significativo de pessoas, para que se reuna um conjunto de testes representativo de vários tipos de pesquisa. Para atingir este objectivo recorrendo a um conjunto limitado de recursos, poderão ser usadas ferramentas como redes sociais, inquéritos remotos ou sistemas de trabalho colaborativo.

Bibliografia

- David Hawking e Nick Craswell, Very Large Scale Retrieval and Web Search, The TREC Book, 2004.
- Miguel Costa e Mário J. Silva, Towards Information Retrieval Evaluation over Web Archives, SIGIR 2009 Workshop on The Future of IR Evaluation, 2009;

- Omar Alonso e 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.