XXXI Congresso de Iniciação Científica Unicamp 2023

PAGERANK: UMA VISÃO DE ÁLGEBRA LINEAR

Palavras-chave: GOOGLE, PAGERANK, ÁLGEBRA LINEAR

Autores:
Leonardo Rangel de Albuquerque (aluno) [Unicamp]
Prof. Dr. Paulo José da Silva e Silva (orientador) [Unicamp]

1 Introdução

O objetivo da pesquisa é a criação de um artigo para introduzir e explicar aos leitores um algoritmo de ranqueamento de páginas da Internet chamado de *PageRank*. O texto tem como foco principal expor a matemática, que principalmente consiste em Álgebra Linear, por trás do algoritmo de maneira intuitiva. Deste modo, O público alvo do artigo são estudantes de graduação que já tenham feito um curso básico de Álgebra Linear.

Além da criação do texto, a pesquisa visa o desenvolvimento de códigos, na linguagem de programação Python, para o cálculo de *ranks* de páginas utilizando o algoritmo do *PageRank*.

2 Feitos

Para a criação do artigo, foi estudado ao longo de um semestre, dois textos: o livro Google's PageRank and Beyond: The Science of Search Engine Rankings ([LM11]), e o artigo The $25,000,000,000 Eigenvector: The Linear Algebra behind Google ([BL06]).

No livro estudado, há códigos na linguagem de programação MATLAB. Assim, para poder ler e entender os códigos, foi preciso fazer um mini-curso de introdução ao MATLAB no site oficial do mesmo. Com o domínio básico da linguagem, foi possível fazer a tradução dos códigos do livro para Python.

Após todos esses feitos, foi iniciado o desenvolvimento do principal objetivo da pesquisa, um artigo. Foi escolhido que o mesmo fosse num formato de *notebook* para que os códigos criados pudessem ser implementados facilmente ao lado do texto. A outra opção seria a criação de um site interativo ao usuário. Porém, ela não foi tomada adiante, já que acarretaria num custo e trabalho a mais que não "valeria a pena".

O texto, ainda, não foi finalizado. Contudo, a sua conclusão é o principal foco da pesquisa no momento.

3 Resumo do Artigo

Mesmo o artigo estando inacabado, já foram escritas suas ideias e argumentos principais. Deste modo, essa seção consistirá num resumo do mesmo.

3.1 O PageRank

O PageRank era um dos algoritmos de classificação de páginas, da Internet, da empresa Google nos seus primeiros anos de vida. A criação e desenvolvimento do algoritmo, no final da década de 90 ao começo dos anos 2000, fez com que o Google se tornasse uma das maiores empresas de tecnologia do planeta. A ideia central deste algoritmo impactante pode ser resumida na seguinte frase: uma página na Internet é importante se páginas importantes levam a ela.

3.2 A Matemática por trás

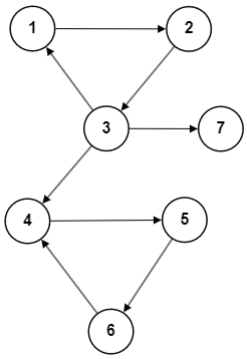

Figure 1: Conjunto P de páginas representadas por nós e links entre as páginas representados por arestas.

Dado o conjunto P de páginas da Internet e o grafo que representa suas conexões da figura (1), criamos uma matriz de Adjacência H' das conexões entre estas páginas. Se a página P_i possuí um *link* para a página P_j o elemento H'_{ij} será 1, caso contrário (não há um link na página P_i que leva a página P_j) será 0. A matriz de adjacência H' do grafo do conjunto P é

$$H' = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{1}$$

Para poder se falar de probabilidade de um usuário ir de uma página à outra, devemos normalizar cada linha de H, exceto a linha 7 o qual é toda composta por 0s. Assim, chamando de H a matriz H' com as linhas normalizadas, obtemos

$$H = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{2}$$

A matriz H agora é uma *matriz subestocástica*. A interpretação, em nosso estudo, para H é a seguinte: se um usuário, por exemplo, estiver na página P_1, a probabilidade dele ir para a página P_2 é 1 e para qualquer outra é 0. Caso o usuário se encontrar na página P_3, a probabilidade dele ir para a página P_7 é 1/3. O grafo da figura (2) ilustra essa ideia.

Porém, há um problema. Caso o usuário estiver na página P_7, a chance dele ir para qualquer outra página é nula. Assim, o mesmo ficará *para sempre* em P_7, um comportamento indesejado para o usuário fictício.

