Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Departamento de Matemática
Licenciatura em Matemática a Distância

A Matemática do buscador do Google: Uma breve introdução

Diogo Micherlon Coelho da Rocha

Orientador: Prof. Dr. Nacib Gurgel Albuquerque

João Pessoa
Dezembro de 2023

Diogo Micherlon Coelho da Rocha

A Matemática do buscador do Google: uma breve introdução

Trabalho de Conclusão de Curso apresentado ao departamento de Matemática da Universidade Federal da Paraíba, como parte integrante dos requisitos necessários para obtenção do título de Licenciado em Matemática.

Orientador: Prof. Dr. Nacib Gurgel Albuquerque

João Pessoa, dezembro de 2023

Catalogação na publicação
Seção de Catalogação e Classificação

R672m Rocha, Diogo Micherlon Coelho da.
A matemática do buscador do Google : uma breve introdução / Diogo Micherlon Coelho da Rocha. - João Pessoa, 2023.
37 p. : il.

Orientação: Nacib Gurgel Albuquerque.
TCC (Curso de Licenciatura em Matemática) - Educação a Distância, Polo João Pessoa-PB - UFPB/CCEN.

1. Matemática. 2. Cadeias de Markov. 3. Ponto fixo de Banach. 4. PageRank. 5. Teoria da Probabilidade. 6. Algebra Linear. 7. Buscador do Google. I. Albuquerque, Nacib Gurgel. II. Título.

UFPB/CCEN CDU 51(043.2)

Elaborado por Josélia Maria Oliveira da Silva - CRB-15/113

Universidade Federal da Paraíba
Unidade de Educação a Distância
Centro de Ciências Exatas e da Natureza
Departamento de Matemática
Coordenação de Licenciatura em Matemática a Distância

Ata da defesa do Trabalho de Conclusão de Curso (TCC), do Licenciando Diogo Micherlon Coelho da Rocha

Aos treze dias do mês de dezembro de dois mil e vinte e três, às nove horas, no polo de apoio presencial da cidade de João Pessoa, em sessão pública, teve início à defesa de Monografia (Trabalho de Conclusão de Curso) do Licenciando **Diogo Micherlon Coelho da Rocha**, intitulada "**A Matemática do buscador do Google: uma breve introdução**". O licenciando cumpriu com o requisito parcial para a obtenção do grau de licenciado em Matemática. Procedeu a defesa diante da Comissão Examinadora constituída pelos seguintes professores: Nacib André Gurgel e Albuquerque, na condição de orientador, Ricardo Burity Croccia Macedo e Renato Burity Croccia Macedo, examinadores. O professor Nacib André Gurgel e Albuquerque, na condição de Presidente, dirigiu os trabalhos e, após as formalidades de praxe, convidou o candidato a discorrer sobre o conteúdo da sua Monografia. Concluída a explanação, o candidato foi arguido pela Comissão Examinadora. Em seguida, a referida comissão reuniu-se para deliberar e atribuir, por unanimidade, a menção **aprovado com nota 8,5 (oito vírgula cinco)**. Nada mais havendo a tratar, foi encerrada a sessão e, para constar, lavrei a presente ata que, depois de lida e aprovada, será assinada por mim e pelos demais membros da Comissão Examinadora.

Membros da Comissão:

Prof. Nacib André Gurgel e Albuquerque
Presidente

Prof. Ricardo Burity Croccia Macedo
Examinador

Prof. Renato Burity Croccia Macedo
Examinador

Agradecimentos

Primeiramente a Deus, pois tudo o que acontece em minha vida é devido a Ele. Sempre tive dificuldades durante a graduação, mas graças a Deus consegui superá-las.

Aos meus pais, Maria do Carmo Coelho Lima da Rocha e Dionizio Marcos da Rocha (in memoriam), por estarem ao meu lado em todos os momentos me apoiando sempre nas minhas decisões e por não desistir de mim quando nem eu mesmo acreditava que poderia conseguir, que juntos lutaram para me proporcionar uma boa educação. Sem vocês, eu nada seria!

À Daniela Monteiro, minha companheira, por ter travado esta batalha ao meu lado, ter me dado força, ter sido meu porto seguro, e ter lutado junto comigo contra dores físicas e psicológicas que enfrentei durante essa licenciatura.

Aos meus filhos, Daphinie Monteiro, Dandhara Monteiro e Danilo Monteiro por proporcionarem um amor inigualável, também me tornarem um homem mais forte, foi incontável a vezes que tive que estudar, fazer prova com um de vocês no colo e esta é a minha maior motivação, proporcionar o melhor de mim para vocês.

Aos meus amigos de graduação que compartilharam comigo seus conhecimentos e suas dificuldades para que juntos pudéssemos crescer. Como também, aqueles que ensinei ou ajudei de alguma forma e me fizeram perceber, por muitas vezes, que suas dúvidas também eram minhas.

Aos professores do Departamento de Matemática - UFPB pelos conhecimentos transmitidos.

Por fim agradeço a Universidade Federal da Paraíba - UFPB, que foi chave de minha formação, abrindo portas para meu crescimento.

Muito Obrigado!

ROCHA, Diogo. **A Matemática do buscador do *Google*: uma breve introdução**. Universidade Federal da Paraíba, João Pessoa, 2023.

Resumo

Neste trabalho, apresentamos uma breve introdução à matemática que fundamenta um dos algoritmos mais famosos e inovadores do mundo: o *PageRank* do *Google*. Esse algoritmo é capaz de calcular a relevância de cada página da Web, baseando-se em vários fatores, entre eles, a pontuação de importância que ele mesmo atribui. Nosso objetivo é explicar como o *PageRank* utiliza conceitos de Álgebra Linear e Teoria da Probabilidade para realizar esse cálculo, e também fornece uma visão intuitiva do seu funcionamento. Para isso, realizamos uma pesquisa bibliográfica, de caráter qualitativo e exploratório, que nos permitiu compreender melhor a lógica e a eficiência do algoritmo.

Palavras chaves: Cadeias de Markov; Ponto fixo de Banach; *PageRank*; Teoria da Probabilidade; Álgebra Linear; *Google*.

ROCHA, Diogo. **The Mathematics behind Google search engine: a brief introduction**. Federal University of Paraíba, João Pessoa, 2023

Abstract

In this work, we present a brief introduction to the mathematics that underlies one of the most famous and innovative algorithms in the world: Google's PageRank. This algorithm is able to calculate the relevance of each web page, based on several factors, including the importance score that it assigns. Our goal is to explain how PageRank uses concepts from Linear Algebra and Probability Theory to perform this calculation, and also provides an intuitive view of its operation. For this, we carried out a bibliographical research, of a qualitative and exploratory nature, which allowed us to better understand the logic and efficiency of the algorithm.

Keywords: Markov chains; Banach fixed point PageRank; Probability Theory; Linear algebra Google.

Lista de Tabelas

Lista de Figuras

Sumário

Introdução

Embora a Internet tenha sido criada originalmente para fins militares e acadêmicos, foi a invenção da *World Wide Web* que a tornou acessível e popular para o público em geral. A Web permitiu que as pessoas compartilhassem informações de forma fácil e rápida, usando uma interface gráfica e um sistema de links.

No entanto, à medida que a quantidade de informações na Web aumentava, também aumentava a dificuldade de encontrá-las. Foi então que surgiram as ferramentas de busca, que usavam algoritmos para indexar e classificar as páginas da Web de acordo com palavras-chave. Os primeiros buscadores eram baseados em diretórios, que organizavam as páginas em categorias hierárquicas, mas logo se mostraram insuficientes para atender à demanda dos usuários.

Por isso, foram desenvolvidos buscadores mais sofisticados, que usavam critérios como relevância, popularidade e autoridade para ranquear os resultados das buscas. Esses buscadores se tornaram essenciais para navegar na Web e encontrar as informações desejadas.

Tendo em vista tudo isso, no primeiro capítulo abordaremos o desenvolvimento matemático do *Pagerank* através dos conceitos de cadeia de Markov, buscando enfatizar o conceito hierárquico das páginas. No segundo capítulo usaremos o teorema do ponto fixo de Banach para reiterar a matemática por traz do funcionamento do buscador do *Google*. Já no terceiro capítulo explanaremos o *Pagerank*, sua história e analisaremos na prática seus conceitos. E por fim no último capítulo vamos explorar o conceito matemático no ensino básico, propondo atividades para a melhor compreensão nos anos finais do ensino Básico.

1 Cadeias de Markov

Estudaremos neste capítulo a definição e resultados de Cadeias de Markov. Na sequência, estes conceitos serão aprofundados. O texto teve como base [14], [10] e [5], onde podem ser extraídas mais informações.

1.1 Breve histórico sobre Andrei Andreyevich Markov

Andrei Andreyevich Markov foi um matemático russo que nasceu em 14 de junho de 1856 em Ryazan, Russia. Ele se formou na Universidade de São Petersburgo em 1878 e se tornou professor na mesma universidade em 1886. Ele é famoso por seus trabalhos na teoria dos números e na teoria das probabilidades, especialmente nos processos estocásticos chamados de cadeias de Markov.

Markov começou sua carreira estudando frações contínuas, limites de integrais, teoria da aproximação e convergência de séries. Ele foi influenciado por seu professor Pafnuty Chebyshev, que desenvolveu o método das frações contínuas e o aplicou à teoria das probabilidades. Markov também provou o teorema do limite central generalizado usando o método de Chebyshev.

Em 1900, Markov se interessou pelos processos estocásticos, que são sequências de eventos aleatórios que dependem uns dos outros. Ele introduziu o conceito de cadeia de Markov, que é um tipo especial de processo estocástico em que a probabilidade do próximo evento depende apenas do evento atual e não dos eventos anteriores. Ele usou as cadeias de Markov para estudar a distribuição das letras em um texto e a probabilidade de uma consoante ou uma vogal aparecer em uma determinada posição.

