Pontifícia Universidade Católica
do Rio de Janeiro

Críston Pereira de Souza

Políticas Eficientes para Revisitação de Páginas Web

Tese de Doutorado

Tese apresentada ao Programa de Pós–graduação em Informática do Departamento de Informática da PUC–Rio como requisito parcial para obtenção do título de Doutor em Informática

Orientador: Prof. Eduardo Sany Laber

Rio de Janeiro
Maio de 2010

Pontifícia Universidade Católica
do Rio de Janeiro

Críston Pereira de Souza

Políticas Eficientes para Revisitação de Páginas Web

Tese apresentada ao Programa de Pós–graduação em Informática do Departamento de Informática do Centro Técnico Científico da PUC–Rio como requisito parcial para obtenção do título de Doutor em Informática. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Eduardo Sany Laber
Orientador
Departamento de Informática — PUC–Rio

Prof. Artur Alves Pessoa
Departamento de Engenharia de Produção — UFF

Prof. Marcus Felipe Fontoura
Yahoo! Research

Prof. Marcus Vinicius Soledade Poggi de Aragão
Departamento de Informática — PUC-Rio

Prof. Raúl Pierre Rentería
Departamento de Informática — PUC-Rio

Prof. José Eugenio Leal
Coordenador Setorial do Centro Técnico Científico — PUC–Rio

Rio de Janeiro, 25 de Maio de 2010

Críston Pereira de Souza

Bacharel em Ciência da Computação pela Universidade Federal da Bahia e Mestre em Informática pela Pontifícia Universidade Católica do Rio de Janeiro. Tem experiência profissional como analista de sistemas, e atuou em projetos de pesquisa nas áreas de Otimização Combinatória, Aprendizado de Máquina e Sistemas Distribuídos.

Ficha Catalográfica

Souza, Críston Pereira de

Políticas eficientes para revisitação de páginas Web / Críston Pereira de Souza; orientador: Eduardo Sany Laber. — Rio de Janeiro : PUC, Departamento de Informática, 2010.

v., 84 f: il. ; 29,7 cm

1. Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui referências bibliográficas.

1. Informática – Teses. 2. Máquinas de Busca. 3. Coletores Web. 4. Internet. 5. Algoritmos Aproximativos. 6. Simulação. I. Laber, Eduardo Sany. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Agradecimentos

Ao meu orientador Professor Eduardo Laber pelo tempo dedicado durante todos estes anos, ajudando com muitas idéias e sua forma objetiva de trabalhar.

Ao CNPq, à FAPERJ e à PUC–Rio, pelos auxílios concedidos que permitiram realizar este trabalho.

À minha mãe Silene, que não mediu esforços para a educação dos filhos.

À Andréia pelo companheirismo e tantas ajudas importantes, sem os quais tudo seria mais difícil.

Aos colegas do LEARN/PUC–Rio pelos momentos de descontração, em especial os alunos Eduardo e Caio que colaboraram com esta pesquisa.

Aos funcionários do departamento de informática pela orientação e auxílio nas questões relacionadas ao curso.

Aos Professores Macêdo, Aline, Flávio e George do LaSiD/UFBA, por despertar em seus alunos o interesse pela pesquisa científica.

Resumo

Souza, Críston Pereira de; Laber, Eduardo Sany. **Políticas Eficientes para Revisitação de Páginas Web**. Rio de Janeiro, 2010. 84p. Tese de Doutorado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Uma máquina de busca precisa constantemente revisitar páginas *Web* para manter seu repositório local atualizado. Uma política de revisitação deve ser empregada para construir um escalonamento de revisitações que mantenha o repositório o mais atualizado possível utilizando os recursos disponíveis. Para evitar sobrecarga de servidores *Web*, a política de revisitação deve respeitar um tempo mínimo entre requisições consecutivas a um mesmo servidor. Esta regra é chamada restrição de *politeness*. Devido ao porte do problema, consideramos que uma política de revisitação é eficiente se o tempo médio para escalonar uma revisitação é sublinear no número de páginas do repositório. Neste sentido, quando a restrição de *politeness* é considerada, não conhecemos política eficiente com garantia teórica de qualidade. Nesta pesquisa investigamos três políticas eficientes que respeitam a restrição de *politeness*, chamadas MERGE, RANDOM e DELAYED. Fornecemos fatores de aproximação para o nível de atualização do repositório quando empregamos as política MERGE ou RANDOM. Demonstramos que 0,77 é um limite inferior para este fator de aproximação quando empregamos a política RANDOM, e apresentamos uma conjectura de que 0,927 é um limite inferior para este fator de aproximação quando empregamos a política MERGE. As políticas também são avaliadas através da simulação da execução destas políticas para manter o nível de atualização de um repositório contendo 14,5 milhões de páginas *Web*. Um repositório contendo artigos da Wikipedia também é utilizado nos experimentos, onde podemos observar que a política MERGE apresenta melhores resultados que uma estratégia gulosa natural para este repositório. A principal conclusão desta pesquisa é que existem políticas simples e eficientes para o problema de revisitação de páginas *Web*, que perdem pouco em termos do nível de atualização do repositório mesmo quando consideramos a restrição de *politeness*.

Palavras–chave

Máquinas de Busca; Coletores Web; Internet; Algoritmos Aproximativos; Simulação.

Abstract

Souza, Críston Pereira de; Laber, Eduardo Sany (Advisor). **Efficient Web Page Refresh Policies**. Rio de Janeiro, 2010. 84p. DSc Thesis — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A search engine needs to continuously revisit web pages in order to keep its local repository up-to-date. A page revisiting schedule must be defined to keep the repository up-to-date using the available resources. In order to avoid web server overload, the revisiting policy must respect a minimum amount of time between consecutive requests to the same server. This rule is called politeness constraint. Due to the large number of web pages, we consider that a revisiting policy is efficient when the mean time to schedule a revisit is sublinear on the number of pages in the repository. Therefore, when the politeness constraint is considered, there are no existing efficient policies with theoretical quality guarantees. We investigate three efficient policies that respect the politeness constraint, called MERGE, RANDOM and DELAYED. We provide approximation factors for the repository's up-to-date level for the MERGE and RANDOM policies. Based on these approximation factors, we devise a 0.77 lower bound for the approximation factor provided by the RANDOM policy and we present a conjecture that 0.927 is a lower bound for the approximation factor provided by the MERGE policy. We evaluate these policies through simulation experiments which try to keep a repository with 14.5 million web pages up-to-date. Additional experiments based on a repository with Wikipedia's articles concluded that the MERGE policy provides better results than a natural greedy strategy. The main conclusion of this research is that there are simple and efficient policies that can be applied to this problem, even when the politeness constraint must be respected, resulting in a small loss of repository's up-to-date level.

Keywords

Search Engines; Web Crawlers; Internet; Approximation Algorithms; Simulation.

Sumário

Lista de figuras

As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality.

Albert Einstein, *Sidelights on Relativity.*