Proposta de projecto de colaboração com o Arquivo da Web Portuguesa

Pesquisa de imagens sobre o Arquivo da Web Portuguesa

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 fins de preservação. No entanto, pouco partido se tirará deste vasto manancial de informação, se não forem disponibilizados mecanismos de pesquisa eficientes. Cada tipo de conteúdo necessita de serviços de pesquisa especializados que permitam aceder à informação de forma eficiente. As imagens são dos tipos de conteúdos mais publicados e pesquisados na Web.

O objectivo deste projecto é o desenvolvimento de um sistema de pesquisa sobre as imagens do Arquivo da Web Portuguesa que permita aos seus utilizadores encontrarem e acederem de forma eficiente às imagens recolhidas ao longo dos anos. A pesquisa sobre um arquivo da web apresenta características que o distinguem de um motor de busca convencional. Um sistema de pesquisa sobre um arquivo da web terá de processar uma maior quantidade de informação e ser capaz de lidar com a componente temporal dos dados. Os motores de busca de imagens actuais baseiam-se nos textos que lhe estão associados. No entanto, esta associação entre textos e imagens não é trivial de derivar. O estudo de mecanismos eficientes que permitam extrair ou associar texto a imagens da web é um tópico que levanta desafios interessantes de investigação.

O sistema seria desenvolvido sobre a plataforma de maquinaria e software do Arquivo da Web Portuguesa, existindo já componentes que poderão servir de base ao trabalho a realizar. O desenvolvimento do projecto seria feito com o apoio da equipa especializada em pesquisa de informação do Arquivo da Web Portuguesa.

O sistema deverá ser implementado na linguagem JAVA sobre a tecnologia Hadoop, uma implementação de código-aberto do paradigma de programação MapReduce proposto pelo Google. Esta tecnologia permite distribuir e paralelizar processamento por grupos 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, inclusivamente na indexação de toda a web para o seu motor de busca