As cadeias de Markov têm muitas aplicações em diversas áreas da ciência, como física atômica, teoria quântica, biologia, genética, comportamento social, economia e finanças. Elas também são usadas para modelar sistemas dinâmicos, sistemas de informação, algoritmos e inteligência artificial.

Markov foi um membro da Academia Russa de Ciências desde 1896 e recebeu vários prêmios por suas contribuições à matemática. Ele morreu em 20 de julho de 1922 em Petrogrado (atual São Petersburgo), na Rússia.

1.2 Conceito de Cadeias de Markov

Um processo de Markov é um processo aleatório do qual o futuro (o próximo passo) depende apenas no estado atual; não tem memória de como o estado atual foi alcançado. Exemplo 1.2.1: Por exemplo, considere a seguinte situação, onde a variável aleatória X_t representa o estado de uma máquina no tempo t, em minutos.

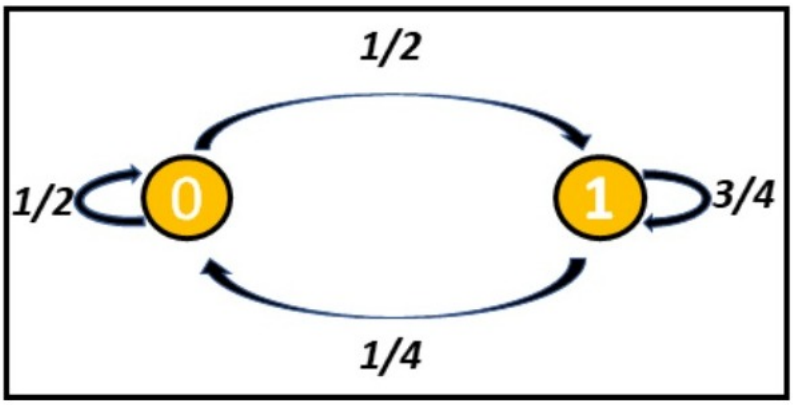
$$X_t = \begin{cases} 1, \text{se a máquina estiver ligada} \\ 0, \text{se a máquina estiver desligada} \end{cases} , com\ t \in \mathbb{N}.$$

Perceba que existe um espaço de estado, no qual é o conjunto E de valores que a variável X_t pede assumir, esse espaço nesse caso é E $=\{0, 1\}$.

Seja X_t a variável aleatória que indica o estado da máquina no minuto t, onde $X_t = 0$ significa que a máquina está desligada e $X_t = 1$ significa que a máquina está ligada. Assim, $X_1,\ X_2,\ X_3,\ ...$ são variáveis aleatórias que representam o estado da máquina nos minutos $1, 2, 3, ...$ respectivamente. O conjunto $\{X_t, t \in \mathbb{N}\}$ é um processo estocástico com espaço de estados discreto (E $= \{0,\ 1\}$) e tempo discreto ($T = \mathbb{N}$). Esse processo estocástico pode ser usado para modelar o comportamento da máquina ao longo do tempo.

Buscando um melhor entendimento deste exemplo podemos representá-lo graficamente através de um diagrama na figura abaixo.

Figura 1: Representação de um processo estocástico



fonte: Autoria própria, 2023

Observe que gráfico demostra uma situação, quando a máquina estiver no estado 0 a probabilidade de permanecer no estado 0 é de 1/2 e de ir para o estado 1 é 1/2, nesse caso a probabilidade de a máquina ligar ou não é a mesma, quando a máquina estiver no estado 1 a probabilidade de permanecer no estado 1 é de 3/4 e de ir para o estado 0 é 1/4.

Perceba que, se a máquina estiver no estado 1 (ligada), então a probabilidade de transição do estado 1 para o estado 0 depende apenas do estado atual (estado 1), independendo do estado em que a máquina estava anteriormente. Quando a ocorrência de um estado futuro depender apenas do estado atual, isto é, a probabilidade dos eventos futuros não dependem dos eventos passados, vindo a depender apenas do estado atual, este é um modelo simples do processo estocástico de Markov (ou markoviano)

$$P(X_t = x_t \mid X_{t-1} = x_{t-1}, \ldots, X_1 = x_1, X_0 = x_0) = P(X_t = x_t \mid X_{t-1} = x_{t-1})$$

A probabilidade de estar no estado x_t no instante t depende somente do estado no instante $t-1$, independendo de todos os estados anteriores $(X_{t-2}, X_{t-3}, \ldots, X_1, X_0)$. Essa propriedade chama-se *propriedade de Markov.* E qualquer processo estocástico em tempo discreto e estado discreto com a propriedade de Markov é definido como *cadeia de Markov.*

Definição 1. *(Matriz Estocástica). Uma matriz quadrada M, de ordem $n \times n$ e $M_{ij} \geq 0$, é chamada de estocástica se, para toda coluna j, $\sum_i M_{ij} = 1$.*

Uma matriz estocástica descreve uma cadeia de Markov X_t sobre um espaço de probabilidades finito "S". Dito de outra maneira, isso significa que as probabilidades de transição que caracterizam um processo de Markov são normalmente agrupadas na matriz de Markov.

Para se mover do estado i para o estado j, ou seja, a probabilidade de se mover de período para o período seguinte é $Pr(j \mid i) = p_{ij}$ então Matriz estocástica "P"é dada por:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1j} & \cdots \\ p_{21} & p_{22} & \cdots & p_{2j} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots \\ p_{i1} & p_{i2} & \cdots & p_{ij} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots \end{bmatrix}$$

Tomando o exemplo das máquina citada anteriormente, suponha um processo estocástico com apenas dois estados (digamos, que a máquina pode esta "ligada"ou "desligada "). Então, a matriz de transição de Markov teria dimensão 2x2:

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}$$

Neste caso a distribuição inicial é dado por um vetor linha. Dada uma distribuição inicial $\boldsymbol{\tau}_0$ e uma matriz de transição P, podemos descobrir diretamente a distribuição do período seguinte. O sistema de equações implícito na multiplicação de vetores abaixo é chamado de sistema de Markov ou processo de Markov:

Se $\boldsymbol{\tau}_0$ for um vetor coluna,$\boldsymbol{\tau}_1 = P \cdot \boldsymbol{\tau}_0$

Se $\boldsymbol{\tau}_0$ for um vetor linha, $\boldsymbol{\tau}_1^T = \boldsymbol{\tau}_0^T \cdot P$, onde "T" indica uma matriz transposta.

A probabilidade de se mudar do estado i para o estado j em duas etapas é dada pelo $(i,j)^0$ elemento da matriz P elevada ao quadrado:

$$(P^2)_{ij} = P^2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} = \begin{bmatrix} \frac{3}{8} & \frac{5}{8} \\ \frac{5}{16} & \frac{11}{16} \end{bmatrix}.$$

Generalizando, a probabilidade de se mudar do estado i para o estado j em "k" etapas é dada pelo $(i,j)^0$ elemento da matriz P elevada a k: $(P^k)_{ij}$. Continuamos alguns cálculos no caso da máquina que pode esta "ligada" ou "desligada":

$$(P^3)_{ij} = P^3 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} \cdot \begin{bmatrix} \frac{3}{8} & \frac{5}{8} \\ \frac{5}{16} & \frac{11}{16} \end{bmatrix} = \begin{bmatrix} \frac{11}{32} & \frac{21}{32} \\ \frac{21}{64} & \frac{43}{64} \end{bmatrix},$$

$$(P^4)_{ij} = P^4 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} \cdot \begin{bmatrix} \frac{11}{32} & \frac{21}{32} \\ \frac{21}{64} & \frac{43}{64} \end{bmatrix} = \begin{bmatrix} \frac{43}{128} & \frac{85}{128} \\ \frac{85}{256} & \frac{171}{256} \end{bmatrix},$$

$$(P^5)_{ij} = P^5 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} \cdot \begin{bmatrix} \frac{43}{128} & \frac{85}{128} \\ \frac{85}{256} & \frac{171}{256} \end{bmatrix} = \begin{bmatrix} \frac{171}{512} & \frac{341}{512} \\ \frac{341}{1024} & \frac{683}{1024} \end{bmatrix}, \cdots, (P^k)_{ij}$$

Ou seja, o estado futuro, para n grande, é independente do estado inicial. Neste caso, temos uma distribuição estacionária dos estados.

Definição 2. *(Probabilidades de transição) As probabilidades condicionais*

$$P(X_{t+1} = j \mid X_t = i), \text{ para } t = 0, 1, 2, \cdots,$$

são chamadas de probabilidades de transição. Se, para cada i e j ,

$$P(X_{t+1} = j \mid X_t = i) = P(X_1 = j \mid X_0 = i), \text{ para todo } t = 0, 1, 2, \cdots,$$

então as probabilidades de transição para um passo são ditas estacionárias e usualmente são denotadas por P_{ij}.

A transição do estado i ao estado j, em um passo simbolizado por P_{ij}, é a probabilidade de após um intervalo de tempo fixo predeterminado e de um objeto que se encontra no estado i ser encontrado no estado j.

Para n passos à frente, como extensão da Definição 2, é possível escrever as probabilidades de transição para cada i e j , com $n = 0, 1, 2, \cdots$, conforme a expressão:

$$P(X_{t+1} = j \mid X_t = i) = P(X_1 = j \mid X_0 = i), \ para\ todo\ t = 0, 1, 2, \cdots ,$$

Usaremos a seguinte simbologia para simplificar a notação:

$$P(X_1 = j \mid X_0 = i)P_{ij},$$

$$P(X_n = j \mid X_0 = i)P_{ij}(n).$$

De acordo com a referência (HILLIER e LIEBERMAN, 1995), a notação $P_{ij}(n)$ introduzida anteriormente implica que, para $n = 0$, $P_{ij}(0)$ é $P(X_0 = j \mid X_0 = i)$, sendo igual a 1 se $i = j$ e igual a 0 em caso contrário.

