**Revista Eletrônica
Paulista de Matemática**

ISSN 2316-9664
Volume 21, dez. 2021
Iniciação Científica

Ariely da Silva Camargo

Faculdade de Engenharia Mecânica
Universidade Federal de Uberlândia
ariely.camargo@ufu.br

Ana Paula Tremura Galves

Faculdade de Matemática
Universidade Federal de Uberlândia
ana.galves@ufu.br

Abordagem matemática por trás do algoritmo PageRank

Mathematical approach behind the PageRank algorithm

Resumo

Este artigo aborda um algoritmo conhecido como PageRank. Criado no final da década de 1990, ele auxiliou a classificação de páginas resultantes das pesquisas realizadas por buscadores da Web, possibilitando a criação do Google. Neste trabalho, será analisado esse algoritmo, entendendo, primeiramente, seu conceito natural através de exemplos simplificados da Web. Após os exemplos, são vistas formas de se modelar matematicamente o problema. Em seguida, são feitas generalizações para Webs com mais páginas. Em relação a modelagem, a maneira mais comum de se abordar esse assunto é utilizando a Cadeia de Markov, modelo que pressupõe conceitos probabilísticos. No entanto, esse artigo mostra uma abordagem mais intuitiva, alicerçada na Álgebra Linear, utilizando conceitos como autovalores, matrizes positivas e normas. Além disso, tendo em vista que esse algoritmo é implementado computacionalmente, será mostrado um método por aproximação para calculá-lo, o Método das Potências adaptado.

Palavras-chave: Algoritmo PageRank. Autovalores. Matriz estocástica. Método das Potências

Abstract

This article discusses an algorithm known as PageRank. Created in the late 1990s, it helped the classification of pages resulting from searches made by Web search engines, enabling the creation of Google. In this paper, this algorithm will be analyzed, understanding, first, its natural concept through simplified examples from the Web. After the examples, ways to mathematically model the problem are seen. Then, generalizations are made for Web sites with more pages. Regarding modeling, the most common way to approach this subject is using the Markov Chain, a model that assumes probabilistic concepts. However, this article shows a more intuitive approach, grounded in Linear Algebra, using concepts such as eigenvalues, positive matrices and norms. Furthermore, since this algorithm is implemented computationally, an approximation method to calculate it, the adapted Power Method, will be shown.

Keywords: PageRank algorithm. Eigenvalues. Stochastic matrix. Power Method.

1 Introdução

O PageRank foi um algoritmo criado em 1997 pelos doutorandos Larry Page e Sergey Brinn. O objetivo dos estudantes era criar um algoritmo que classificasse páginas advindas de resultados de buscadores Web. Nesse sentido, a classificação é essencial, pois faz com que páginas relevantes sejam mostradas primeiramente ao usuário, reduzindo o tempo necessário para encontrar algo. A fim de implementar o algoritmo, Larry Page e Sergey Brinn criaram o buscador Google, o qual se transformou em uma das empresas mais ricas da história.

Primeiramente, é preciso entender como funciona um buscador Web. Suponha que determinado usuário da Web deseje obter informações sobre o calendário acadêmico da Universidade Federal de Uberlândia. Desse modo, ele pode digitar em um buscador a frase “calendário acadêmico UFU”. As palavras “calendário”, “acadêmico” e “UFU” são palavras-chaves e serão usadas pelo buscador.

Neste processo, o primeiro passo feito pelo buscador é agrupar todas as páginas encontradas com as palavras-chaves. Após isso, o programa deve listar o resultado para o usuário. O PageRank foi desenvolvido para atuar nesse segundo passo, visto que o algoritmo fornece uma pontuação para cada página do conjunto das páginas agrupadas. Por fim, o buscador lista as páginas em ordem decrescente de pontuação exibindo-as na tela.

Assim, o PageRank tem o objetivo de entender quais páginas são mais relevantes para o usuário e, portanto, devem ter uma pontuação maior. O funcionamento do algoritmo é baseado em três conceitos intuitivos e usados cotidianamente: indicação, confiabilidade e certeza. Com esses conceitos e algumas ferramentas matemáticas, foi possível a criação de um algoritmo que tornasse as buscas na Web mais rápidas e com melhor qualidade. Na próxima seção, esses conceitos serão discutidos com maiores detalhes.

O intuito do trabalho não é explicar o funcionamento do Google nos dias de hoje, e sim apresentar os conceitos matemáticos envolvidos na ideia inicial de Page *et al* (1998) para o desenvolvimento do algoritmo PageRank, para maiores detalhes consulte a referência 4, o qual fez com que o Google se tornasse um dos mais eficientes buscadores da internet.

2 Traçando o algoritmo

O primeiro conceito a ser entendido é o de indicação. Na Web essa ação é traduzida através dos hyperlinks, também conhecidos como links. Nesse sentido, uma página 1 indica uma página 2 se 1 possui um hyperlink para 2. Dessa forma, é importante que a quantidade de hyperlinks que uma página recebe influencie a sua pontuação. Além disso, impõe-se que uma página não pode possuir um hyperlink para si mesma. É importante ressaltar que essa ideia já era utilizada na análise da relevância de trabalhos acadêmicos disponibilizados digitalmente na época, segundo Page e Brinn. Nesse caso, o cálculo da pontuação de determinado trabalho era dado pelo somatório da quantidade de outros trabalhos acadêmicos que o citavam. No entanto, segundo Page e Brinn essa ideia não poderia ser exclusivamente aplicada as demais páginas da Web. Isso porque, segundo eles, diferentemente dos trabalhos acadêmicos que possuíam revisamento de conteúdo feito por humanos e alto grau de qualidade, as páginas da Web cresciam sem qualquer supervisão ou compromisso com a qualidade da informação.

