ALGORITMOS DE ESCALONAMENTO PARA ATUALIZAÇÃO DE BASES DE PÁGINAS WEB

Alunos: Eduardo Teixeira Cardoso e Caio Dias Valentim
Orientador: Eduardo Sany Laber

Introdução

O Crawler é um dos componentes mais importantes de uma ferramenta de busca. Cabe a este visitar os diferentes sítios para fazer a coleta das páginas que serão indexadas pela ferramenta. Devido a importância comercial dos Crawlers, pouco se sabe sobre como, de fato, as principais ferramentas de busca os implementam apesar de diversos aspectos gerais dos Crawlers estarem disponíveis na literatura [1-3].

Alguns dos trabalhos na literatura estudam os problemas de escalonamento confrontados pelos crawlers. A maioria destes assume o crawler em um momento inicial, onde a topologia da rede não é conhecida. O problema, neste caso, consiste em projetar uma política de visitação para as páginas que tenha bom desempenho em relação a um dado critério, como exemplo, soma das importâncias das páginas coletadas em um dado intervalo de tempo. Uma série de restrições devem ser respeitadas ao realizar as visitas. Em particular, deve haver um intervalo mínimo entre requisições ao mesmo sítio.

Após o crawler ter coletado o conjunto inicial de páginas que serão indexadas pela ferramenta de busca, faz se necessário manter este conjunto de páginas atualizado e incrementá-lo periodicamente. Neste projeto estamos interessados em problemas de escalonamento confrontados pelos crawlers exatamente nesta fase. A diferença fundamental desta fase, em relação à inicial, é que nesta estão (podem estar) disponíveis uma série de informações sobre as páginas já visitadas que podem ser úteis para definir boas políticas de escalonamento [1,2].

Entendemos que existe um espaço para formulação de modelos que tenham a capacidade de capturar os diferentes aspectos necessários para definir políticas otimizadas para atualização de bases Web e que possam ser resolvidos eficientemente do ponto de vista computacional.

Objetivos

Os principais objetivos do projeto são:

- Desenvolver modelos de otimização para escalonar atualizações.
- Projetar algoritmos eficientes para escalonar atualizações e incremento de bases de páginas Web
- Comparar experimentalmente os algoritmos desenvolvidos com outros algoritmos propostos na literatura.

Resultados Obtidos

Através do desenvolvimento de um Simulador conseguimos criar um ambiente controlado no qual podemos testar o progresso e a qualidade das diversas políticas de escalonamento da literatura diante de circunstâncias variadas, assim como criar nossas próprias políticas e compará-las de forma mais direta. O simulador permite lidar com páginas com características diferentes (importância, tamanho) e distribuídas em diversos servidores heterogêneos (velocidades diferentes).

O principal resultado que alcançamos foi o desenvolvimento de uma política eficiente para revisitação de páginas que leva em conta a restrição de politeness. A restrição de politeness estabelece que duas requisições consecutivas a um mesmo servidor não devem ocorrer dentro de um intervalo de tempo pré-determinado (em geral, 15 segundos). Não conhecemos outro trabalho na literatura que tenha levado em conta esta restrição para determinar com que freqüência as páginas devem ser revisitadas. Realizamos experimentos que demonstram que:

- Considerar a restrição de politeness na construção da política tem grande impacto na qualidade da solução final;
- Nossa política tem um desempenho melhor que a política proposta por Wolf [1], quando o ambiente exige o cumprimento do tempo de politeness (como ocorre na prática).
- Nossa política se afastou pouco da solução ótima. Este afastamento foi calculado através de um limite inferior que desenvolvemos para medir o melhor frescor do repositório que pode ser obtido por uma política de revisitação, dado que a restrição de politeness deve ser respeitada. Este limite inferior pode ser utilizado como ferramenta de análise de outras políticas que venham a ser propostas.

Referências

1 - WOLF, J. L. e SQUILLANTE, M. S. e Yu, P. S. e SETHURAMAN, J. e OZSEN, L. **Optimal crawling strategies for web search engines**. In WWW ’02, pages 136-147, New York, NY, USA, 2002. ACM Press.

2 - GARCIA-MOLINA, H. e CHO, J. **Effective page refresh policies for Web crawlers**. ACM Trans. Database Syst., 28(4):390-426, 2003.

3 - MANNING, C. D. e RAGHAVAN, P. e SCHÜTZ, H. **Introduction to Information Retrieval**, Cambridge University Press. 2008.