As probabilidades de estados são definidas como a seguir.

Definição 3. *: (Probabilidade de estado no instante n) A probabilidade do estado i tomada no instante n é a probabilidade de um objeto ocupar o estado i após um número n finito de passos.*

$$p_i(n) = P(X_n = i), \ para\ i = 0, 1, 2, \cdots , M.$$

De posse das definições estabelecidas nesta seção, com exemplos já citados, vamos apresentar os números.

Exemplo 1.2.1: Representa o estado de uma máquina, 1, se a máquina estiver ligada e 0, se a máquina estiver desligada. A partir de observações históricas, foram obtidas para um passo as probabilidades de transição supostas constantes. A máquina estar desligada e continuar desligada com probabilidade igual a 3/4 e, de a máquina estar desligada e ser ligada é com probabilidade 1/2. De a máquina estar ligada e ser desligada a probabilidade é de 1/4 e, de estar ligado e permanecer ligado a probabilidade é de 1/2.

$$-desligado- \ permanecer- \ desligado- > P(X_1 = 0 \mid X_0 = 0) = P_{00} = 1/2;$$

$$-desligado- \ sucede- \ ligamento- > P(X_1 = 1 \mid X_0 = 0) = P_{01} = 1/2;$$

$$-ligado- \ permanecer- \ ligado- > P(X_1 = 1 \mid X_0 = 1) = P_{11} = 3/4;$$

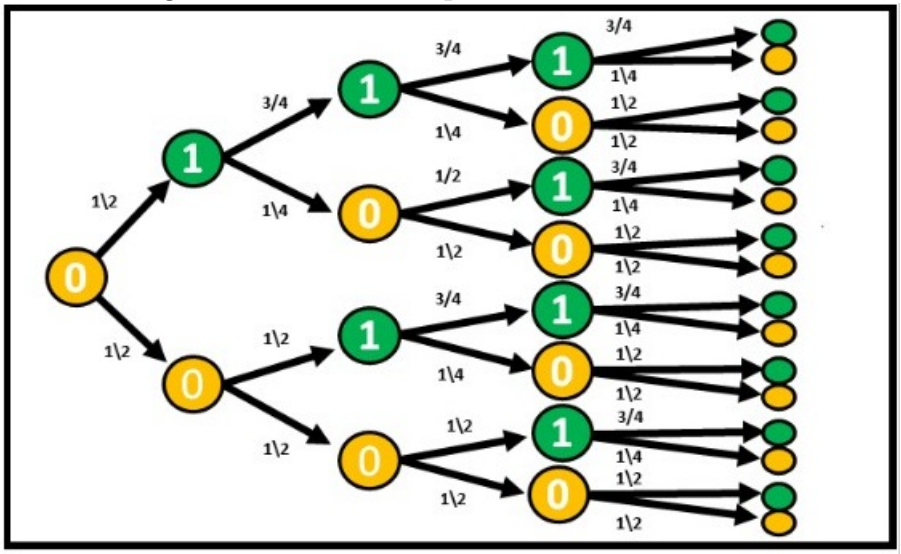
$$-ligado- \ sucede- \ desligadamento- > P(X_1 = 1 \mid X_0 = 1) = P_{10} = 1/4.$$

Para resolver o exemplo lançamos mão de um artifício conhecido por árvore de probabilidades. A Figura 2 ilustra o diagrama em árvore partindo do estado 0, onde são considerados apenas quatro passos. Outro diagrama poderia ser construído, porém, partiria do estado 1. Mostraremos como se calcula a probabilidade do estado 1 após quatro

passos, isto é, p_1.

Notamos na árvore de probabilidades (Figura 2) que, dado que os eventos são independentes, precisamos multiplicar todas as probabilidades dos 'caminhos' que levam ao estado 1 para calcular a probabilidade do estado 1 em quatro passos (BILLINTON, 1970).

Figura 2: : Diagrama em árvore de probabilidades iniciando no estado 0



fonte: Autoria própria, 2023

Se estendermos os cálculos para mais passos não é difícil concluir que a probabilidade do estado 1 encaminhará para 2/3 , enquanto que a probabilidade do estado 0 será de 1/3, preservadas as condições.

Definição 4. *Um estado i se diz recorrente se $f_i = 1$. Um estado i se chama transitório se $f_i < 1$*

Notamos que se o estado i é recorrente, então a cadeia de Markov saindo do estado i, sempre vai voltar para este estado, ou o número de vezes que a cadeia está no estado i é infinito. Consequentemente, a média do número de vezes em que a cadeia está em tal estado é infinito. Por outro lado, se i é transitório, então a probabilidade de que, saindo deste estado i, a cadeia nunca mais volte neste estado é positiva; a probabilidade é igual a $1 - f_i$. Não é difícil ver que o número de vezes que a cadeia volta para tal estado tem distribuição geométrica:

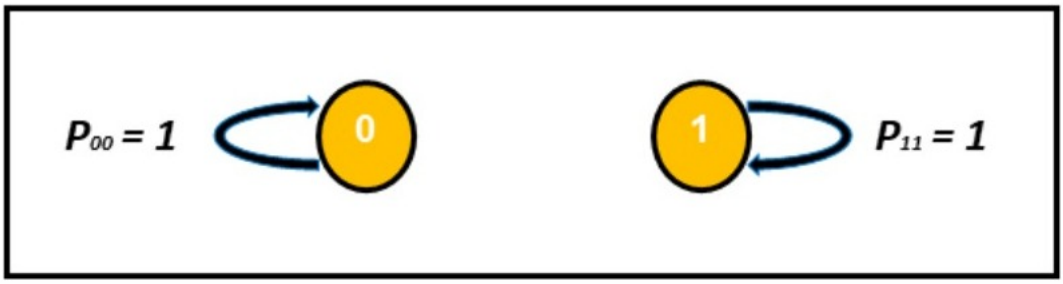
$$P\{X_n \textit{ exatamente } k \textit{ vezes volta em estado } i \mid X_0 = i\} = f_i^k(1 - f_i), k = 0, 1, \cdots$$

Estado recorrente; Um estado i é recorrente se e somente se, partindo do estado i , o processo eventualmente retornará ao estado i com probabilidade $f_{ii} = 1$.

O processo cuja cadeia de Markov foi apresentada no Exemplo 1.2.1 possui ambos os estados recorrentes. Outro exemplo de estados recorrentes corresponde à matriz de probabilidades de transição mostrada a seguir

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Figura 3: : Diagrama de transição de uma cadeia de Markov com dois estados recorrentes.

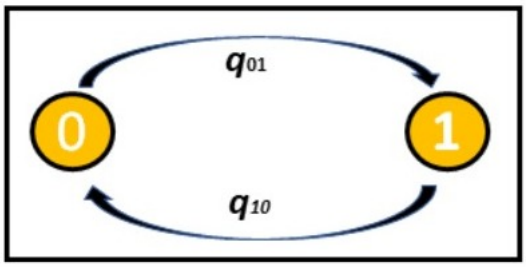


fonte: Autoria própria, 2023

Um estado i é transitório se existir um estado j que seja acessível a partir de i, mas o estado i não seja acessível a partir do estado j.

O processo cuja cadeia de Markov foi apresentada no Exemplo 1.2.1 possui ambos os estados transitórios.

Figura 4: Diagrama de transição de uma cadeia de Markov com dois estados transitórios.



fonte: Autoria própria, 2023

A média do número de vezes em que a cadeia está no estado i é finita e igual a $1/(1 - f_i)$.

Seja I_n a função indicadora de que a cadeia está no estado i :

$$I_n = \begin{cases} 1, se\ X_n = i \\ 0, se\ X_n \not= i \end{cases}$$

Agora, o número de vezes em que a cadeia está no estado i pode ser representado através da soma $\sum_{n=0}^{\infty} I_n$. Logo, a esperança pode ser representada da seguinte forma:

$$E\left[\sum_{n=0}^{\infty} I_n \mid X_0 = i\right] = \sum_{n=0}^{\infty} P_{ii}^{(n)}.$$

Notemos que o estado i é recorrente se $\sum_{n=1}^{\infty} P_{ii}^{(n)} = \infty$, transitório se $\sum_{n=1}^{\infty} P_{ii}^{(n)} < \infty$.

Notamos que se uma cadeia é finita (o número de estados dela é finito), então, não podem ser todos os estados transitórios.

Se o estado i é recorrente, e o estado i se comunica com o estado j, então o estado j também é recorrente.

Explanamos algumas definições da cadeia de Markov, enfatizamos as de mais relevância para o tema abordado, vale salientar que essas definições estão entre outras que caracterizam uma cadeia de Markov. Podem ser extraídas mais informações nas referencias [4], [11] e [12]. Com a base da Cadeia de Markov, iremos aborda posteriormente um tema contextualizado para o melhor entendimento do tema abordado.