Dessa forma, é preciso inserir no algoritmo a ideia de confiabilidade, o objetivo era que páginas bem classificadas fossem aquelas indicadas por outras páginas também bem classificadas. Para modelar isso matematicamente, vale-se o uso do exemplo a seguir:

Neste exemplo, o sentido das setas mostra a indicação, então, por exemplo, 1 possui um hiperlink para as páginas 2, 4 e 5. Considere então x_1, x_2, x_3, x_4 e x_5 as pontuações das páginas 1, 2, 3, 4 e 5, respectivamente. A pontuação de cada página será a soma das pontuações das páginas que a indicaram. Desse modo:

$$\begin{cases} x_1 = x_2 \\ x_2 = x_1 + x_3 + x_5 \\ x_3 = x_2 + x_5 \\ x_4 = x_1 + x_3 \\ x_5 = x_1 + x_4 \end{cases}$$

Esse modelo, no entanto, não é adequado, já que acumula pontuações. No exemplo acima, a página 2 ao indicar a página 3, fornece sua pontuação a ela, a qual por sua vez terá uma pontuação maior do que a página 2, já que $x_5 > 0$. Contudo, não parece justo a página 3 ser melhor classificada que 2, visto que ambas foram indicadas pela página 5 e, além disso, a página 2 teve uma indicação a mais, vinda da página 1. Assim, é importante a aplicação do conceito de certeza.

Para solucionar o problema da acumulação de pontos, sempre que uma determinada página j indicar mais de uma página, cada uma das páginas indicadas devem receber o mesmo acréscimo de pontuação, porém menor do que a pontuação da página j que as indicaram. Elas devem receber a mesma pontuação, pois considera-se que j , ao fazer mais de uma indicação, não tem preferência por nenhuma das páginas. Além disso, impor que a pontuação seja menor, impede que uma determinada página se auto prejudique. Assim, se uma página j indica n páginas, então cada página de 1 até n receberá um acréscimo de $\frac{1}{n}$ na sua pontuação. Representando em forma de sistemas e na forma matricial $BX = X$, tem-se:

$$\begin{cases} x_1 = \frac{x_2}{2} \\ x_2 = \frac{x_1}{3} + \frac{x_3}{2} + \frac{x_5}{2} \\ x_3 = \frac{x_2}{2} + \frac{x_5}{2} \\ x_4 = \frac{x_1}{3} + \frac{x_3}{2} \\ x_5 = \frac{x_1}{3} + \frac{x_4}{1} \end{cases} \iff \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{3} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}$$

Assim, o problema consiste em encontrar um autoespaço da matriz B , associado ao autovalor 1, denotado por $Aut_B(1)$. Além disso, o autovetor a ele associado precisa ter todas as **coordenadas estritamente positivas**. Para esse exemplo, é possível encontrar o polinômio característico de B , denotado por $p_B(\lambda)$, e verificar se 1 é raiz do polinômio e, por fim, encontrar um autovetor com coordenadas positivas, se possível. Fazendo isso, encontra-se o seguinte autovetor:

$$\left(\frac{9}{14}, \frac{9}{7}, \frac{8}{7}, \frac{11}{14}, 1 \right).$$

Dessa forma, as páginas ficam classificadas da seguinte maneira: 2, 3, 5, 4 e 1. Observe a classificação, e perceba que ela faz sentido. De fato, a melhor classificada foi a página 2, já que

recebeu mais indicações. Em seguida, vem a página 3 que, apesar de ter recebido duas indicações, uma delas foi da bem classificada página 2.

Esse exemplo motiva as definições da próxima subseção.

2.1 Matriz coluna estocástica

Antes de iniciar a análise das matrizes coluna estocástica, é preciso explicar algumas observações acerca das notações e dos espaços trabalhados. Os espaços vetoriais trabalhados são $\mathbb{R}^n$ e $M_n(\mathbb{R})$. No entanto, será usada, com frequência, a matriz de coordenadas de um elemento em $\mathbb{R}^n$ em relação à base canônica do $\mathbb{R}^n$. Nesse sentido um vetor qualquer $\vec{x}$ em $\mathbb{R}^n$ é representado por uma letra minúscula sob uma seta e sua matriz de coordenada é representada pela mesma letra em maiúscula. Assim, a notação $X \in \mathbb{R}^n$ simboliza, na verdade, um vetor coluna.

Definição 1 *Sejam $\mathbb{W}$ um conjunto de n páginas agrupadas de uma pesquisa na Web e $x_1, x_2, \dots, x_n$ as pontuações, respectivamente, das páginas 1, 2, ..., n . Essas pontuações são escritas em um vetor $\vec{x} \in \mathbb{R}^n$, em que $\vec{x} = (x_1, x_2, \dots, x_n)$.*

Definição 2 *Seja $\mathbb{W}$ um conjunto de n páginas agrupadas de uma pesquisa na Web. Considere a matriz $A = (a_{ij})_{n \times n} \in M_n(\mathbb{R})$, tal que*

$$a_{ij} = \begin{cases} 0, & \text{se } j \text{ não indica } i; \\ \frac{1}{n_j}, & \text{caso contrário, onde } n_j \text{ é o número de páginas que } j \text{ indicou.} \end{cases}$$

Observe que a soma das entradas de cada coluna da matriz A é 1 ou 0.

De fato, fixe uma coluna k e considere que a página k indicou m outras páginas, com $m \neq 0$. Logo, na coluna k , há m entradas com valor $\frac{1}{m}$ e $(i - m)$ entradas com valores 0. A soma dessas entradas é, portanto, $\frac{1}{m} \cdot m + (i - m) \cdot 0 = 1$. No entanto, se k não indicou nenhuma página, então a soma das entradas da coluna k é $k \cdot 0 = 0$.

