Proposta de projecto de colaboração com o Arquivo da Web Portuguesa

Classificação automática de 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, restringir o espaço de procura e extrair documentos relevantes para pesquisa.

A classificação de documentos contribui para responder a estas necessidades, permitindo navegar hierarquicamente por uma árvore de classes onde os documentos se encontram agrupados. Os directórios de pesquisa Yahoo e Dmoz, são exemplos deste paradigma, permitindo pesquisar informação de uma forma alternativa aos motores de busca, agrupando conteúdos por tópico e oferecendo assim cenários de navegação.

O objectivo deste projecto é criar um sistema automático de classificação de documentos web armazenados ao longo do tempo no Arquivo da Web Portuguesa. A classificação será feita por tópico e sub-tópico, identificando o assunto que o documento descreve (ex. desporto→futebol, política→internacional).

Desta forma, será possível procurar e extrair documentos, requisitando ao sistema todos os documentos de um tópico. Esta classificação permitirá também agrupar no Arquivo da Web os resultados de pesquisa por termo, com duas finalidades: a primeira como pista visual para a pesquisa do utilizador e a segunda para aumentar a variedade de resultados por tópico, da mesma forma que hoje se restringe os resultados por sítio web.

Um aspecto particular num arquivo da web, é que diferentes versões de um documento podem sofrer evoluções tanto a nível visual como do seu conteúdo, logo os tópicos atribuídos podem também evoluir.

O sistema deverá ser implementado na linguagem JAVA sobre a tecnologia Hadoop, uma implementação de código-aberto do paradigma de programação MapReduce desenvolvido pelo Google. Esta tecnologia permite distribuir e paralelizar processamento por *clusters* com milhares de processadores, sobre quantidades de dados na ordem de grandeza dos Petabytes. Esta escalabilidade é atingida com reduzido esforço para o programador e está actualmente a ser utilizada pelo Yahoo em mais de 10.000 máquinas, para diversos estudos e tarefas, inclusive na indexação de toda a web para o seu motor de busca.