1.3 O RÃ DÕ e o algoritmo Pagehank

No ano de 2009 na XXXI Olimpíada Brasileira de Matemática, desenvolveu um problema descrito da seguinte maneira; Sejam A, B, e C os vértices do triângulo no sentido anti-horário. Seja Q_n a probabilidade de, após n saltos, Dõ estar no vértice B. Temos $P_0 = 1, Q_0 = R_0 = 0$ e, para todo $n \geq 0$,

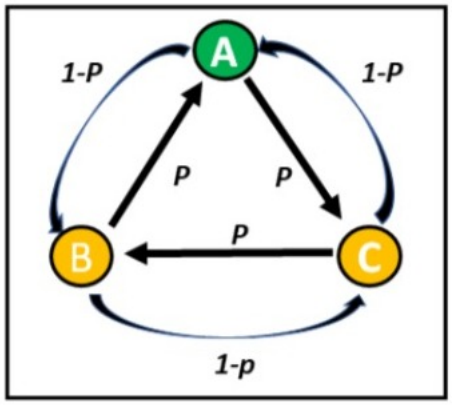
$$(*) \begin{cases} P_{n+1} = (1-P)R_n + pQ_n \\ Q_{n+1} = (1-P)P_n + pR_n \\ R_{n+1} = (1-p)Q_n + pP_n. \end{cases}$$

Em primeira vista, embora não percebemos, esse problema tem relação com o algoritmo *PageRank*.

Para ilustrar melhor o problema e a solução de uma forma mais didática para o leitor, podemos incluir um diagrama do triângulo equilátero com os vértices A, B *e* C e as probabilidades de salto da Rã Dõ. Podemos mostrar que a rã pode saltar de um vértice para outro com as probabilidades indicadas nas arestas do triângulo. Por exemplo, se a rã está no vértice A, ela pode saltar para o vértice B com probabilidade p ou para o vértice C com probabilidade $1-p$. O diagrama também mostra que a rã nunca pode saltar para o mesmo vértice em que está. Vejamos na figura 1.

Quando estamos na posição inicial da Rã Dõ o valor de $n = 0$, isso que dizer que a probabilidade de Rã estar no vértice A é de 100%, que é fornecida no problema. Donde, $a_0 = 1$ e $b_0 = c_0 = 0$. Isto é, para $n = 0$, podemos observar que a probabilidade de estar

Figura 5: Representação do problema da rã Dõ



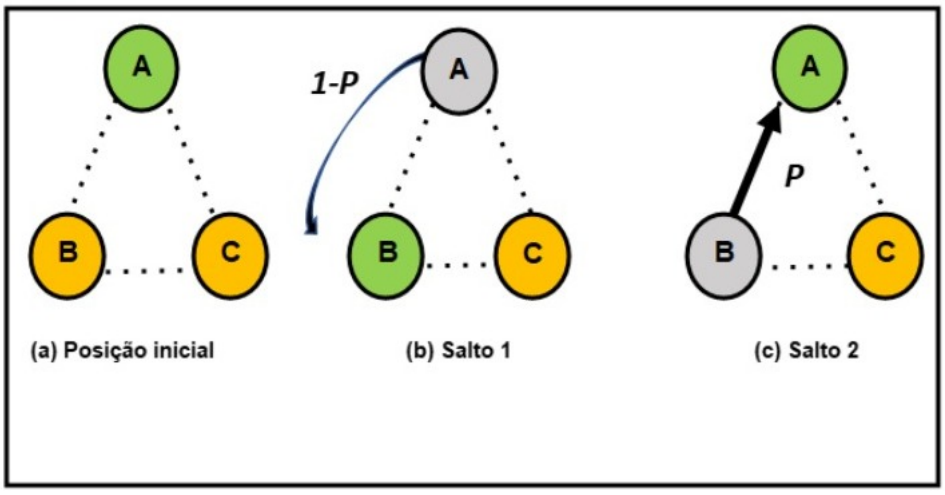
fonte: Autoria própria, 2023

no vértice B é de 0% e de estar no vértice C também é de 0%. Quando $n = 1$, temos $a_1 = 0, b_1 = 1 - p$ e $c_1 = p$. Quando $n = 2$, devemos fazer uma análise mais cautelosa.

Podemos mostrar duas probabilidades para os saltos de RÃ DÕ, a primeira de salto de A para B, e também de B para A como ilustra figura 6 abaixo. Se P é a probabilidade de Dõ seguir esse caminho, então $P = (1 - p)p$. Note que $b_1 = 1 - p$, assim

$$P = b_1.p.$$

Figura 6: Primeira alternativa de saltos até o vértice A



fonte: Autoria própria, 2023

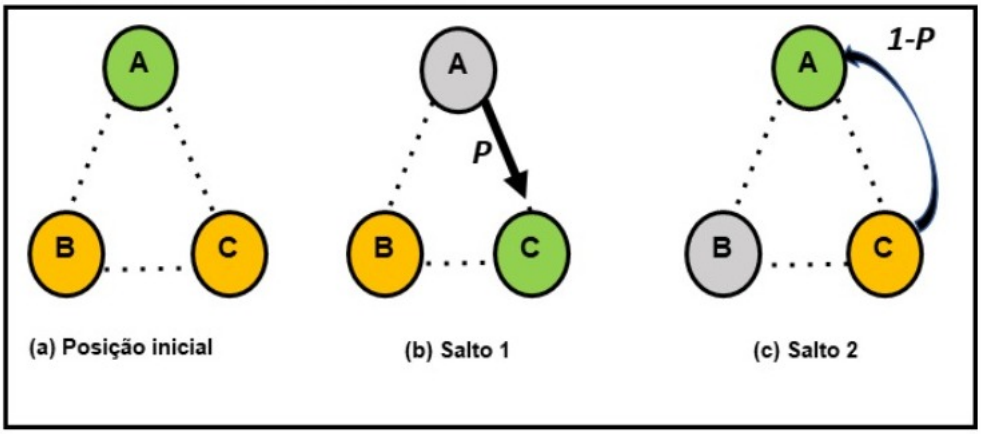
Podemos da mesma forma observar a segunda alternativa, na qual o primeiro salto de A para C e posteriormente de C para A.

Novamente, se Q é a probabilidade de Dõ seguir esse caminho até o vértice A, então $Q = p(1 - p)$.

Ou ainda, como $c_1 = p$, $Q = c_1(1 - p)$. como ilustra a figura 7 abaixo;

Sabendo que o objetivo é determinar as probabilidades futuras de a_2, b_2 e c_2 conhecendo

Figura 7: Segunda alternativa de saltos até o vértice A



fonte: Autoria própria, 2023

o vértice em que RÃ se encontra no atual momento, isto é, posterior o salto 1.

Portanto,

$$a_2 = b_1.p + c_1.(1-p)$$

$$b_2 = a_1.(1-p) + c_1.p$$

$$c_2 = a_1.p + b_1.(1-p).$$

Vamos calcular os passos a_3, b_3 e c_3

• Se o estado atual da RÃ após o segundo salto é o vértice B, cuja probabilidade é b_2 , então, para se chegar ao vértice A, o terceiro salto deve ser no sentido horário (p).

• Se o estado atual da rã após o segundo salto é o vértice C, cuja probabilidade é c_2, então, para se chegar ao vértice A, o terceiro salto deve ser no sentido anti-horário $(1-p)$.

Temos

$$a_3 = b_2.p + c_2.(1-p)$$

$$b_3 = a_2.(1-p) + c_2.p$$

$$c_3 = a_2.p + b_2.(1-p)$$

As probabilidades futuras a_n, b_n e c_n dependem apenas do vértice em que se encontra no salto imediatamente anterior, $n-1$. As seguintes generalizações nos permitem escrever de maneira analogamente;

$$a_n = p \cdot b_{n-1} + (1-p) \cdot c_{n-1}$$

$$b_n = p \cdot c_{n-1} + (1-p) \cdot a_{n-1}$$

$$c_n = p \cdot a_{n-1} + (1-p) \cdot b_{n-1}$$

Dessa forma escrevemos n probabilidades futuras de RÃ DÕ, através desta genera-

lização.

No problema da RÃ, queríamos mostrar que, para $n \to \infty$, a probabilidade dela estar no vértice A seria de $1/3$. Mas como chegamos a esse valor? Usamos um método chamado de cadeias de Markov, que consiste em modelar o comportamento da RÃ como uma sequência de estados aleatórios. Cada estado representa um vértice do diagrama da Figura 5, e a transição entre os estados depende de uma matriz de probabilidades. Essa matriz nos diz qual a chance do RÃ saltar de um vértice para outro a cada instante. Ao aplicarmos esse método, encontramos um vetor de estado estacionário, que é o limite do vetor de estado quando n tende ao infinito. Esse vetor nos fornece a probabilidade do RÃ estar em cada vértice no longo prazo, e é assim que obtemos o valor de $1/3$ para o vértice A.

Esse mesmo método pode ser usado para resolver outros problemas envolvendo probabilidades no longo prazo, como o algoritmo *PageRank*, que mede a importância das páginas da Web. Nesse caso, os estados representam as páginas, e as transições dependem dos *links* entre elas. O vetor de estado estacionário nos fornece a probabilidade de um usuário visitar cada página no longo prazo, e isso é usado para ranquear as páginas nos mecanismos de busca. Assim, podemos ver que o problema da rã é mais do que um simples exercício de matemática, mas uma ferramenta poderosa para modelar fenômenos aleatórios.

2 Ponto fixo de Banach

Estudaremos neste capítulo a definição e resultados de Ponto fixo de Banach. Na sequência esses conceitos serão aprofundados e o texto teve como base [1] e [6], onde podem ser extraídas mais informações

2.1 Uma breve história sobre Stefan Banach

Stefan Banach foi um dos maiores matemáticos do século XX e o fundador da moderna análise funcional. Ele criou uma abordagem geral e abstrata para o estudo de espaços de funções, que hoje são chamados de espaços de Banach em sua homenagem. Ele também provou teoremas importantes sobre séries de Fourier, equações integrais e diferenciais, topologia e teoria da medida.