Definição 3 *Seja $B = (b_{ij})_{n \times n} \in M_n(\mathbb{R})$ uma matriz como na Definição 2, se a soma dos elementos de cada coluna de B for igual a 1, dizemos que B é uma matriz coluna estocástica.*

As matrizes coluna estocástica podem ser utilizadas na modelagem de um tipo de sistema dinâmico denominado Cadeia de Markov. Os sistemas dinâmicos são caracterizados por variáveis cujo valores dependem do tempo, por exemplo. Há sistemas dinâmicos denominados processos estocásticos em que os fenômenos que os influenciam são aleatórios. Dessa maneira, é preciso utilizar o conceito de variáveis aleatórias. A Cadeia de Markov é um tipo de processo estocástico no qual os valores que as variáveis assumem em determinado instante de tempo dependem exclusivamente do valor assumido no instante imediatamente anterior. Anton e Rorres (2012) mostram algumas situações que podem ser modeladas utilizando Cadeia de Markov na referência 1.

Neste trabalho serão consideradas apenas as matrizes coluna estocástica, excluindo-se os casos em que a matriz associada a uma Web possui pelo menos uma coluna com soma zero, isto é, toda página k possui pelo menos um hyperlink para outra página.

A proposição abaixo mostra uma propriedade importante das matrizes coluna estocástica.

Proposição 4 *Seja B uma matriz coluna estocástica, então 1 é um autovalor de B .*

Prova: Seja $B^T = (b'_{ij})_{n \times n}$ a matriz transposta de B .

Note que a matriz $(B - \lambda I)^T = B^T - \lambda I$. Desse modo,

$$p_B(\lambda) = \det(B - \lambda I) = \det(B^T - \lambda I) = p_{B^T}(\lambda).$$

Logo, as matrizes B e B^T possuem os mesmos autovalores.

Assim, para provar que 1 é um autovalor da matriz B , basta provar que é autovalor de B^T . Para isso, basta encontrar um único vetor não nulo $X = (x_{i1})_{n \times 1}$, tal que $B^T X = X$.

Tome o vetor X , tal que $x_{i1} = 1, \forall i \in \{1, \dots, n\}$ e considere $B^T X = (c_{i1})_{n \times 1}$.

Assim, tem-se que $c_{i1} = \sum_{j=1}^n b'_{ij} \cdot x_{j1} = \sum_{j=1}^n b'_{ij} \cdot 1 = \sum_{j=1}^n b'_{ij}, \forall i \in \{1, \dots, n\}$.

Como $b'_{ij} = b_{ji}, \forall i, j \in \{1, \dots, n\}$, e sendo B , por hipótese, uma matriz coluna estocástica

$$c_{i1} = \sum_{j=1}^n b'_{ij} = \sum_{j=1}^n b_{ji} = 1 = x_{i1} \Leftrightarrow B^T X = X.$$

Portanto, foi provado que 1 é um autovalor de B^T e, portanto, de B . ■

2.2 Falha do algoritmo

A Proposição 4 mostra que toda matriz coluna estocástica tem 1 como autovalor. No entanto, a fim de que esse algoritmo tenha função prática, é preciso que $\dim(Aut_B(1)) = 1$ e que exista um autovetor com todas as coordenadas positivas. Porém, isso nem sempre ocorre, como ilustrado no exemplo a seguir:

Construindo a matriz B , como na definição 2, desse conjunto, tem-se:

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & \frac{1}{2} & 0 \end{bmatrix}$$

Fazendo os cálculos, tem-se que $Aut_B(1) = \left\{ \vec{x} \in \mathbb{R}^5 \mid \vec{x} = \alpha(1, 1, 0, 0, 0) + \beta\left(0, 0, \frac{3}{4}, \frac{1}{2}, 1\right), \alpha, \beta \in \mathbb{R} \right\}$.

Como visto no exemplo anterior, nem sempre o subespaço próprio $Aut_B(1)$ é unidimensional, fazendo com que haja ambiguidade no resultado do algoritmo.

De fato, é possível provar que se uma Web é desconexa e formada por m subwebs então $\dim \text{Aut}_B(1) \geq m$. Essa demonstração é feita por Batti (2015) na referência 2. Assim, não se pode utilizar diretamente a Definição 3 para o algoritmo PageRank.

A solução para esta questão pode ser encontrada em tópicos da Álgebra Linear que estudam matrizes cujas entradas são positivas. De fato, o Teorema de Perron-Frobenius, muito utilizado em várias aplicações, como a Cadeia de Markov, afirma que matrizes positivas possuem autovetores com entradas positivas. Além disso, esse teorema garante que o autoespaço de uma matriz positiva associado ao seu autovalor de maior magnitude é unidimensional. Mais detalhes sobre tal teorema é feito por Soares (2018) na referência 5. No entanto, neste trabalho, não será necessário o estudo do caso geral, apenas serão provados alguns resultados particulares para matrizes coluna estocástica.

3 Corrigindo a falha do algoritmo

Nesta seção são apresentados os resultados importantes para auxiliar em uma melhora na falha do algoritmo, e a definição da Matriz Google.

3.1 Matriz positiva

Definição 5 Seja $B = (b_{ij})_{n \times n} \in M_n(\mathbb{R})$, B é dita matriz positiva quando $b_{ij} > 0, \forall i, j \in \{1 \dots, n\}$.

