PageRank

Caio de Moraes Braz, Gustavo Perez Katague 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
katague@linux.ime.usp.br
coelho@ime.usp.br

17 de setembro de 2012

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 mais conhecidos para classificar páginas na Internet é PageRank que atribui um valor (`rank`) a cada página, percorrendo sua estrutura de *links*.

Neste trabalho de formatura supervisionado estudamos o PageRank, sua modelagem implementação e aplicações, assim como comparação com outros métodos.

Sumário

1 Introdução

1.1 Motivação

A necessidade de organizar informação nos segue desde os tempos mais primórdios. Estudos históricos mostram que na antiga biblioteca de Pergamum, os "livros" foram inventados após a falta de papiro vindo do Egito. Estes "livros" eram mais fáceis de se manusear do que os pergaminhos de papiro e logo os substituiram.

Atualmente, a internet proporcionou ao mundo uma revolução no que diz respeito à criação e recuperação de informação. A quantidade de dados aumenta rapidamente, de forma não organizada e muitas vezes com conteúdo equivocado. Seria então conveniente que hovesse alguma forma de classificar a confiabilidade das informações, assim como sua relevância. Neste contexo, essa classificação se torna altamente prioritária, tornando-a um excelente alvo de estudos de como resolver este problema.

É possível perceber que junto com a popularidade da internet vieram mecanismos de busca via web (Web Search Engines). Estes mecanismos se utilizam de diversos 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 Larry Page (daí o nome *PageRank*) como parte de um protótipo de um mecanismo de busca textual em páginas da web chamado *Google* [3], que usa de maneira crucial a estrutura de links das páginas.

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

1.3 Objetivos

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

Em seguida, pretendemos comparar o *PageRank* com outros métodos de classificação, fazendo um estudo detalhado de performance e qualidade das buscas, assim como propriedades e características a respeito de cada um.

2 Grafo da Web

3 PageRank

Antes de 1998, não haviam mecanismos de busca que levavam em conta a estrutura de *hyperlinks* da rede, no entanto, alguns pesquisadores, como Brin e Page já tinham ideias de como retirar informações desta estrtura da web.

3.1 Ideia

A ideia do *PageRank* consiste em ver cada hyperlink como sendo uma recomendação, isto é, um link de uma página para outra significa que a primeira está dando um aval para a segunda, portanto, uma página com mais recomendações provavelmente é mais importante que uma outra com menos recomendações.

No entanto, devemos também levar em conta a fonte da recomendação, pois dependendo de quem ela veio, pode ser mais relevante. Por exemplo, uma recomendação de Donald Knuth para uma página sobre algoritmos parece ser mais importante do que uma recomendação dele para uma página sobre esportes. Por outro lado, se descobrirmos que o Donald Knuth é uma pessoa muito gentil e recomenda muitas páginas sobre algortimos, então o valor desta recomendação deve ser menor que o esperado.

Resumidamente, é assim que o *PageRank* funciona: "uma página é importante se ela é recomendada por outras páginas importantes".

3.2 Matemática

Vamos agora entender como o *PageRank* é definido matemáticamente.

A primeira fórmula, apresentada em [1], define o *PageRank* de uma página P_i, denotado como $r(P_i)$, como:

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|P_j|}$$

Onde B_{P_i} é o conjunto das páginas que apontam para P_i e $|P_j|$ é o número de links contidos em P_j.

Temos em mãos um processo iterativo, que com um pouco mais de notação, fica mais claro:

$$r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|}$$

O processo é iniciado com $r_0(P_i) = 1/n$, para todo P_i, onde n é o número total de páginas e é iterado com esperança de que este valor convirja para algo estável.

3.3 Representação Matricial

As equações vistas acima podem ser escritas na forma de um produto matriz-vetor. Para isso, vamos utilizar uma matriz $H \in \mathbb{R}^{n \times n}$ e um vetor $\pi \in \mathbb{R}^n$.

O vetor π é onde estarão guardados os valores do *PageRank* e a matriz H é definida como:

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

Com esta notação matricial, podemos reescrever a fórmula inicial como:

$$\pi^{(k+1)T} = \pi^{(k)T} H$$

E da fórmula acima, podemos observar que a complexidade de cada iteração do algoritmo é $O(n^2)$, dado a multiplicação matriz-vetor. Entretanto, a matriz H é esparsa, e apenas suas entradas não-nulas são armazenadas. Segundo [4], a média de links é de aproximadamente 10 por página. Mais especificamente, H possui algo próximo de $10n$ entradas não-nulas, ao invés de n^2 de uma matriz densa. Dessa maneira, a complexidade da iteração é da ordem de $O(n)$.

3.4 Ajustes

Devemos agora contornar alguns problemas provenientes da modelagem escolhida. O primeiro são os *rank sinks*, que são páginas que não tem links para nenhuma outra, acumulando *rank* a cada iteração sem repassar para outras páginas nas próximas iterações.

Para corrigir o problema do *rank sink*, é feito um ajuste estocástico na matriz H original. As linhas nulas em H serão substituídas pelo vetor $\frac{1}{n} \cdot e^T$, gerando uma nova matriz que chamaremos de S. Esta nova matriz é estocástica.

3.5 Teoria de Cadeias de Markov

O vetor π do *PageRank* é uma distribuição estacionária em tempo discreto.

Definição 1. *Uma* ***matriz estocástica*** $P_{n \times n}$ *em que cada linha a soma de suas entradas é igual à 1.*

Definição 2. *Um* ***processo estocástico*** *é um conjunto de variáveis aleatórias* $\{X_t\}_{t=0}^{\infty}$ *que contêm um espaço de estados* $\{S_1, S_2, ..., S_n\}$ *em comum. O parâmetro* t *é geralmente*

associado ao tempo e X_t representa o estado do processo no tempo t. No contexto do PageRank consideraremos o tempo discreto.

Definição 3. *Uma **cadeia de Markov** é um processo estocástico que satistaz a propriedade de Markov*

$$P(X_{t+1} = S_j | X_t = S_{i_t}, X_{t-1} = S_{i_{t-1}}, \ldots, X_0 = S_{i_0}) = P(X_{t+1} = S_j | X_t = S_{i_t})$$

para cada $t = 0, 1, 2 \ldots$. A notação $P(E|F)$ denota probabilidade condicional do evento E ocorrer dado o acontecimento do evento F.

A propriedade de Markov garante que o processo não possui memória. As transições de estado ao longo do tempo dependem apenas do estado atual, sendo sua história irrelevante. O processo do usuário aleatório na web é uma cadeia de Markov.

Definição 4. *A probabilidade de transição $P_{ij}(t) = P(X_t = S_j | X_{t-1} = S_i)$ é a probabilidade de mudança do estado S_i para o estado S_j*

Definição 5. *Uma cadeia de Markov estacionária é uma cadeia cujas probabilidades de transição não são alteradas com o tempo.*

Definição 6. *Um estado S_j é dito acessível de um estado S_i, ou $S_i \to S_j$, se $P(X_{n+k} = S_j | X_k = S_i) = p_{ij}^{(n)} > 0$*

Definição 7. *Um estado S_i se comunica com S_j se $S_i \leftrightarrow S_j$.*

3.6 Implementação

3.7 Aplicações

4 Resultados

5 Conclusão

6 Análise Subjetiva

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.