Universidade de São Paulo

Instituto de Matemática e Estatística

MAC0499 - Trabalho de Formatura Supervisionado

Mecanismos de Busca

Autor:
Caio de Moraes Braz
Gustavo Perez Katague

Supervisor:
José Coelho de Pina Jr.

4 de Junho de 2012

1 Tema

Recuperação de informação no contexto web.

2 Resumo

Organizar e recuperar informação sempre foi algo presente na história humana. Desde a antiguidade, foram sendo criados métodos para agrupar as informações existentes como pergaminhos e chegando aos livros e enciclopédias. Como consequência, isso gerou jeitos de catalogar e recuperar essas informações quando fosse necessário, um bom exemplo são os sistemas de busca tradicionais em bibliotecas.

A busca tradicional teve um grande salto com o surgimento dos primeiros computadores digitais, em meados da década de 1940, que permitiram buscas mais rápidas, além de recuperar informações sobre os documentos relevantes à busca.

Porém a maior revolução ocorreu em 1989, com o surgimento da *World Wide Web* (WWW), na qual pela primeira vez foram criadas coleções de documentos que possuiam apontadores (*hyperlinks*) para outros documentos. A WWW se tornou rapidamente o principal repositório de informações, crescendo em uma velocidade jamais vista.

Com este crescimento da WWW, um antigo problema tomou novas proporções. Como recuperar informações relevantes em um ambiente novo e muito diferente? Os novos desafios deste ambiente incluem:

- Tamanho: A WWW é grande, muito grande, a ponto de ser atualmente a maior coleção de informação existente.
- Dinamismo: A WWW é dinâmica, isto é, os documentos nela mudam! Mudam de conteúdo, de lugar, os *hyperlinks* são alterados, novos documentos são adicionados.
- Auto-organização: Na WWW, não há um controle sobre o conteúdo de documentos. Documentos aparecem e desaparecem, o conteúdo deles é muitas vezes incerto e cada um deles pode apontar para qualquer outro documento.
- *Hyperlinks*: Porém, a WWW possui *hyperlinks*. Eles fazem ser possível "navegar" na WWW e com isso, teremos a base para encontrar documentos relevantes.

Os primeiros mecanismos de busca concentravam-se no conteúdo para encontrar os documentos, porém nem sempre estes documentos eram relevantes, embora tivessem um conteúdo próximo ao desejado.

Usando a estrutura de *hyperlinks*, um novo modelo para encontrar informações relevantes foi pensado, um deles sendo o *PageRank*, que calcula a relevância de um documento, baseado na estrutura de *links* da WWW. Estes métodos são chamados de "classificadores por popularidade".

3 Objetivos

Neste projeto pretendemos, inicialmente, estudar a teoria sobre mecanismos de busca, em especial sobre os métodos de classificação por popularidade, sendo um deles o *PageRank*.

Após ter a base formada, a foco central será implementar um mecanismo de busca sobre domínios restritos que utilize algum método de classificação por popularidade, exibindo esta classificação.

4 Atividades realizadas

As atividades realizadas até agora foram:

- Reuniões semanais com o supervisor.
- Estudo de materiais sobre o assunto.
- Implementações básicas (Web Crawler, PageRank).

5 Cronograma

	mar	abr	mai	jun	jul	ago	set	out	nov
Estudos	x	x	x	x	x	x			
Implementação					x	x	x	x	
Monografia						x	x	x	x
Apresentação									x
Pôster									x

6 Estrutura esperada da monografia

Atualmente, uma estrutra que é viável para a monografia é:

1. Introdução
2. Grafo da Web
3. Mecanismo de busca
4. PageRank
5. Implementações e simulações
6. Desafios e dificuldades
7. Parte Subjetiva

Esta estrutura poderá ser modificada a medida que o trabalho for avançando durante o ano.

7 Bibliografia

Referências

[1] Sergey Brin, Rajeev Motwani, Lawrence Page e Terry Winograd, *The pagerank citation ranking: Bringing order to the web*, Technical report, Stanford University, 1998.

[2] ———, What can you do with a web in your pocket?, *IEEE Data Engineering Bulletin* **21** (1998), no. 2, 37–47.

[3] Sergey Brin e Lawrence Page, The anatomy of a large-scale hypertextual web search engine, *Proceedings of the Seventh International World Wide Web Conference*, vol. 30, April 1998, `http://infolab.stanford.edu/~backrub/google.html`, pp. 107–117.

[4] Amy N. Langville e Carl D. Meyer, *Google's pagerank and beyond: The science of search engines rankings*, Princeton University Press, 2006.