PageRank

Caio de Moraes Braz e José Coelho de Pina Jr.
Bacharelado em Ciência da Computação
Instituto de Matemática e Estatística,
Universidade de São Paulo,
Rua do Matão 1010, 05508–900 SãoPaulo, SP
caiobraz@linux.ime.usp.br
coelho@ime.usp.br

19 de setembro de 2011

Resumo

Classificar informações por relevância é um grande desafio, principalmente por se tratar de algo subjetivo, que depende do interesse do usuário, do nível de conhecimento relativo ao assunto, entre outros.

Um dos métodos para classificar páginas na Internet é o conhecido PageRank que atribui um valor (`rank`) a cada página, percorrendo sua estrutura de *links*.

Neste trabalho de formatura supervisionado estudamos o PageRank, suas implementações e aplicações.

Sumário

1 Introdução

1.1 Motivação

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

Logo, é muito 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.

Um dos métodos existentes para classificar páginas na Internet é o **PageRank**, que classifica as páginas atribuindo um valor (que chamaremos de *rank*) a cada página, calculado a partir da estrutura de links destas páginas imersas na Internet.

1.2 Histórico

O PageRank foi desenvolvido na Universidade de Stanford por Sergey Brin e Lawrence "Larry"Page (daí o nome PageRank) como parte de um protótipo de um mecanismo de busca textual em páginas na Internet de nome *Google* [1], que usa de maneira crucial a estrutura de links das páginas.

Este protótipo é a base para o *Google* que conhecemos hoje. O endereço `google.stanford.edu` ainda existe, e hoje é o mecanismo de busca oficial da Universidade de Stanford.

1.3 Objetivos

Neste trabalho pretendemos estudar o PageRank, suas características, as dificuldades de sua implementação, suas aplicações, assim como produzir um pequeno *software* que aplica o PageRank para domínios específicos, como por exemplo, a rede interna do IME ou a rede Linux.

2 PageRank

Nesta seção, vamos detalhar vários aspectos do PageRank, desde sua modelagem como uma cadeia de markov, o algoritmo em si, a implementação e as aplicações onde pode ser utilizado.

2.1 Modelagem

Modelando a Internet como uma cadeia de markov em tempo discreto, temos as páginas sendo estados e os *links* sendo as ligações entre eles. A probabilidade de cada ligação é fixada em $1/N_i$, onde N_i é o número de ligações que sai do estado i.

Logo, a matriz estocástica (i.e. $\sum_i A_{i,j} = 1, \text{para todo } i$) que representa esta cadeia é:

$$A_{i,j} = \begin{cases} 1/N_j, & \text{se existe um arco ligando } j \text{ a } i \\ 0, & \text{caso contrário} \end{cases}$$

Onde N_j é o número de arcos que sai do estado j.

O PageRank é uma distribuição estacionária desta cadeia, onde o rank de cada página é a probabilidade de estar nela após um grande número de mudanças de estado, começando em qualquer página.

O modelo supõe um passeio aleatório nesta cadeia, onde o usuário acessa um link ao acaso em cada passo.

2.2 Algoritmo

Em [2], temos uma primeira definição de PageRank:

$$PR(u) = c \cdot \sum_{v \in B_u} \frac{PR(v)}{L_v}$$

Onde, B_u é o conjunto de páginas que aponta para u, L_v é o número de links que sai de v e finalmente c é uma constante de normalização.

Esta definição recursiva já mostra bem a idéia de que a importância de uma página está diretamente ligada à importância das páginas que apontam para ela, porém, como veremos com mais detalhes na seção 2.3, teremos um problema quando a cadeia possuir o que se chama de *Rank Sink*, (definir BEM isso aqui)!!

Para resolver este problema, é introduzido um fator de amortecimento d que representa a probabilidade do usuário continuar clicando em links. Logo, a ação na qual o usuário eventualmente "cansa"de avançar em links e decide acessar uma página qualquer tem uma probabilidade $(1 - d)$.

Assim sendo, a nova versão do PageRank fica sendo:

$$PR(u) = c \cdot ((1 - d) + d \cdot \sum_{v \in B_u} \frac{PR(v)}{L_v})$$

Garantindo assim que todos as páginas tenham sempre uma chance de ser acessada, fazendo com que o processo funcione.

Uma outra melhoria que pode ser feita é usar um valor d diferente para cada página, gerando assim um PageRank customizado. Veremos mais sobre isso na seção 2.4.

2.3 Implementação

2.4 Aplicações

3 Resultados

4 Conclusão

5 Análise Subjetiva

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.