Proposta de projecto de colaboração com o Arquivo da Web Portuguesa

Arquivo de Navegrafias de páginas durante a recolha

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 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.

O AWP periodicamente recolhe e armazena a web portuguesa. Este processo é realizado por um componente denominado batedor que iterativamente recolhe, extrai e segue ligações para novos conteúdos (*crawler*). No entanto, é necessário dotar o sistema com mecanismos que permitam preservar a informação arquivada para que se mantenha acessível a longo prazo. Um problema com que se debatem os arquivos da web é como reproduzir as páginas da Web da forma o mais fiel possível à original, passados vários anos.

Os arquivos da web recolhem a informação disponível na Web, no entanto, frequentemente as páginas são codificadas para funcionar apenas em determinados navegadores (*browsers*) e não segundo as normas de formato, como é recomendado.

Este facto levanta principalmente dois problemas. O primeiro é que o batedor por vezes não consegue interpretar o código das páginas e falha na recolha de alguns conteúdos, o que resulta, por exemplo, em páginas arquivadas onde faltam imagens. O segundo problema, é que passados alguns anos, torna-se difícil aceder às páginas arquivadas. Por exemplo, há alguns anos atrás foram publicadas páginas com codificação específica para o navegador Netscape. Hoje em dia, este navegador caiu em desuso e os navegadores actuais têm dificuldade em apresentar correctamente as páginas arquivadas que contém estas codificações específicas. Além disso, os navegadores são desenvolvidos para responderem às características e normas estabelecidas para a Web actual, sendo secundárias as preocupações de retro-compatibilidade com codificações específicas de páginas, que raramente ocorrem na Web actual e apenas existem em arquivos da web.

Embora as páginas apresentem problemas de codificação que causam problemas ao processo de recolha e acesso a longo prazo, normalmente estas páginas funcionam nos navegadores mais comuns à data em que estão disponíveis da Web. O objectivo deste projecto é

criar um componente adicional a incluir no batedor que tirará "fotografias" da aparência das páginas nos navegadores mais comuns durante a recolha, as navegrafias (*browsershots*). Assim sendo, mesmo que um navegador no futuro não consiga interpretar correctamente o código de uma página antiga para apresentá-la correctamente, será sempre possível apresentar uma imagem da sua aparência num navegador antigo à data em que foi recolhida. O problema da preservação a longo prazo das páginas será assim mitigado, aumentando as hipóteses de acesso aos conteúdos a longo prazo.

Além disso, as navegrafias permitirão vir a dotar o sistema de pesquisa do AWP de novas funcionalidades, tais como a apresentação de miniaturas das páginas nos resultados das pesquisas (*thumbnails*) ou desenvolvimento de novas interfaces de utilização que permitam tirar partido desta informação adicional.

O desenvolvimento do projecto será feito em colaboração com a equipa especializada em recolha de informação do Arquivo da Web Portuguesa, recorrendo à sua plataforma de maquinaria e software. O batedor actual do AWP usa o Heritrix, que está implementado em Java. Existem várias tecnologias para gerar navegrafias, tais como o Screengrab, WebpageDump ou o Browsershots. A biblioteca britânica publicou uma lista de software para geração de navegrafias. A geração de navegrafias foi usada em trabalhos de investigação que acedem a informação existente em arquivos da web.