PageRank - Estudo e Aplicações

Caio de Moraes Braz

Orientador: José Coelho de Pina Jr.

Instituto de Matemática e Estatística - Universidade de São Paulo

15 de novembro de 2011

Introdução

Classificar informações sempre foi algo importante, ainda mais no contexto da Internet onde essa classificação, se torna altamente prioritária, tornando-a um excelente alvo de estudos de como enfrentar este problema, que possui várias dificuldades, como o problema de escala (a Internet é muito grande), problemas de qualidade de informação (muitas páginas existentes tem conteúdo irrelevantes ou errôneos).

Logo, é bastante interessante estudar métodos para classificar a importância relativa entre essas páginas e com isso conseguir distinguir os conteúdos mais relevantes em uma busca, por exemplo.

Objetivos

- Estudar um método de classificar estas informações, no caso, o `PageRank` [2] que é a base da criação do *Google* que conhecemos hoje.
- Desenvolver um software que calcule o PageRank para domínios específicos

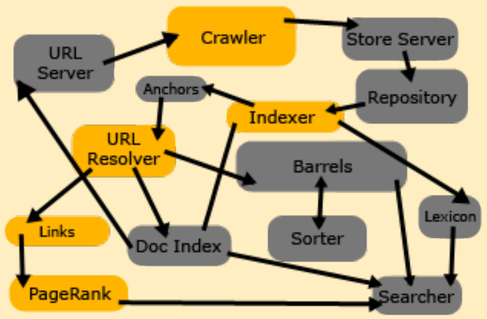
PageRank

O PageRank consiste em atribuir um valor para cada página, de modo que este valor reflita a importância relativa da página, em relação às outras pertencentes ao domínio.

Essa classificação é feita por meio da estrutura de links, determinando a importância de cada página a partir da importância das páginas que apontam para ela.

Anatomia de um mecanismo de busca

Basicamente, um mecanismo de busca, como o *Google* segue uma anatomia como esta, descrita em [1]:



As partes destacadas em amarelo foram alvos de estudo durante este projeto.

Software

O desenvolvimento do software, foi estruturado nas seguintes partes:

- **Web Crawler:** percorrer as páginas de um modo automatizado e eficiente, usando sua estrutira de `hyperlinks`.
- **Parser:** para cada página percorrida, descobrir os `hyperlinks` dela para continuar o processo no Web Crawler.
- **Estruturação dos dados:** montagem da estrutura de links como um grafo, o qual é necessário para o cálculo do PageRank.
- **PageRank:** calcular o PageRank para o grafo do domínio, de forma eficiente.

Modelagem

Basicamente, podemos modelar o domínio em questão como um digrafo, onde os vértices são as páginas e os arcos são os links. Assim sendo podemos dizer que este digrafo representa uma Cadeia de Markov onde para cada estado, os links tem probabilidades iguais. O PageRank é a distribuição estacionária desta cadeia.

Desafios

Alguns dos desafios encontradas durante o projeto:

- lidar com a escala (a quantidade de informação é grande!).
- eficiência dos algoritmos.
- modelagem correta do problema e das estruturas de dados.

Resultados

Um protótipo que calcula o PageRank para o dominio do IME-USP [http://www.ime.usp.br], futuramente, a ideia é extender para domínios quaisquer.

Referências

1. 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.
2. Rajeev Motwani Lawrence Page, Sergey Brin e Terry Winograd, *The pagerank citation ranking: Bringing order to the web*, Tech. report, Stanford University, 1999.

Contato

http://www.linux.ime.usp.br/∼caiobraz/mac499/ e-mail: `caiobraz@linux.ime.usp.br` e-mail: `coelho@ime.usp.br`