Banach nasceu em 30 de março de 1892, em Cracóvia, na Polônia. Ele foi criado por sua mãe e nunca conheceu seu pai. Ele mostrou talento para a matemática desde cedo, mas não teve uma educação formal nessa área. Em 1910, ele se mudou para Lviv (atualmente na Ucrânia) e se matriculou na faculdade de engenharia da Universidade Técnica local. Ele teve que trabalhar como tutor para se sustentar, o que lhe deixava pouco tempo para estudar. Ele se formou em 1914, mas logo depois a Primeira Guerra Mundial eclodiu e ele teve que fugir de Lviv.

2.2 Ponto fixo de Banach

Da mesma forma em que relatamos a contribuição de Makov para o desenvolvimento do *Google*. Neste capítulo iremos também prestigiar a contribuição de Stefan Banach com o desenvolvimento do ponto fixo de Banach

É nítido que a matemática de Banach e Marcov deram frutos para Page e Brin lançarem o algoritmo *PageRank*, essa matemática foi o grande diferencial para que o *Google* tenha seu sucesso notório.

Um teorema de ponto fixo é um resultado matemático que garante a existência e unicidade de um elemento x que satisfaz a condição $f(x) = x$, onde f é uma aplicação de um espaço métrico em si mesmo. Um exemplo importante desse tipo de teorema é o Teorema do Ponto Fixo de Banach, que afirma que se f é uma contração uniforme, ou seja, se existe uma constante $0 \leq \beta < 1$ tal que $d(f(x), f(y)) \leq \beta d(x, y)$ para todos x e y no espaço métrico, então f tem um único ponto fixo.

O Teorema do Ponto Fixo de Banach tem diversas aplicações em análise matemática, especialmente na resolução de equações não lineares e equações diferenciais ordinárias. Além disso, o teorema fornece um método iterativo para aproximar o ponto fixo de uma contração uniforme, bastando escolher um ponto arbitrário X_0 e definir $X_n + 1 = f(X_n)$ para todo $n \geq 1$.

Neste capítulo, vamos apresentar alguns conceitos básicos sobre espaços métricos e demonstrar o Teorema do Ponto Fixo de Banach. Em seguida, vamos mostrar algumas aplicações desse teorema.

2.2.1 Espaços Métricos.

Definição 5. *Um espaço métrico é um par (M, d), onde M é um conjunto não vazio e d é uma função que associa a cada par de elementos de M um número real não negativo, chamado de distância entre eles. A função d deve satisfazer as seguintes propriedades para quaisquer x, y e z em M :*

i) $d(x, y) = 0$ *se e somente se* $x = y$ *(identidade dos indiscerníveis);*

ii) $d(x, y) = d(y, x)$ *(simetria);*

iii) $d(x, z) \leq d(x, y) + d(y, z)$ *(desigualdade triangular).*

Os elementos de M são chamados de *pontos*. A função d é chamada de *métrica* ou *função distância* em M.

Um exemplo de espaço métrico é, sejam M um conjunto qualquer não vazio e consideremos a função $d : M$ x $M \to \mathbb{R}$ definido por

$$d(x, y) = \begin{cases} 0, se\ x = y \\ o, se\ x \neq y. \end{cases}$$

È imediata a verificação de que d é uma métrica, essa é chamada *métrica zero-um.*

Um outro exemplo é que considere $M = \mathbb{R}$ o conjunto dos números reais e a função $d : M$ x $M \to \mathbb{R}$ definida por

$$d(x, y) =\mid x - y \mid .$$

Usando as propriedades de módulo de um número real, segue que d é uma métrica. De fato, sejam $x, y\ e\ z \in \mathbb{R}$ quaisquer. se $x = y$, é imediato que $\mid x - y \mid = 0$, o que implica em $d(x - y) = 0$. Além disso, se $x \neq y$, temos $x - y \neq oqueimplicaem \mid x - y \mid > 0$, ou seja, que $d(s, y) > 0$. também, temos que

$$d(x, y) =\mid x - y \mid =\mid -(x - y) \mid =\mid y - x \mid = d(x, y).$$

Por fim, temos que

$$\mid x - y \mid =\mid -(x - y) + (y - z) \mid \leq \mid x - y \mid + \mid y - z \mid$$

Logo, $d(x,z) \leq d(x,y) + d(y,z)$. Portanto, d é um métrica a qual é chamada de métrica usual de $\mathbb{R}$

2.2.2 Convergência em Espaços Métricos.

Definição 6. *Uma sequência (x_n) em um espaço métrico (M,d) chama-se limitada quando o conjunto de seus termos é limitado, isto é, quando existe $c > 0$ tal que*

$$d(x_m, x_n) \leq c,$$

para quaisquer $m, n \in \mathbb{N}$.

Um exemplo de sequência limitada é a sequência (x_n), definida por $x_n = (-1)^n$ para todo n natural. Nesse caso, podemos tomar $c = 2$, pois para quaisquer m e n naturais, temos

$$d(x_m, x_n) = |(-1)^m - (-1)^n| \leq |(-1)^m| + |(-1)^n| \leq 1 + 1 = 2.$$

Definição 7. *Uma sequência (x_n) em um espaço métrico (M,d) é dita convergente em M se existir $x \in M$ tal que*

$$\lim_{n\to\infty} d(x_n, x) = 0.$$

isto é, para todo ϵ, existe n_0 tal que $n > n_0$ implica $d(x_n, x), < \varepsilon$.

Quando for necessário, usaremos a notação $x_n \to x$ para indicar a convergência.

2.2.3 O Teorema do Ponto Fixo de Banach.

Definição 8. *Seja(M,d) um espaço métrico. Uma função $fM \to M$ de contração sobre M se existe um número real positivo $k < 1$ tal que:*

$$d(f(x), f(y)) \leq kd(x,y), \text{ para todo } x, y \in M.$$

Para exemplificar, Considere $M = \mathbb{R}$ com a métrica usual. A função $f : [1, +\infty) \to \mathbb{R}$ definida por $f(x) = \sqrt{x}$ é uma contração. De fato:

$$d(f(x), f(y)) = |\sqrt{x} - \sqrt{y}| = |\sqrt{x} - \sqrt{y}| \cdot \frac{|\sqrt{x} + \sqrt{y}|}{|\sqrt{x} + \sqrt{y}|} = \frac{1}{|\sqrt{x} + \sqrt{y}|} \cdot |x - y|.$$

Com o $x, y \geq 1$ temos que $\sqrt{x} + \sqrt{y} \geq 2$, ou ainda, $\frac{1}{\sqrt{x}+\sqrt{y}} \leq \frac{1}{2}$.
Logo,

$$d(f(x), f(y)) \leq \frac{1}{2}|x - y|.$$

Observe que não é uma contração quando definida no intervalo fechado $[0,1]$, pois $\lim_{x\to 0^+} f'(x) = +\infty$.

Teorema 1. *(Teorema do Ponto Fixo de Banach) Considere $(M.d)$ um espaço métrico completo e uma contração $f : M \to M$. Então f possui um único ponto fixo.*

Prova 1. *Considere $x_0 \in M$ e a sequência (X_n) em M. definida por $x_{n+1} = f(x_n)$*
Então,

$$d(x_1, x_2) = d(f(x_0), f(x_1)) \leq kd(x_0, x_1) \Rightarrow d(x_1, x_2) \leq kd(x_0, x_1)$$

$$d(x_2, x_3) = d(f(x_1), f(x_2)) \leq kd(x_1, x_2) \leq k^2 d(x_0, x_1) \Rightarrow d(x_2, x_3) \leq k^2 d(x_0, x_1).$$

Continuando o processo, usando um argumento indutivo, chegamos à conclusão que $d(x_n, x_{n+1}) \leq k^n d(x_0, x_1)$. Como o nosso interesse é mostrar que (x_n) é uma sequência de Cauchy, da desigualdade triangular, temos:

$$d(x_n, x_{n+p}) \leq d(x_n, x_{n+1}) + d(x_{n+1}, x_{n+2}) + \cdots + d(x_{n+p-1}, x_{n+p}).$$

Pooutro lado,

$$d(x_n, x_{n+1}) \leq k^n d(x_0, x_1),$$

$$d(x_{n+1}, x_{n+2}) \leq k^{n+1} d(x_0, x_1),$$

$$\vdots$$

$$d(x_{n+p-1}, x_{n+p}) \leq k^{n+p-1} d(x_0, x_1).$$

Assim, temos:

$$d(x_n, x_{n+1}) + d(x_{n+1}, x_{n+2}) + \cdots + d(x_{n+p-1}, x_{n+p}) \leq (k^n + k^{n+1} + \cdots + k^{n+p-1}) d(x_0, x_1)$$

Sabendo que $k < 1$ para qualquer p, fixado, temos

$$k^n + k^{n+1} + \cdots + k^{n+p-1} - k^n \cdot \frac{1 - k^p}{1 - k} \leq \frac{k^n}{1 - k},$$

logo, obtemos:

$$d(x_n, x_{n+p}) \leq \frac{k^n}{1 - k} \cdot d(x_0, x_1).$$

Tomando o limite quando $n \to \infty$ chegamos a:

$$\lim_{n\to\infty} d(x_n, x_{n+p}) \leq \lim_{n\to\infty} (\frac{k^n}{1 - k} d(x_0, x_1))$$

$$\lim_{n\to\infty} d(x_n, x_{n+p}) \leq d(x_0, x_1) \lim_{n\to\infty} \frac{k^n}{1 - k}$$

Como $0 < k < 1, k^n \to 0$, vale

$$\lim_{n\to\infty} \frac{k^n}{1-k} = 0.$$

ou seja,

$$\lim_{n\to\infty} d(x_n, x_{n+p}) = 0$$

donde concluímos que (x_n) *é de fato uma sequência de Cauchy em* M.

Ora como (M, d) *é um Espaço Métrico Completo então* (x_n) *converge em* M.