Proposição 6 Seja $B = (b_{ij})_{n \times n} \in M_n(\mathbb{R})$ uma matriz positiva e coluna estocástica, então todo autovetor de B associado ao autovalor 1 tem todas as entradas positivas ou todas negativas.

Prova: A fim de provar essa proposição, usa-se a desigualdade triangular para uma quantidade de k parcelas.

Seja $X = (x_{i1})_{n \times 1}$ um autovetor de B associado ao autovalor 1, então para cada $i = 1 \dots, n$, verifica-se que $x_{i1} = \sum_{j=1}^n b_{ij} \cdot x_{j1}$.

Aplicando o módulo em ambos os lados da igualdade anterior, obtém-se $|x_{i1}| = \left| \sum_{j=1}^n b_{ij} \cdot x_{j1} \right|$.

Supondo, agora, que existam coordenadas de sinais opostos e usando a desigualdade triangular para n parcelas, lembrando que as entradas de B são não negativas, tem-se que

$$|x_{i1}| < \sum_{j=1}^n |b_{ij} \cdot x_{j1}| = \sum_{j=1}^n |b_{ij}| \cdot |x_{j1}| = \sum_{j=1}^n b_{ij} \cdot |x_{j1}|.$$

Fazendo a soma dessas n desigualdades, fica-se com

$$\sum_{i=1}^n |x_{i1}| < \sum_{i=1}^n \sum_{j=1}^n b_{ij} \cdot |x_{j1}| = \sum_{j=1}^n \sum_{i=1}^n b_{ij} \cdot |x_{j1}| = \sum_{j=1}^n |x_{j1}| \cdot \sum_{i=1}^n b_{ij} = \sum_{j=1}^n |x_{j1}|.$$

O que é um absurdo, logo, não há coordenadas com sinais opostos.

Resta provar que não há coordenadas iguais a zero.

Suponha que exista um autovetor $X = (x_{i1})_{n \times 1}$, com uma coordenada $x_{k1} = 0$.

Assim, $x_{k1} = 0 = \sum_{i=1}^n b_{ik} \cdot x_{k1}$, como $b_{ik} > 0$ e x_{kj} são ou não negativos ou não positivos, ambos para todo $k \in \{1, \dots, n\}$, então $x_{k1} = 0, \forall k \in \{1, \dots, n\}$, logo X é um vetor nulo.

Como o vetor nulo não é autovetor, a proposição está provada. ■

Proposição 7 *Sejam $B = (b_{ij})_{n \times n} \in M_n(\mathbb{R})$ e $X, Y \in \mathbb{R}^n$, de modo que $\{X, Y\}$ é linearmente independente. Sob essas condições, é possível encontrar $\alpha, \beta \in \mathbb{R}$, não simultaneamente nulos, tal que $Z = \alpha X + \beta Y$ possua pelo menos duas entradas com sinais opostos.*

Prova: Para provar essa proposição, é preciso utilizar o fato de que uma soma com mais de uma parcela é igual a zero se, e somente se, então todos os elementos são zero ou há elementos positivos e negativos. Assim, para provar esta proposição, basta determinar α e β tais que a soma das entradas de Z seja 0.

Tome $X = (x_{i1})_{n \times 1}, Y = (y_{i1})_{n \times 1}$ e $Z = (z_{i1})_{n \times 1}$, vetores coluna.

Considere $\sum_{i=1}^n x_{i1}$. Se $\sum_{i=1}^n x_{i1} = 0$, então tome $\alpha = 1$ e $\beta = 0$. Assim,

$$\sum_{i=1}^n z_{i1} = \sum_{i=1}^n (\alpha x_{i1} + \beta y_{i1}) = \alpha \sum_{i=1}^n x_{i1} + \beta \sum_{i=1}^n y_{i1} = \alpha \sum_{i=1}^n x_{i1} = 0.$$

Caso contrário, se $\sum_{i=1}^n x_{i1} \neq 0$, tome $\beta = 1$ e $\alpha = \frac{-\sum_{i=1}^n y_{i1}}{\sum_{i=1}^n x_{i1}}$. Dessa maneira a soma das entradas

de $Z = \alpha X + \beta Y$ é

$$\sum_{i=1}^n z_{i1} = \sum_{i=1}^n \left(\frac{-\sum_{i=1}^n y_{i1}}{\sum_{i=1}^n x_{i1}} \cdot x_{i1} + y_{i1} \right) = \left(\frac{-\sum_{i=1}^n y_{i1}}{\sum_{i=1}^n x_{i1}} \right) \cdot \sum_{i=1}^n x_{i1} + \sum_{i=1}^n y_{i1} = -\sum_{i=1}^n y_{i1} + \sum_{i=1}^n y_{i1} = 0.$$

■

Com as Proposições 6 e 7 é provado o seguinte teorema:

Teorema 8 *Seja B uma matriz positiva coluna estocástica, então $\dim \text{Aut}_B(1) = 1$*

Prova: Suponha, por absurdo, que existam $X, Y \in \mathbb{R}^n$ autovetores, associados ao autovalor 1, tais que o conjunto $\{X, Y\}$ é linearmente independente.

Pela Proposição 7, existem α e $\beta \in \mathbb{R}$ tais que $Z = \alpha X + \beta Y$ tenha entradas com sinais opostos.

Além disso, como Z é um autovetor de B , pela Proposição 6, suas coordenadas têm o mesmo sinal.

Por esta contradição, chega-se que a hipótese é falsa, isto é, não existem X e Y tais que $\{X, Y\}$ é linearmente independente.

Assim, qualquer conjunto com mais de um elemento é linearmente dependente.