Para resolver esse problema fazemos um *ajuste estocástico* em H. Este consiste na mudança de, nas linhas as quais todos seus elementos são 0, substituímos ela por uma em que todos seus elementos são $1/n$, em que n é o número de páginas do conjunto P (dimensão da matriz H). Chamando de S essa nova *matriz Estocástica*, obtemos

$$S = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \end{bmatrix}. \tag{3}$$

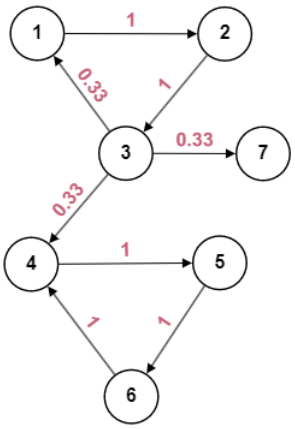

Figure 2: Grafo ilustrando a interpretação probabilística de H. Os valores nas arestas representam a probabilidade de ir de uma página a outra.

A interpretação que difere S de H é que, caso o usuário estiver na página P_7, a probabilidade dele ir para qualquer outra página é a mesma. O grafo da figura (3) representa essa ideia.

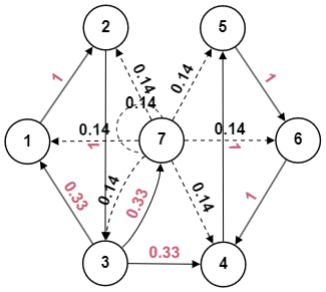

Figure 3: Grafo ilustrando a interpretação probabilística de S. Os valores nas arestas representam a probabilidade de ir de uma página a outra. A aresta circular no nó 7 representa a possibilidade do usuário ir à página 7 já estando na página 7.

Note o seguinte fato sobre S. Caso o usuário estiver na página P_4, ele irá para P_5. Caso estiver em P_5, ele irá para P_6. E caso estiver em P_6, ele irá para P_4. Criando assim um *loop*. Portanto, caso o usuário estiver tanto em P_4, P_5 ou P_6, ele `nunca` irá chegar, por exemplo, em P_1. Isso, para o nosso objetivo, é um problema e deve ser ajustado.

Para que o fato discutido acima não ocorra, consideramos uma leve modificação no comportamento do usuário que navega as páginas de P. Agora, antes de simplesmente seguir algum dos links, da página em que se encontra no momento, ele tem uma probabilidade $1 - \alpha$ de ir para qualquer página de P. Assim criaremos uma nova matriz a partir de S com essa nova propriedade. Essa matriz é a chamada *matriz Google* G. Ela é obtida pela seguinte equação:

$$\mathbf{G} = \alpha\mathbf{S} + (1 - \alpha)1/n\mathbf{e}\mathbf{e}^{\mathbf{T}}. \tag{4}$$

Em que $\alpha \in (0, 1)$ e $1/n\mathbf{e}\mathbf{e}^{\mathbf{T}}$ é uma matriz de "teleportação aleatória" em que todos seus elementos são $1/n$. Em nosso exemplo, escolhendo $\alpha = 0.85$ a matriz G é

$$G = 0.85 \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \end{bmatrix} + 0.15 \begin{bmatrix} 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \end{bmatrix}. \tag{5}$$

O problema matemático em si do PageRank é encontrar um *vetor de densidade de probabilidade estacionário* π associado a G. Esse vetor, que é chamado de *Vetor do PageRank*, é o autovetor à esquerda associado ao autovalor $\lambda = 1$, de norma 1 igual à 1, da matriz G ($\pi^T = \pi^T G$). Na questão de existir e ser único, temos que, dado G, sua existência e unicidade são garantidas graças a um teorema chamado de Teorema de Perron.

Como, na vida real, um conjunto P de páginas da Internet pode atingir um número muito grande de páginas (e.g. $n = 10^9$), a melhor maneira de se calcular π é por meio de um método iterativo. Por razões numéricas, o *Método da Potência* ($\pi^{(k+1)T} = \pi^{(k)T} G$) se torna o favorito para seu cálculo.

A convergência ao *Vetor do PageRank* pelo Método da Potência é garantida para qualquer vetor inicial $\pi^{(0)}$ devido ao fato do autovalor $\lambda = 1$ ser o único autovalor no raio espectral de G. Além disso, a velocidade de convergência dependerá do valor de α, quanto mais próximo de 1, mais lento será a convergência, e quanto mais próxima de 0, mais rápida será.

Por fim, utilizando a linguagem de programação Python para criar um código que simula o Método da Potência no cálculo do Vetor do PageRank, obtemos que, para o grafo da figura (1), o mesmo é dado por,

$$\pi^T = [0.05352337 \quad 0.07342271 \quad 0.09033715 \quad 0.25251666 \quad 0.24256699 \quad 0.23410976 \quad 0.05352337]. \tag{6}$$

Assim, temos que, no grafo, a página mais importante segundo o PageRank é a P_4.

References

[BL06] Kurt Bryan and Tanya Leise. "The $25,000,000,000 Eigenvector: The Linear Algebra behind Google". In: *SIAM Review* 48.3 (Jan. 2006), pp. 569–581. ISSN: 0036-1445. DOI: 10.1137/050623280.

[LM11] Amy N. Langville and Carl D. D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, July 2011.