Assim tomando o limite na equação $x_{n+1} = f(x_n)$ *teremos:*

$$\lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} f(x_n)$$

Bem como $\lim_{n\to\infty} x_n = a$ *e como a aplicação* f *é contínua, obtemos que*

$$\lim_{n\to\infty} f(x_n) = f(lim_{n\to\infty} x_n) = f(a)$$

Assim, temos a igualdade desejada

$$f(a) = a$$

Assim, provamos a existência do ponto fixo. Agora provemos a unicidade. Sejam a *e* b *em* M *tais que* $f(a) = a$ *e* $f(b) = b$. *Assim,*

$$d(a, b) = d(f(a), f(b)) \leq kd(a, b)$$

Isso leva á desigualdade

$$(1-k)d(a, b) \leq 0$$

Como $k < 1$ *entas* $1 - k < 0$ *donde concluímos que* $d(a, b) \leq 0$. *Como* $d(a, b)$ *é um número real não negativo, segue que* $d(a, b) = 0$, *e isso só acorre se , e somente se* $a = b$. *Assim,* f *só possui um único ponto fixo, o que completa a demostração do Teorema.*

2.3 Funcionamento do Buscador do *Google*

O funcionamento de um buscador é baseado essencialmente em dois passos:

1) Matching Busca

2) Ranking Seleciona e Ordena

A forma de ordenar as páginas encontradas é o segredo do sucesso do *Google*.

Sabemos que as páginas se conectam através de links e podemos pensá-las como nós de um grafo direcionado cuja arestas são dadas pelos links. Seja G um grafo direcionado, conosco $1, 2, ..., n$. Nosso objetivo é, para cada nó i, atribuir um valor real X_i que traduza a relevância do nó i.

Quando j é um nó que aponta para i, chamaremos de *link* de i para j e denotaremos $j \to i$. O número de links saindo de j será denotado por l_j. Assim, definiremos a relevância de um nó ou site da seguinte forma:

Um site é relevante se for apontado por outros sites relevantes.

Essa definição é recursiva e intuitiva, pois significa que quanto mais *links* um site recebe, mais relevante ele é. Porém, nem todos os *links* têm o mesmo peso. Um *link* vindo de um site muito relevante vale mais do que um *link* vindo de um site pouco relevante. Portanto, a relevância de um site depende não só da quantidade, mas também da qualidade dos *links* que ele recebe.

Para calcular a relevância de cada site. O *PageRank* atribui um valor numérico entre 0 e 1 para cada site, sendo que quanto maior o valor, mais relevante é o site. O valor do *PageRank* de um site é determinado pela seguinte fórmula:

$$X_i = \sum_{j=i} X_j L_j$$

Desse modo, definiremos m_{ij} como o número de *links* de $j \to i$, que pode inclusive ser zero. Com isso, temos o sistema linear

$$\begin{cases} x_1 = \frac{m_{11}}{l_1}x_1 + \frac{m_{12}}{l_2}x_2 + \cdots + \frac{m_{1n}}{l_n}x_n \\ x_2 = \frac{m_{21}}{l_1}x_1 + \frac{m_{22}}{l_2}x_2 + \cdots + \frac{m_{1n}}{l_n}x_n \\ \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \\ x_n = \frac{m_{n1}}{l_n}x_1 + \frac{m_{n2}}{l_2}x_2 + \cdots + \frac{m_{nn}}{l_n}x_n \end{cases}$$

Uma transformação linear é uma função que preserva as operações de adição vetorial e multiplicação por escalar entre dois espaços vetoriais.

Se $A = (a_{ij})$, onde $a_{ij} = \frac{m_{ij}}{l_j}$, então A pode ser vista como uma transformação linear de $A : \mathbb{R}^n \to \mathbb{R}^n$, e $x = [x_1, x_2, ..., X_n]$ um autovetor de autovalor unitário. Isso significa que x é um ponto fixo de A, ou seja, $A_x = x$.

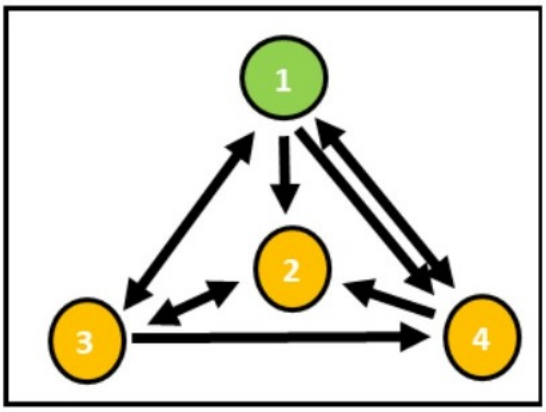
Por exemplo, se$l_1 = 4, l_2 = 1, l_3 = 3$ *e* $l_4 = 2$ observamos sistema linear:

$$\begin{cases} x_1 = \frac{0}{4}x_1 + \frac{0}{l}x_2 + \frac{1}{3}x_3 + \frac{1}{2}x_4 \\ x_2 = \frac{1}{4}x_1 + \frac{0}{1}x_2 + \frac{1}{3}x_3 + \frac{1}{2}x_4 \\ x_3 = \frac{1}{4}x_1 + \frac{1}{1}x_2 + \frac{0}{3}x_3 + \frac{0}{2}x_4 \\ x_4 = \frac{2}{4}x_1 + \frac{0}{l}x_2 + \frac{1}{3}x_3 + \frac{0}{2}x_4 \end{cases}$$

Uma possível solução desse sistema é dada pelo vetór [4, 5, 6, 4], que indicam, respectivamente, a relevância das páginas 1, 2, 3 e 4. Portanto, a página 3 é a que possui maior relevância.

Uma maneira de explicar o funcionamento do modelo de *PageRank* é usar o conceito

Figura 8: Grafo de relevância de pagina.



fonte: Autoria própria, 2023

de probabilidade. Podemos interpretar a_{ij} como a probabilidade de, ao sair da página j, chegar à página i. Por exemplo, se um usuário escolhe uma das n páginas aleatoriamente, digamos $v_0 = [1, 0, ..., 0]$, o vetor v_1, obtido na equação $\mathrm{A}v_0 = v_1$, indica a probabilidade do usuário estar na página i após um clique, partindo de v_0. Assim, sucessivamente, após n cliques a probabilidade do usuário estar na página i é dada pela equação $\mathrm{A}v_n - 1 = v_n$. Esse modelo seria ideal se o usuário sempre encontrasse um *link* de uma página para outra, mas isso não ocorre. A partir daí, a ideia de Page e Brin foi introduzir um fator probabilístico p de começar tudo de novo e, evidentemente, $1 - p$ de continuar nos *links*. Desse modo, a aplicação que indica o percurso aleatório do usuário em um grafo de n vértices é

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \mapsto p. \begin{bmatrix} \frac{1}{n} \\ \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{bmatrix} + (1-p) \begin{bmatrix} \frac{m_{11}}{l_1} & \frac{m_{12}}{l_2} & \cdots & \frac{m_{1n}}{l_n} \\ \frac{m_{21}}{l_1} & \frac{m_{22}}{l_2} & \cdots & \frac{m_{2n}}{l_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{m_{n1}}{l_n} & \frac{m_{n2}}{l_2} & \cdots & \frac{m_{nn}}{l_n} \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Podemos escrever, de forma mais simples, como $T : \mathbb{R}^n \to \mathbb{R}^n$, onde $y_7 \to$ p + (1 - p)A_y, com

$$e = \begin{bmatrix} \frac{1}{n} \\ \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{bmatrix}$$

Note que a função T é contínua. Se mostrarmos que T é uma contração, e lembrando que $\mathbb{R}^n$ é um espaço métrico completo, o Teorema do ponto fixo de Banach nos garante que T possui um único ponto fixo, ou seja, que o internauta chega sempre à página desejada. Para tanto, vejamos que, dados y, z $\in \mathbb{R}^n$ e, notando que temos que

$$\sum_{i=1}^{n} a_{ij} = 1$$

$$\| T(y) - T(z) \| = \| (1-p)A(y-z) \| = (1-p) \sum_{i=1}^{n} (\sum_{i=1}^{n} a_{ij} \mid y_j - z_j \mid) =$$

$$= (1-p) \sum_{i=1}^{n} \mid y_j - z_J \mid = (1-p) \| y - z \| .$$

Já que $1-p$ é menor que 1, temos que T é uma contração. Finalmente, por a aplicação possuir um único ponto fixo, o Teorema do ponto fixo de Banach garante que tudo isso funciona indicando que a relevância de cada página está bem definida.

3 Pagerank

Estudaremos neste capítulo um pouco sobre a historia e o funcionamento de *Pagerank*. Na sequência esses conceitos serão aprofundados e o texto teve como base [2], [7], [10], [11] e [14], onde podem ser extraídas mais informações

3.1 Breve histórico sobre *Pagerank*

PageRank é um algoritmo usado pelo *Google* para classificar as páginas da web em seus resultados de busca. Ele mede a importância de uma página levando em conta a quantidade e a qualidade dos *links* que apontam para ela. O nome vem do sobrenome de Larry Page, um dos fundadores do *Google*, e não do termo "página web".

O algoritmo foi desenvolvido na Universidade de Stanford por Larry Page e Sergey Brin em 1996, como parte de um projeto de pesquisa sobre um novo tipo de motor de busca. Eles tiveram a ideia de que a informação na web poderia ser ordenada em uma hierarquia de "popularidade de *links*": uma página é mais importante se tiver mais *links* apontando para ela. O trabalho contou com a colaboração de Rajeev Motwani e Terry Winograd. O primeiro artigo sobre o projeto, descrevendo o *PageRank* e o protótipo inicial do *Google*, foi publicado em 1998. Logo depois, Page e Brin fundaram a *Google* Inc., a empresa por trás do *Google*.