Portanto, $\dim \text{Aut}_B(1) = 1$. ■

3.2 Matriz Google

Perceba que o Teorema 8 fornece uma ferramenta importante nas classificações de páginas. Isso porque ao provar que o autoespaço associado a 1 de uma matriz positiva e coluna estocástica é unidimensional, ele possibilita uma única classificação para páginas. Ademais, a Proposição 6 garante a existência de autovetores com coordenadas não negativas, uma das condições das pontuações das páginas.

No entanto, até agora, a matriz do algoritmo não é positiva, já que algumas entradas podem ser 0. Assim, a seguir, será definida uma matriz G , a qual será a matriz B modificada de maneira que todas as entradas 0 de B sejam substituídas por entradas maiores do que 0, mas ao mesmo tempo próximas de 0. Desse modo, será possível utilizar o Teorema 8 e trabalhar com a matriz G no PageRank.

Definição 9 Seja $B = (b_{ij})_{n \times n} \in M_n(\mathbb{R})$ uma matriz como na Definição 3, e considere $S = (s_{ij})_{n \times n} \in M_n(\mathbb{R})$, em que $s_{ij} = \frac{1}{n}$, $\forall i, j \in \{1, \dots, n\}$. Defina-se $G = (g_{ij})_{n \times n}$, como

$$G = (1 - m) \cdot B + m \cdot S, \text{ onde } 0 < m < 1, \text{ isto é, } g_{ij} = (1 - m) \cdot b_{ij} + m \cdot s_{ij}.$$

A matriz G será chamada **Matriz Google**.

Perceba, que a matriz S da definição 9 é coluna estocástica.

Agora, é preciso mostrar que a Matriz Google se encaixa no Teorema 8, isto é, que G é uma matriz positiva e coluna estocástica.

Como $g_{ij} = (1 - m) \cdot b_{ij} + m \cdot s_{ij}$, fixando uma coluna k qualquer, como B e S são coluna estocásticas, chega-se em

$$\begin{aligned} \sum_{i=1}^n g_{ik} &= \sum_{i=1}^n [(1 - m) \cdot b_{ik} + m \cdot s_{ik}] = \sum_{i=1}^n (1 - m) \cdot b_{ik} + \sum_{i=1}^n m \cdot s_{ik} = \\ &= \sum_{i=1}^n b_{ik} - m \sum_{i=1}^n b_{ik} + m \sum_{i=1}^n s_{ik} = 1 - m + m = 1. \end{aligned}$$

Assim, a matriz G é coluna estocástica.

Agora, é preciso provar que ela é positiva.

Como $(1 - m) \cdot b_{ij} \geq 0$ e $m \cdot s_{ij} > 0$, então $(1 - m) \cdot b_{ij} + m \cdot s_{ij} > 0$.

Portanto, foi provado que a matriz G é positiva e coluna estocástica, implicando, pelo teorema 8, que $\dim \text{Aut}_G(1) = 1$. Além disso, pela Proposição 6, G possui um autovetor com coordenadas estritamente positivas.

Um fato importante sobre a matriz G é o valor de m . Como foi tratado, m pertence ao intervalo $(0, 1)$. O valor escolhido pelo Google, inicialmente, era $m = 0.15$. Tal valor está relacionado a velocidade de convergência da sequência das matrizes G , assunto a ser discutido na próxima seção.

4 Cálculo do autovetor das pontuações

Para calcular todos os autovetores λ de uma matriz A , pode-se usar seu polinômio característico $p_A(\lambda)$. No entanto, considerando que a Web pode ter bilhões de páginas, o uso dessa ferramenta torna-se inviável, já que o grau de $p_A(\lambda)$ seria na casa dos bilhões. Assim, a fim de calcular o autovetor das pontuações é preciso usar um método de aproximação por iteração.

Neste sentido, um método iterativo muito utilizado para o cálculo de autovetores específicos de matrizes é o **Método das Potências**. Nesse método,

$$\lim_{k \rightarrow \infty} A^k X_0 = Q, \text{ também escrito como } \lim_{k \rightarrow \infty} AX_k = Q, \text{ em que } X_{k+1} = AX_k, k \geq 0,$$

no qual X_0 é, geralmente, um vetor qualquer e Q é um autovetor associado ao autovalor de A de maior módulo, sendo A diagonalizável.

Visto que, para aplicar o método das potências, é preciso que a matriz seja diagonalizável, então não é possível aplicá-lo diretamente na Matriz Google, uma vez que ela pode não ser diagonalizável. No entanto, pode-se adaptar esse método, utilizando suas ideias principais.

Ao trabalhar com limites de matrizes, uma ferramenta importante é definir a norma vetorial para matrizes de n linhas e uma coluna, como definida a seguir:

Definição 10 Seja $X = (x_{i1})_{n \times 1} \in M_{n \times 1}(\mathbb{R})$, define-se $\|X\|_1 = \sum_{i=1}^n |x_{i1}|$.

$\|X\|_1$ é chamado norma 1 de X .

Sejam $Q = (q_{i1})_{n \times 1}$, $X = (x_{i1})_{n \times 1} \in M_{n \times 1}(\mathbb{R})$, tome $V = Q - X$. Perceba, portanto, que quando X se aproxima de Q , isto é, cada x_{i1} se aproxima de q_{i1} , $\|V\|_1$ se aproxima de 0.

