Título

Estratégias de partição multi-objectivo para a descarga eficiente de conteúdos Web

Autor

José Luis Padrão Exposto

Orientadores

António Pina e Joaquim Macedo

Resumo

A descarga de conteúdos Web é uma operação concretizada por robôs, tendo diversas aplicações, tais como: motores de pesquisa, geradores de históricos da Web, *mirroring* e *backups*. No entanto, quando a escala de descarga é aumentada as soluções de robôs centralizadas tornam-se pouco eficientes, principalmente devido a limitações da largura de banda disponível no local onde o robô opera.

A solução imediata para este problema é a descentralização dos robôs através de um sistema distribuído e cooperativo em que diversos robôs são espalhados geograficamente de modo a tirarem partido de múltiplas conexões à Internet, aumentando, desta forma a largura de banda global disponível.

Numa solução distribuída de robôs existem, no entanto, algumas questões importantes a ter em conta: por um lado, (1) é necessário criar um mecanismo que permita a atribuição de quais URLs cada robô deve descarregar; (2) na existência de informação adicional, os robôs devem descarregar os URLs que estão mais próximos dele; (3) os URLs descobertos nas hiperligações das páginas podem ser encaminhados para outros robôs e, por isso, é necessário minimizar este intercâmbio.

São bem conhecidas soluções em que a atribuição dos URLs aos robôs é efectuada por intermédio de uma função de dispersão aplicada ao nome da máquina que alberga o URL, no entanto, é nossa convicção que este mecanismo "cego" pode ser consideravelmente melhorado se se utilizar informação adicional sobre a topologia de rede em que os servidores Web e os robôs assentam.

Através da representação da topologia de rede através de grafos, em que são considerados elementos, tais como, tempos médios de ida e volta (RTT) e distâncias geográficas entre os robôs e os servidores, quantidade de URLs por servidor e número de hiperligações entre as páginas, torna-se, assim possível a aplicação de mecanismos de partição multi-objectivo de tal forma em que seja maximizado o tempo de descarga das páginas pelos robôs e minimizado o intercâmbio de hiperligações entre os robôs.

Nesta comunicação serão apresentados resultados provenientes de experiências que provam esta teoria utilizando como medida principal de pesagem dos grafos o RTT e a distância geográfica entre os servidores. Estamos neste momento, a refinar a parametrização de parâmetros de configuração dos grafos de modo a melhorar os resultados obtidos. Está também a ser realizada uma comparação mais exaustiva da utilização isolada ou combinada de diferentes métricas de modo a encontrar os ganhos obtidos e a concluir a melhor combinação possível.