O *PageRank* foi inspirado na análise de citações, desenvolvida por Eugene Garfield em 1950 na Universidade da Pensilvânia, e pelo método *"Hyper Search"*, desenvolvido por Massimo Marchiori, da Universidade de Pádua. No mesmo ano em que o *PageRank* foi introduzido (1998), Jon Kleinberg publicou seu trabalho sobre HITS. Os fundadores do *Google* citaram Marchiori e Kleinberg em seu artigo original.

Um motor de busca chamado *"RankDex"* da *IDD Information Services*, desenhado por Robin Li, desde 1996, já explorava uma estratégia semelhante para pontuação e ranking de páginas. A tecnologia utilizada em *RankDex* foi patenteada em 1999 e usada mais tarde quando Li fundou a *Baidu* na China. O trabalho de Li está referenciado em algumas patentes do *Google*, de métodos de pesquisa de Larry Page.

3.2 Pagerank

Para ilustrar o funcionamento do *PageRank*, imagine uma rede de apenas quatro páginas: A, B, C e D. As ligações de uma página a si própria e as ligações múltiplas entre duas páginas são ignoradas. Inicialmente, a soma dos valores de *PageRank* de todas as páginas da web correspondia ao número de páginas na web. Em versões posteriores, o *PageRank* passou a assumir valores entre 0 e 1, representando uma distribuição probabilística, ou seja, a probabilidade de um usuário, percorrendo ligações aleatoriamente, chegue a uma determinada página.

No primeiro passo do processo de cálculo iterativo do *PageRank*, todas as páginas têm o mesmo valor de *PageRank.* No nosso exemplo de quatro páginas, o primeiro passo consiste em atribuir o valor 0,25 de *PageRank* a cada uma das quatro páginas. Note-se que a soma dos valores de *PageRank* de todas as páginas é 1.

Neste capítulo, vamos explorar como o algoritmo *PageRank* funciona para medir a relevância de cada página da Web. A ideia básica é que os *links* que apontam para uma página são como votos que indicam a sua qualidade e importância. Quanto mais *links* uma página recebe, maior é o seu *PageRank*.

Uma possível forma de se atribuir um valor numérico a cada página da Web é baseada na quantidade e na qualidade dos *links* que apontam para ela. Assim, uma página que recebe muitos *links* de outras páginas importantes teria um valor maior do que uma página que recebe poucos *links* de páginas pouco relevantes. Podemos tomar com exemplo, dados os números 0,35 e 0,4 sabemos que $0,4 > 0,35$. Portanto isso implica que uma página com valor 0,4 seria mais importante que a página de importância 0, 35. Esse valor numérico poderia ser usado para ordenar as páginas da Web de acordo com a sua importância relativa

A Web é formada por uma grande rede de páginas interligadas por *links.* Esses *links* permitem que os usuários naveguem entre as páginas e encontrem as informações que procuram. Mas como medir a importância de cada página na Web? Essa foi a questão que motivou Page e Brin a desenvolverem um algoritmo baseado na estrutura de *links* da Web.

A ideia básica do algoritmo é que cada página recebe uma pontuação de importância proporcional ao número de *links* que apontam para ela, chamados de *backlinks*. Podemos pensar nos *backlinks* como votos: quanto mais votos uma página recebe, mais importante ela é. Além disso, os votos não são todos iguais: os votos de páginas mais importantes valem mais do que os votos de páginas menos importantes. Assim, a Web se torna uma espécie de democracia, onde as páginas elegem as mais relevantes por meio dos *links* (BRYAN; LEISE, 2006).

O grafo da Web é composto por n páginas (vértices) numeradas de 1 a n, conforme ilustrado na Figura 8. A importância de cada página é denotada por X_k, onde k é o índice da página. Além disso, d_k e b_k indicam, respectivamente, o número de *links* que saem e que entram na página k. Por exemplo, na página 2 do grafo da Web da Figura 8, temos que $d_2 = 2$ e $b_2 = 1$

Como mencionamos, nossa proposta inicial é usar k como a quantidade de *backlinks* da página k. Quanto mais *links* uma página recebe, mais relevante ela se torna. Por exemplo, na Figura 8, temos que $X_1 = 2, X_2 = 1, X_3 = 2$ e $X_4 = 3$. Isso significa que a página 4 é a mais relevante, as páginas 1 e 3 têm a mesma relevância e, finalmente, a página 2 é a menos relevante. Então, um mecanismo de busca poderia ordenar as páginas nos seus resultados de pesquisa com base nessa classificação.

Classificação	Páginas	Valor de importância
1°	4	3
2°	1 e 3	2
3°	2	1

Tabela 1: Classificação 1

No entanto, observe que a página 1 tem um *backlink* da página mais relevante. Em outras palavras, a página mais relevante indica a página 1. Isso não deveria fazer a página 1 mais relevante do que a página 3. Estamos assumindo que *links* de páginas relevantes têm mais peso. Isso corresponde à ideia de que o usuário não vai preferir apenas as páginas com mais *backlinks*, mas também as páginas que têm *backlinks* de qualidade, de páginas relevantes. Porém, o nosso método ignora esse aspecto. Além disso, ele poderia ser facilmente manipulado, pois poderíamos criar novas páginas, intencionalmente, e adicionar um *link* para a página que quiséssemos aumentar a relevância. Assim, *links* de páginas irrelevantes teriam o mesmo peso que *links* de páginas relevantes. Um *backlink* é um *link* de outro site (o referenciador) para um recurso da web (o referente). Um recurso da web pode ser (por exemplo) um site, uma página da web ou um diretório da web. Um *backlink* é uma referência comparável a uma citação.

Sendo assim, esperamos que um bom método de ranqueamento não leve em consideração apenas o número de *backlinks* das páginas, mas, também, que um *backlink* de uma página importante deve ter um peso maior do que um *backlink* de uma página menos importante. Ou seja, passamos a nos preocupar, além da quantidade, com a qualidade dos *links*.

3.3 Calculando o *Pagerank* na Prática

Até agora, vimos como o *PageRank* funciona para Webs de tamanho reduzido. Mas como podemos lidar com a Web real, que tem quase 2 bilhões de páginas. (De acordo com a página (Internet Live Stats, 2022). Será que é possível encontrar o vetor de estado estacionário de uma matriz tão grande? A resposta é sim, mas não é trivial. O *PageRank* é obtido por meio de iterações sucessivas entre a matriz do *Google* e um vetor de estado inicial, que geralmente tem todas as entradas iguais a $1/n$, onde n é o número total de páginas da Web. De acordo com (PAGE et al., 1999), foram necessárias cerca de 52 iterações para chegar a uma boa aproximação para os *PageRanks* de um conjunto de páginas com cerca de 322 milhões de *links*. Na tabela a seguir, mostramos o número

n	$G_n \cdot X_0$	X_n
1	$G_1 \cdot X_0$	$[0,200\ 0,200\ 0,285\ 0,200\ 0,115]^T$
2	$G_2 \cdot X_0$	$[0,200\ 0,200\ 0,213\ 0,272\ 0,115]^T$
3	$G_3 \cdot X_0$	$[0,200\ 0,200\ 0,243\ 0,211\ 0,148]^T$
4	$G_4 \cdot X_0$	$[0,200\ 0,200\ 0,243\ 0,235\ 0,120]^T$
5	$G_5 \cdot X_0$	$[0,200\ 0,200\ 0,232\ 0,237\ 0,131]^T$
6	$G_6 \cdot X_0$	$[0,200\ 0,200\ 0,242\ 0,228\ 0,131]^T$
7	$G_7 \cdot X_0$	$[0,200\ 0,200\ 0,238\ 0,236\ 0,127]^T$
8	$G_8 \cdot X_0$	$[0,200\ 0,200\ 0,238\ 0,232\ 0,130]^T$
9	$G_9 \cdot X_0$	$[0,200\ 0,200\ 0,239\ 0,232\ 0,129]^T$
10	$G_{10} \cdot X_0$	$[0,200\ 0,200\ 0,238\ 0,233\ 0,129]^T$

Tabela 2: Iterações entre a matriz G e o vetor de estado inicial x_0

de iterações necessárias para alcançar os mesmos valores de importância, com três casas decimais, que encontramos ao calcular o vetor de estado estacionário da matriz G da Web com *subwebs*. Para isso, usamos o vetor de estado inicial

$$x_0 = \begin{bmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{bmatrix}.$$

Portanto, o algoritmo *PageRank* consiste em aplicar a matriz G sucessivas vezes a um vetor de probabilidade inicial para obter aproximações cada vez melhores do vetor de estado estacionário (SANTIAGO, 2021). Esse algoritmo é usado pelo *Google* Search para classificar as páginas da web em seus resultados de pesquisa. O algoritmo *PageRank* mede a importância das páginas da web contando e avaliando a qualidade dos *links* que apontam para elas. A suposição subjacente é que as páginas mais importantes tendem a receber mais *links* de outras páginas (Wikipedia, 2023). O algoritmo *PageRank* pode ser calculado para qualquer coleção de documentos com citações e referências recíprocas. O valor numérico que ele atribui a cada elemento E é chamado de *PageRank* de E e denotado por PR(E) (GeeksforGeeks, 2023). O algoritmo *PageRank* confere a cada página uma classificação de sua importância, que é uma medida recursivamente definida pela qual uma página se torna importante se páginas importantes apontarem para ela (Stanford University, 2021).