Assim, considerando a ideia presente na Definição 10, a fórmula do Método das Potências e a especificidade da Matriz Google, queremos tomar um vetor X_0 , aplicar um método iterativo da forma $X_{k+1} = GX_k$, de modo que $\lim_{k \rightarrow \infty} \|Q - X_{k+1}\|_1 = 0$. Mas,

$$\begin{aligned} \lim_{k \rightarrow \infty} \|Q - X_{k+1}\|_1 = 0 &\iff \lim_{k \rightarrow \infty} \|GQ - G^{k+1}X_0\|_1 = 0 \iff \\ &\iff \lim_{k \rightarrow \infty} \|G(Q - G^kX_0)\|_1 = 0 \iff \lim_{k \rightarrow \infty} \|G(Q - X_k)\|_1 = 0. \end{aligned}$$

Simplificando a expressão, tome $V = Q - X_k$. Como o $\lim_{k \rightarrow \infty} \|GV\|_1 = 0$, é natural estudar o produto $\|GV\|_1 = 0$. Neste produto, como G é positiva e coluna estocástica, conclui-se que $\sum_{i=1}^n v_{i1} = 0$. Isso motiva o seguinte estudo:

Proposição 11 Seja $G \in M_n(\mathbb{R})$ uma matriz positiva e coluna estocástica. Seja $\mathbb{U}$ um subespaço de $\mathbb{R}^n$ tal que $\vec{v} \in \mathbb{U}$ se, e somente, se $\sum_{i=1}^n v_{i1} = 0$. Sob essas condições, pode-se afirmar que

- (i) $GV \in \mathbb{U}$, para todo $V \in \mathbb{U}$.
- (ii) $\|GV\|_1 \leq c\|V\|_1$, para todo $V \in \mathbb{U}$, com $c = \max_{1 \leq j \leq n} |1 - 2 \cdot \min_{1 \leq i \leq n} g_{ij}| < 1$.
- (iii) $\|G^kV\|_1 \leq c^k\|V\|_1$, para todo $k \in \mathbb{N} - \{0\}$.

Prova: (i) Considere $GV = (h_{i1})_{n \times 1}$, logo é preciso provar que $\sum_{i=1}^n h_{i1} = 0$.

Sabendo que $h_{i1} = \sum_{j=1}^n (g_{ij} \cdot v_{j1})$ e que $\sum_{i=1}^n g_{ij} = 1$ para $j = 1, \dots, n$, obtém-se que

$$\sum_{i=1}^n h_{i1} = \sum_{i=1}^n \sum_{j=1}^n (g_{ij} \cdot v_{j1}) = \sum_{j=1}^n \sum_{i=1}^n (g_{ij} \cdot v_{j1}) = \sum_{j=1}^n v_{j1} \cdot \sum_{i=1}^n g_{ij} = \sum_{j=1}^n v_{j1} = 0.$$

(ii) $\|GV\|_1 = \sum_{i=1}^n |h_{i1}|$. Sem perda de generalidade, considere que as m primeiras entradas de GV são positivas e as outras $n - m$ são negativas. Perceba que $m \neq n$, pois pelo item (i) a soma de suas coordenadas é zero, logo há coordenadas com sinais opostos. Assim,

$$\sum_{i=1}^n |h_{i1}| = \sum_{i=1}^m h_{i1} - \sum_{i=m+1}^n h_{i1} = \sum_{i=1}^m \sum_{j=1}^n (g_{ij} \cdot v_{j1}) - \sum_{i=m+1}^n \sum_{j=1}^n (g_{ij} \cdot v_{j1}).$$

Ao inverter a ordem dos somatórios, tem-se v_{j1} constante: $\sum_{j=1}^n v_{j1} \sum_{i=1}^m g_{ij} - \sum_{j=1}^n v_{j1} \sum_{i=m+1}^n g_{ij}$.

Colocando $\sum_{j=1}^n v_{j1}$ em evidência, tem-se $\sum_{j=1}^n v_{j1} \left(\sum_{i=1}^m g_{ij} - \sum_{i=m+1}^n g_{ij} \right)$. Além disso, ao analisar a expressão $\sum_{i=1}^m g_{ij} - \sum_{i=m+1}^n g_{ij}$, conclui-se que

$$-1 < \sum_{i=1}^m g_{ij} - \sum_{i=m+1}^n g_{ij} < 1, \text{ pois } GV \in \mathbb{U} \text{ e } m, n \neq 0.$$

Denotando $\sum_{i=1}^m g_{ij} - \sum_{i=m+1}^n g_{ij} = a_{ij}$, é válido estudar os valores extremos de a_{ij} , isto é, o maior e menor valor que assume para cada $j \in \{1, \dots, n\}$.

O maior valor é encontrado quando GV possui $m = n - 1$ coordenadas positivas ou nulas. Nessas circunstâncias: $a_{ij} = (1 - \min_{1 \leq i \leq n} g_{ij}) - \min_{1 \leq i \leq n} g_{ij} = 1 - 2 \cdot \min_{1 \leq i \leq n} g_{ij}$.

Já o menor valor é encontrado quando GV possui $n - 1$ coordenadas negativas. Nessas circunstâncias: $a_{ij} = -1 + 2 \cdot \min_{1 \leq i \leq n} g_{ij}$.

$$\text{Portanto, } -1 < -1 + 2 \cdot \min_{1 \leq i \leq n} g_{ij} \leq a_{ij} \leq 1 - 2 \cdot \min_{1 \leq i \leq n} g_{ij} < 1 \iff |a_{ij}| \leq \left| 1 - 2 \cdot \min_{1 \leq i \leq n} g_{ij} \right|.$$

Assim, utilizando o resultado acima e aplicando a desigualdade triangular

$$\sum_{i=1}^n |h_{i1}| = \sum_{j=1}^n v_{j1} \cdot a_{ij} = \left| \sum_{j=1}^n v_{j1} \cdot a_{ij} \right| \leq \sum_{j=1}^n |v_{j1} \cdot a_{ij}| = \sum_{j=1}^n |v_{j1}| \cdot |a_{ij}|.$$

Fixando $c = \max_{1 \leq j \leq n} a_{ij} = \max_{1 \leq j \leq n} |1 - 2 \cdot \min_{1 \leq i \leq n} g_{ij}|$, tem-se que

$$\sum_{j=1}^n |v_{j1}| \cdot |a_{ij}| \leq \sum_{j=1}^n |v_{j1}| \cdot c = c \sum_{j=1}^n |v_{j1}| = c \|V\|_1.$$

(iii) Esse item pode ser provado através da indução finita.

