Partição do espaço Web para a descarga optimizada de conteúdos

José Exposto[1] António Pina[2] Joaquim Macedo[2]

[1]Escola Superior de Tecnologia e Gestão de Bragança
5301-857 Bragança, Portugal
exp@ipb.pt

[2]Universidade do Minho
4710-057 Braga, Portugal
{pina, macedo}@di.uminho.pt

Resumo

As abordagens tradicionais de Recuperação de Informação, baseadas em sistemas centralizados, não têm sido capazes de responder da forma mais eficaz aos problemas de escala, criados pela incomensurável quantidade de informação, actualmente, existente na Web e que se prevê que continue a crescer exponencialmente.

Face a esta situação, têm vindo a ser apresentadas alternativas, baseadas em sistemas distribuídos, com resultados positivos mas, ainda assim, pouco significativos em relação aos objectivos prosseguidos.

As razões para o insucesso das soluções distribuídas é, quanto a nós, em larga medida, reflexo da inexistência de mecanismos de distribuição adaptados às infra-estruturas informacional e comunicacional intrínsecas à Web, não entrando em consideração com informação valiosa, passível de ser obtida através da análise e do tratamento estatístico de dados previamente recolhidos.

Com efeito, o conhecimento aprofundado da topologia física de distribuição de servidores, da topologia lógica do encadeamento de hiperligações e da organização dos conteúdos temáticos das páginas Web, quando usado como suporte da descoberta dos padrões de organização presentes naqueles três níveis organizacionais, pode ser usada para a definição de políticas e heurísticas que tornem efectivas as soluções de recuperação baseadas em sistemas distribuídos.

Nesta perspectiva, a nossa abordagem tem em vista a partição do espaço Web português em múltiplas secções, cada uma das quais formadas por grupos de páginas e respectivos servidores que constituem pontos óptimos de uma medida de localidade, determinada a partir da aplicação de técnicas de aglomeração aos padrões combinados de distribuição física e lógica dos recursos Web.

Uma vez determinadas as secções, cada uma poderá ser atribuída a um ou mais robôs distribuídos e cooperantes numa correspondência tal que o resultado final seja a optimização de uma função global de descarga de páginas, em termos de largura de banda e tempo de comunicação.

Nesta comunicação, são apresentadas as metodologias utilizadas para a recolha de dados do nível físico, o que inclui a informação geográfica e topológica e as distâncias temporais medidas para os servidores Web, dando particular atenção aos aspectos relacionados com a criação e a validação da estrutura de interligação completa de todos os servidores. Esta estrutura representada por um grafo completo é um instrumento experimental fundamental para a criação de um modelo simples de representação que esperamos vir a contribuir de forma decisiva no processo que irá permitir inferir as distâncias temporais entre dois quaisquer servidores.

Posteriormente, com base na informação condensada naquele grafo serão conduzidas experiências de aplicação das técnicas de aglomeração, com vista à avaliação da qualidade das partições produzidas. Procedimento semelhante será aplicado para cada um dos demais níveis organizacionais tendo o cuidado de incorporar de forma incremental em cada nível os resultados obtidos no nível imediatamente anterior, de modo a poder evidenciar os benefícios ou prejuízos resultantes da aplicação das diferentes metodologias ao processo de partição.