Proposta de projecto de colaboração com o Arquivo da Web Portuguesa

Reconhecimento de entidades mencionadas em conteúdos web arquivados

A FCCN tem em curso o projecto de Arquivo da Web Portuguesa e procura colaborar com entidades de Investigação e Desenvolvimento, que tenham interesse em participar na realização de projectos inovadores. 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.

Periodicamente a web portuguesa é recolhida e armazenada para preservação futura. Esta grande quantidade de dados requer mecanismos que permitam aceder à informação arquivada, restringir o espaço de procura e extrair documentos relevantes para as pesquisa realizadas.

O objectivo deste projecto é criar um sistema automático de geração de meta-dados adicionais, descritivos dos conteúdos arquivados. Esta geração irá basear-se na identificação de entidades mencionadas (EM). Ou seja, localizar e identificar elementos no texto associados a classes pré-definidas, tais como pessoa ou organização (ex. Fernando Pessoa) ou expressões temporais (ex. 6 de Agosto de 1966).

O volume de informação de um arquivo da web faz que até ao momento não exista tecnologia capaz de processar os dados históricos e oferecer um sistema de pesquisa por termo, com qualidade semelhante à dos motores de busca.

Como estas EM identificam informação relevante dos documentos, contribuem para filtrar ruído e reduzir o tamanho dos índices que suportam as pesquisas. Permitem também desambiguar pesquisas realizadas pelos utilizadores, por exemplo, diferenciando JAVA, como linguagem de programação da ilha na Indonésia, e perceber que alguns conjuntos de termos têm significado próprio (ex. Presidente da República).

Num arquivo da Web é fundamental localizar no tempo a informação armazenada. Os mecanismos actuais baseiam-se na data de recolha para localizar um conteúdo no tempo, contudo esta abordagem é frágil, pois um conteúdo poderá referir-se a acontecimentos do passado. A identificação de EM temporais, permitirá atribuir com maior exactidão uma data a um conteúdo e consequentemente melhorar a qualidade dos resultados das pesquisas.

O sistema deverá ser implementado na linguagem JAVA sobre a tecnologia Hadoop, uma implementação open source do paradigma de programação MapReduce desenvolvido pelo Google. Esta tecnologia permite distribuir e paralelizar processamentos por clusters com milhares de processadores, sobre quantidades de dados na ordem de grandeza dos Petabytes. Esta escalabilidade quase ímpar e atingida com reduzido esforço para o programador, está actualmente a ser aproveitada por o Yahoo em mais de 10.000 máquinas, para processar até 1 Petabyte de dados em diversos estudos e tarefas, inclusive na indexação de toda a web para o seu motor de busca.