Primeiramente, é preciso mostrar que $G^{k-1}V \in \mathbb{U}$.

Para $k = 1$ é verdade pelo item (i).

Suponha que vale para $k - 2$, isto é, $G^{k-2}V \in \mathbb{U}$, então $G^{k-1}V = G(G^{k-2}V) \in \mathbb{U}$ pelo item (i) e pela hipótese da indução.

Agora, pode-se provar que $\|G^k V\|_1 \leq c^k \|V\|_1$.

Para $k = 1$ é verdade pelo item (i).

Suponha que vale para $k - 1$, então pelo resultado anterior

$$\|G^k V\|_1 = \|G(G^{k-1} V)\|_1 \leq c \|G^{k-1} V\|_1 \leq c \cdot c^{k-1} \|V\|_1 = c^k \|V\|_1.$$

■

Com esses resultados, pode-se enunciar e provar a seguinte proposição:

Proposição 12 *Seja $G \in M_n(\mathbb{R})$ uma Matriz Google e Q um autovetor de coordenadas estritamente positivas com as pontuações das n páginas. Seja $X \in M_{n \times 1}(\mathbb{R})$, de coordenadas estritamente positivas, com $\|X\|_1 = \|Q\|_1$. Assim, $Q = \lim_{k \rightarrow \infty} G^k X$.*

Prova: Para qualquer vetor X não nulo, pode-se encontrar um autovetor de pontuações Q com mesma norma 1. De fato, se Q é um autovetor de pontuações qualquer, basta multiplicar cada entrada por $\frac{\|X\|_1}{\|Q\|_1}$.

Sendo assim, dados X e Q de coordenadas estritamente positivas e com mesma norma 1, é possível encontrar $V \in \mathbb{U}$ de forma que $Q = V + X$.

Pelo item (iii) da Proposição 11, $0 \leq \|G^k V\|_1 \leq c^k \|V\|_1$.

Como $\lim_{k \rightarrow \infty} c^k = 0$, pois $0 < c < 1$ e $\lim_{k \rightarrow \infty} 0 = 0$, pelo Teorema do Confronto, $\lim_{k \rightarrow \infty} \|G^k V\|_1 = 0$.
Porém,

$$\begin{aligned} \lim_{k \rightarrow \infty} \|G^k V\|_1 = 0 &\Leftrightarrow \lim_{k \rightarrow \infty} \|G^k (Q - X)\|_1 = 0 \Leftrightarrow \lim_{k \rightarrow \infty} G^k (Q - X) = 0 \Leftrightarrow \lim_{k \rightarrow \infty} (G^k Q - G^k X) = 0 \Leftrightarrow \\ &\Leftrightarrow \lim_{k \rightarrow \infty} (Q - G^k X) = 0 \Leftrightarrow \lim_{k \rightarrow \infty} Q - \lim_{k \rightarrow \infty} G^k X = 0 \Leftrightarrow Q = \lim_{k \rightarrow \infty} G^k X. \end{aligned}$$

■

Como visto acima, quando k tende ao infinito, pode-se encontrar um autovetor com as pontuações das páginas. No entanto, um computador não pode executar os cálculos infinitamente. Por isso, ele utiliza um critério de parada. Normalmente, define-se um erro ϵ , e para cada iteração é feita a diferença entre os erros relativos ϵ_{k-1} e ϵ_k . Caso a diferença entre esses erros relativos for menor que o erro ϵ escolhido, o computador para de executar as iterações e o resultado $G^k X$ é tomado como a solução aproximada do problema.

A velocidade de convergência, isto é, o quanto de iterações são necessárias para que $\epsilon_{k-1} - \epsilon_k < \epsilon$, depende do valor escolhido para m . Como já foi dito, inicialmente, Page *et al* (1998) utilizaram $m = 0.15$, já que, segundo seus testes, esse era o valor que conciliava velocidade de convergência e um valor considerado adequado para Q . Bryan e Leise (2006) também tratam da velocidade de convergência na referência 3.

Outro fato importante sobre X é que se considerar $\|X\|_1 = 1$, pode-se computar o autovetor Q sem utilizar diretamente a matriz G . Nesse sentido, as novas matrizes manipuladas pelos computadores conterão mais entradas 0, tornando os cálculos mais rápidos.

Lembrando da Definição 9, $G = (1 - m).B + m.S$, seja X um vetor qualquer em que $\|X\|_1 = 1$, então $GX = ((1 - m).B + m.S)X = (1 - m).BX + m.SX = (1 - m).BX + m.Z$, onde $Z = \left(\frac{1}{n}\right)_{n \times 1}$.

Bryan e Leise (2006) também abordam essa questão na referência 3.

Na próxima seção, será retomado o exemplo inicial do trabalho, descrito na Seção 2. Usando as ideias desenvolvidas no artigo, como a Matriz Google e o método de iteração explicitado, será encontrado uma aproximação para um autovetor com as pontuações das 5 páginas.

5 Aplicando o algoritmo

Primeiramente, monta-se a matriz B para a determinada Web.

Como $n = 5$, a matriz G é dada por: $G = (0.85) \cdot B + (0.15) \cdot S$, com $S = \left(\frac{1}{5}\right)_{n \times n}$.