4 Exploração o pagerank no ensino Básico

4.1 Explorando no ensino médio

Uma forma de introduzir o conceito de Webs fortemente conectadas no ensino médio é mostrar exemplos reais de como elas são usadas para ordenar páginas da internet, baseando-se na solução de sistemas lineares homogêneos. Esse assunto é muito atual e relevante, pois envolve aspectos da Matemática, da Computação e da Comunicação. Depois de explorar os conceitos teóricos, os alunos podem usar alguns softwares, como o *Winmat*, para realizar os cálculos mais complexos e verificar os resultados obtidos.

Uma forma de introduzir o conceito de Webs fortemente conectadas no ensino médio é mostrar exemplos reais de como elas são usadas para ordenar páginas da internet, baseando-se na solução de sistemas lineares homogêneos. Esse assunto é muito atual e relevante, pois envolve aspectos da Matemática, da Computação e da Comunicação. Depois de explorar os conceitos teóricos, os alunos podem usar alguns softwares, como o *Winmat*, para realizar os cálculos mais complexos e verificar os resultados obtidos.

Vamos exemplificar algumas atividade que acreditamos serem interessante para abordagem em sala de aula.

4.1.1 Atividade 1

Observe a matriz abaixo de *links* e faça os itens que seguem:

$$A = \begin{bmatrix} 0 & \frac{1}{3} & 1 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{3} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} & 0 & 0 \end{bmatrix}$$

a) A rede tem quantos sites? Essa rede é admissível? Por quê? A rede é fortemente conectada? Sugestão: gere o grafo associado à Web representada pela matriz, a fim de visualizar o problema.

b) Existe *link* do site 4 para o site 1? E do site 1 para o site 4?

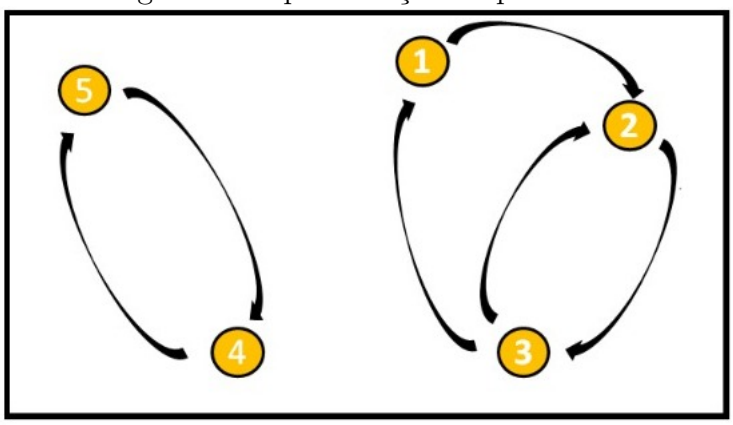
c) Estabeleça o ranking das páginas. Qual delas é a mais importante? E a menos importante?

4.1.2 Atividade 2

Considere a rede associada ao grafo abaixo:

a) A rede é admissível? É fortemente conectada? Determine a matriz de *links*, denotando-a por A. Proponha uma ordenação para os sites.

Figura 9: Representação do problema



fonte: Autoria própria, 2023

b) Utilize o “truque” da matriz do tipo M = (1 – m)A + mS para estabelecer o ranking das páginas dessa rede. Use m = 0,1.

4.1.3 Atividades 3

Suponha que os responsáveis pelo site 3 na Web da figura apresentada no texto ficaram chateados com o fato de o índice de importância do site 3 ter ficado menor do que o do site 1 e decidiram criar um site 5, com *link* para o site 3, colocando um *link* também de 3 para 5.

a) A rede continua sendo admissível? E fortemente conectada?

b) Isso torna a importância da página 3 maior do que a da página 1? Justifique

4.2 Comentários sobre as atividades

Essas atividades tem como foco ”Matrizes e o ranking de sites na internet”, que apresenta uma forma de aplicar os conceitos de matrizes e sistemas lineares na análise da relevância de páginas web. Propõe três atividades para os alunos do ensino médio, que envolvem a construção e a manipulação de matrizes associadas a redes de sites, bem como a resolução de sistemas lineares para obter as importâncias relativas de cada site. O objetivo é mostrar como a Matemática pode ser usada para entender e melhorar o posicionamento de um site nos mecanismos de busca.

O texto também discute algumas questões que podem despertar a curiosidade e o espírito investigativo dos estudantes, como a relação entre a estrutura da rede e a solução do sistema linear. Busca oferecer uma alternativa ao ensino tradicional de matrizes, que muitas vezes se limita aos aspectos operacionais e não explora as possibilidades de aplicação e contextualização desse conteúdo.

- A primeira atividade tem como objetivo introduzir o conceito de matriz e sua relação com a representação geométrica de uma rede.

- A segunda atividade mostra como alterar a matriz A para obter as importâncias dos sites na rede, usando a resolução de sistemas lineares.

- A terceira atividade exemplifica como um administrador de sites pode usar esse conhecimento para melhorar o posicionamento do seu site no ranking.

- Propõe uma aplicação atual e interessante de matrizes e sistemas lineares, que pode despertar a curiosidade e o espírito investigativo dos alunos.

- Também sugere algumas questões que podem ser exploradas pelos alunos ou pelo professor, para ampliar o entendimento do tema e sua relevância na Matemática e na vida.

4.3 Respostas das atividades

4.3.1 Resposta atividade 1

Atividade 1: A rede possui quatro sites, é admissível e fortemente conectada. Não existe *link* do site 4 para o site 1, mas existe *link* do site 1 para o site 4. Ranqueamento pelo método *PageRank* do mais importante para o menos importante: 1, 2, 4, 3.

4.3.2 Resposta atividade 2

A rede é admissível, mas não é fortemente conectada. A matriz de *links* é dada por:

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

Usando o truque proposto no segundo item, o site 3 é o mais importante, seguido do 2. A seguir vêm os sites 4 e 5, ambos considerados igualmente importantes, e o menos importante é o site 1.

4.3.3 resposta atividade 3

A rede continua sendo admissível e fortemente conectada. A importância da página 3 torna-se maior do que a da página 1 com a nova configuração.

Considerações finais

A matemática é uma ciência fascinante que está presente em tudo o que nos rodeia. Mas nem sempre conseguimos perceber a sua beleza e utilidade. Por isso, precisamos questionar, explorar e descobrir como a matemática funciona e se aplica à realidade. Um exemplo muito interessante é o *Google*, o maior e mais popular motor de busca da internet. Como é que o *Google* consegue encontrar as páginas mais relevantes para cada pesquisa que fazemos? A resposta está no Algoritmo *PageRank*, uma criação genial que usa a matemática para medir a importância de cada página da Web. O Algoritmo *PageRank* atribui um valor numérico a cada página, baseado na quantidade e qualidade dos *links* que apontam para ela. Esse valor é usado para ordenar as páginas por ordem de relevância quando o *Google* mostra os resultados de uma pesquisa. Assim, a matemática é a chave do sucesso do *Google* e uma ferramenta poderosa para organizar e acessar a informação na internet. Neste trabalho, explicamos, de forma clara e didática, como a matemática está envolvida no processo de ranqueamento das páginas da Web.

O algoritmo *PageRank* é uma fascinante e elegante aplicação da Álgebra Linear, que também envolve conceitos da Teoria da Probabilidade. Este trabalho não pretende ser exaustivo, nem abordar todos os aspectos do algoritmo. Porém, esperamos que o leitor possa ter uma boa ideia de como o *PageRank* funciona, e da matemática que está por trás dele.

Ademais, além deste trabalho servir como um incentivo ao estudo de assuntos do campo da Álgebra Linear e da Probabilidade, tais como matrizes, sistemas de equações, processos estocásticos, ele também evidencia, a medida que expõe como o algoritmo foi modelado, a beleza do processo que é traduzir ideias em linguagem matemática

Referências

[1] BARROS, Cícero Demétrio Vieira. *O Teorema do Ponto Fixo de Banach e algumas Aplicações.*2013. 46 p. Dissertação de Mestrado (PROFMAT) - Universidade Federal da Paraíba, João Pessoa, 2013

[2] BERNERS-LEE, T. J. The world-wide web.*Computer networks and ISDN systems,* Elsevier, v. 25, n. 4-5, p. 454Ű459, 1992

[3] BILLINTON R.*Power System Reliability Evaluation, Routledge,1970,*

[4] BOLDRINI, J. L. et al *Álgebra Linear I.* São Paulo: Harper & Row do Brasil, 1980.

[5] FILHO, J. R. R.*A matemática por trás do algoritmo Pagerank do Google,* Instituto Federal da Praíba,2022

[6] HILLIER, Frederick S.; LIEBERMAN, Gerald J.*Introdução à Pesquisa Operacional,Stanford University,1995.*

[7] LEHMAN, E.; LEIGHTON, T.; MEYER, A. R.*Mathematics for computer science.*Relatório Técnico, 2010. Acesso em: 18 ago. 2018.

[8] MALAJOVICH, G. *Álgebra linear.* UFRJ: [s.n.], 2021

[9] MORAIS, E. A. M.; AMBRÓSIO, A. P. L. *Ferramentas de busca na internet.*Relatório Técnico: Universidade Federal de Goiás, 2007.

[10] NORRIS, J.R.*Markov Chains,*Cambridge University Press,1998.

[11] PAGE, L. et al. *The PageRank citation ranking:* Bringing order to the web. [S.l.], 1999.

[12] POOLE, D.*Linear algebra:*a modern introduction. Trent University: Cengage Learning, 2014

[13] SILVA, T. C. M. da; JÚNIOR, V. V. *Cadeias de Markov:* conceitos e aplicações em modelos de difusão de informação. [S.l.], 2011.

[14] STOLFI, A. d. S. *World wide web:* forma aparente e forma oculta: webdesign da interface ao código. Tese (Doutorado) Ů Universidade de São Paulo, 2010