Classificando Páginas em Redes Desconexas com PageRank

Veronica Susana Fichera de Werneck , **Fábio Borges**
Instituo Superior de Tecnologia, FAETEC
25.651-075, Petrópolis, RJ
E-mail: veronica@lncc.br, borges@lncc.br.

Resumo: Os primeiros mecanismos de busca na Internet se baseavam apenas em informações de suas páginas. Com o exponencial crescimento do número de páginas, isto se tornou insuficiente, pois várias páginas poderiam ter o mesmo conteúdo. Neste momento, ao final dos anos noventa, foi desenvolvido o algoritmo PageRank para ajudar a classificar-las baseando na centralidade [2] do autovetor associado aos seus *links*.

Este resumo apresenta uma pesquisa teórica, em desenvolvimento através de um trabalho de conclusão de curso de graduação em computação. Tal trabalho está codificando um programa para catalogar as páginas da Internet e simular o funcionamento do PageRank.

O algoritmo conhecido como PageRank é dado pela fórmula

$$x_k = \sum_{j \in L_k} \frac{x_j}{n_j},$$

onde x_k representa a importância da página k, sendo L_k o conjunto de páginas de 1 a n, n_j representa o número *links* contidos na página j.

Para acharmos a pontuação de uma página, e considerando que esta pontuação se resume a encontrar um autovetor correspondente ao autovalor principal $\lambda = 1$ para a matriz que representa uma determinada rede [1], podemos observar que, no caso de redes com páginas isoladas o autovalor será $\lambda < 1$ o que complicará a classificação desta página. Note que as páginas isoladas formam uma rede desconexa, considerando os vértices de um grafo como páginas na Internet e as arestas como seus *links*.

O desejável para o cálculo desta pontuação, tendo uma matriz A representando a rede Internet, é que $\dim(V_1(A)) = 1$ o que faz com que exista um único autovetor x com $\sum_i x_i = 1$.

Mas numa rede desconexa, a matriz A que representará esta situação terá uma ou mais colunas de zeros e desta forma não será coluna-estocástica (matriz coluna estocástica significa que a soma de todas as entradas de cada coluna soma 1 e devem ser positivas), e desta forma o autovalor será menor ou igual a 1 e $V_1(A)$ será bidimensional.

Uma solução para este problema é substituir a matriz A pela M, aplicando a seguinte fórmula:

$$M = (1 - m)A + mS,$$

onde S é uma matriz $n \times n$ com todas as entradas $1/n$, e m é um valor entre 0 e 1. O valor de m utilizado pelo Google é 0.15. A substituição da matriz A pela M permite comparar páginas de redes desconexas.

A idéia principal do *Power method* [2] é começar com um vetor típico x_0 e gerar uma seqüência $x_k = Mx_{k-1} = M^k x_0$ aproximando k ao infinito. O vetor x_k se aproxima bastante ao autovetor para o autovalor de maior magnitude de M. Com tudo, dependendo da magnitude do autovalor, o vetor x_k pode crescer sem limite ou decair para o vetor zero. Assim, a cada iteração podemos calcular

$$x_k = \frac{Mx_{k-1}}{||Mx_{k-1}||},$$

onde $|| \cdot ||$ pode ser um vetor normal.

Referências

[1] K. Bryan and T. Leise. The $25,000,000,000 eigenvector: The linear algebra behind google. *SIAM Rev.*, 48(3):569–581, 2006.

[2] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.