Dessa forma,

$$G = \begin{bmatrix} 0.030000 & 0.455000 & 0.030000 & 0.030000 & 0.030000 \\ 0.313333 & 0.030000 & 0.455000 & 0.030000 & 0.455000 \\ 0.030000 & 0.455000 & 0.030000 & 0.030000 & 0.455000 \\ 0.313333 & 0.030000 & 0.455000 & 0.030000 & 0.030000 \\ 0.313333 & 0.030000 & 0.030000 & 0.880000 & 0.030000 \end{bmatrix}$$

Começando as iterações pelo vetor $X = (1, 0, 0, 0, 0)$, foram feitas 19 iterações, algumas estão mostradas abaixo. Esses dados foram obtidos a partir do algoritmo exibido na Tabela 1 a seguir.

$$X_1 = \begin{bmatrix} 0.030000 \\ 0.313333 \\ 0.030000 \\ 0.313333 \\ 0.313333 \end{bmatrix} \quad X_6 = \begin{bmatrix} 0.1454 \\ 0.2491 \\ 0.2372 \\ 0.1573 \\ 0.2110 \end{bmatrix} \quad X_{13} = \begin{bmatrix} 0.1386 \\ 0.2562 \\ 0.2283 \\ 0.1665 \\ 0.2105 \end{bmatrix} \quad X_{19} = \begin{bmatrix} 0.1388 \\ 0.2559 \\ 0.2283 \\ 0.1663 \\ 0.2107 \end{bmatrix}.$$

Após o término dessas 19 iterações, o autovetor com as pontuações Q foi tomado como X_{19} , sendo, portanto, as páginas classificadas na seguinte ordem: 2, 3, 5, 4 e 1 exatamente como encontrado no exemplo inicial.

O algoritmo exibido na Tabela 1 foi programado utilizando o GNU Octave, software de código aberto com uma linguagem de programação própria que fornece ferramentas para análise numérica. O código deve ser escrito no editor do software e ser executado na sua janela de comando.

```
%MATRIZ B
B = [0,1/2,0,0,0;1/3,0,1/2,0,1/2;...
0,1/2,0,0,1/2;1/3,0,1/2,0,0;1/3,0,0,1,0];

%MATRIZ S
n = 5;
S= 1/n*ones(n);

%MATRIZ GOOGLE G
G = 0.85*B + 0.15*S;

%VETOR INICIAL
X0 = [1;0;0;0;0];

%PRIMEIRA ITERAÇÃO
X = G*X0;
n=1;

%DEFININDO O CRITÉRIO DE PARADA
e = 10e-5;
erro = 0;

%ERRO
for j=1:5
    soma = abs(X0(j) - X(j));
    erro = erro + soma;
endfor

%DEMAIS ITERAÇÕES
while erro > e
    erro = 0;
    Y = X;
    X0 = X;
    X = G*X0;
    for j=1:5
        soma = abs(X0(j) - X(j));
        erro = erro + soma;
    endfor
    n=n+1;
endwhile

%X:AUTOVETOR DAS PONTUAÇÕES
X,n
```

Tabela 1: Algoritmo utilizado.

6 Conclusão

Este artigo trouxe uma aplicação da Álgebra Linear na computação. Além disso, foi explicado detalhadamente um raciocínio para se modelar matematicamente um problema simplificado da realidade, já que considerando uma Web com cinco páginas, foi possível encontrar os valores que as pontuações das páginas assumiriam, e verificar se esses valores interpretavam bem a confiabilidade de uma página da Web, algo, a princípio, subjetivo. Ademais, foi feito um embasamento teórico, utilizando ramos da Álgebra Linear, a fim de encontrar uma nova matriz para o algoritmo que não alterasse a essência da ideia proposta inicialmente e que solucionasse os problemas de pontuações negativas e ambiguidades nas pontuações das páginas. Também foi explicado o porquê não seria possível encontrar um autovetor com as pontuações utilizando a definição de autovalores e, portanto, utilizando o polinômio característico. Além disso, como não era necessário determinar todos os autovalores da Matriz Google, foi exibido um método iterativo inspirado no Método das Potências. Por fim, na última seção, foi exibido um exemplo numérico em que é possível observar a convergência da sequência, bem como o resultado final da pontuação, o qual estava de acordo com o exemplo inicial.

7 Referências Bibliográficas

- [1] ANTON, H; RORRES, C. **Álgebra linear com aplicações**. 10. ed. Porto Alegre: Bookman, 2012.
- [2] BATTI, J. C. B. **Um pouco da matemática por trás do algoritmo PageRank do Google**. 2015. 61 f. Dissertação (Mestrado Profissional em Matemática)- Centro de Ciências Físicas e Matemáticas, Universidade Federal de Santa Catarina, Florianópolis, 2015.
- [3] BRYAN, K. ; LEISE, T. The 25,000,000,000 eigenvector: the linear algebra behind Google. **Society for Industrial and Applied Mathematics Review**, v. 48, n. 3, p. 569-581, 2006
- [4] PAGE, L. *et al.* **The PageRank citation ranking: bringing order to the web**. Technical Report. [Stanford]: Stanford InfoLab, 1998. Disponível em: <http://ilpubs.stanford.edu: 8090/422/>. Acesso em: 28 abr. 2021.
- [5] SOARES, J. **Um simples estudo da matemática por trás do Google**. 2018. 49 f. Dissertação (Mestrado Profissional em Matemática)- Programa de Mestrado Profissional em Matemática em Rede Nacional, Universidade Federal de São João del-Rei, São João del-Rei, 2018.