PageRank - Estudo e Aplicações

Caio de Moraes Braz & Gustavo Perez Katague
Supervisor: José Coelho de Pina Jr.

Instituto de Matemática e Estatística - Universidade de São Paulo

9 de novembro de 2012

Introdução

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 e relevância das informações.

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.

Objetivos

Estudar métodos de classificação de conteúdo baseados em estrutura de *hyperlinks*, analisando características, performance computacional e sensibilidade a parâmetros.

Nosso principal alvo de estudos foi o algoritmo *PageRank*.

Desafios

Alguns dos desafios encontradas durante o trabalho:

- Lidar com a escala dos dados.
- Implementação eficiente dos algoritmos.
- Problemas de precisão numérica.

Digrafo da Web

A estrutura de *hyperlinks* da internet forma um grafo dirigido. Os nós do grafo representam as páginas e as arestas dirigidas representam os *links*. Podemos representar um grafo através de uma matriz de adjacência. Seja P_i uma página da web indexada com o inteiro i, então tomemos a matriz L como representação de um conjunto de páginas e seus *links*:

$$L_{ij} = \begin{cases} 1, & \text{se existe um } \textit{link} \text{ de } P_i \text{ para } P_j \\ 0, & \text{caso contrário} \end{cases}$$

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. O *PageRank* é definido 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.

Modelagem

A primeira modelagem da matriz de adjacência, que chamaremos de H, é uma breve adaptação do grafo da web. H é semi-estocástica, pois podem existir páginas que não possuem nenhuma outra como referência.

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

Para tornar H estocástica, é feita outra adaptação, onde indiretamente as páginas sem referências apontam para todas as outras. A nova matriz S pode ser escrita:

$$S = H + (1/n)ae^T$$

$$a_i = \begin{cases} 1, & \text{se a } i\text{-ésima linha de } H \text{ for nula} \\ 0, & \text{caso contrário} \end{cases}$$

E onde e é um vetor no qual todas as entradas valem 1. Para calcular o *PageRank*, precisamos de uma matriz aperiódica e irredutível, pois queremos calcular a distribuição estacionária da cadeia de Markov por ela representada. O último ajuste necessário transforma S na matriz G:

$$G = \alpha S + (1 - \alpha)(1/n)ee^T$$

Onde α é a probabilidade de seguir a estrutura de links original.
Calcularemos um vetor π, onde π_i representa o rank atribuido à página P_i. Temos então a seguinte formula que descreve o *PageRank*:

$$\pi^T = \pi^T G$$

Implementação

Neste trabalho, foram implementados dois algoritmos, o ***PageRank*** e o ***HITS***.
Em ambos os algoritmos utilizamos o método da potência para realizar os cálculos. Este método encontra uma boa aproximação para o autovetor associado ao maior autovalor em escala da matriz em questão por meio de um processo iterativo, no qual em cada passo estamos com uma aproximação melhor deste autovetor.
No caso do *PageRank*, a matriz G foi construída de modo que o método da potência termine, convergindo para o vetor no qual estamos interessados.

Sensibilidade do *PageRank*

Um dos parâmetros mais relevantes quando calculamos o *PageRank* é o $\alpha \in [0, 1]$, que influencia diretamente a taxa de convergência do processo, de modo que quanto mais próximo de 1, mais iterações são necessárias.

Referências Bibliográficas

- Sergey Brin, Rajeev Motwani, Lawrence Page e Terry Winograd, *The pagerank citation ranking: Bringing order to the web*, Technical report, Stanford University, 1998.
- Amy N. Langville e Carl D. Meyer, *Google's pagerank and beyond: The science of search engines rankings*, Princeton University Press, 2006.

Contato

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