

Felipe Moraes da Silva

Uma Prova Conceitual Aplicada a um Modelo de Rede para Ordenação de Páginas da Internet

Rio Grande, Rio Grande do Sul, Brasil

Junho, 2018

Felipe Moraes da Silva

Uma Prova Conceitual Aplicada a um Modelo de Rede para Ordenação de Páginas da Internet

Trabalho de Conclusão de Curso, Matemática Aplicada Bacharelado, submetido por Felipe Moraes da Silva junto ao Instituto de Matemática, Estatística e Física da Universidade Federal do Rio Grande.

Universidade Federal do Rio Grande - FURG

Instituto de Matemática, Estatística e Física - IMEF

Curso de Matemática Aplicada Bacharelado

Orientador: Dra. Catia Maria dos Santos Machado

Coorientador: Dr. André Andrade Longaray

Rio Grande, Rio Grande do Sul, Brasil

Junho, 2018

Este trabalho é dedicado aos meus pais, com todo meu amor e gratidão, por tudo que fizeram por mim ao longo de minha vida.

Agradecimentos

Quero agradecer, em primeiro lugar, a Deus, pela força e coragem durante toda esta longa caminhada.

A Universidade Federal do Rio Grande, pela oportunidade de fazer o curso.

A todos os professores do curso, que foram tão importantes na minha vida acadêmica.

A minha orientadora Dra. Catia M^a dos Santos Machado, pelo empenho dedicado à elaboração deste trabalho, pela amizade, pelo incentivo e carinho. Fico muito grato de ter me aceitado como seu orientando.

Aos membros da banca examinadora, composta pelos professores André e Celiane, pela disponibilidade de participar e pelas contribuições feitas para melhoria do trabalho.

Ao Ms. Alessandro Saadi, pela confiança de me aceitar como bolsista no Programa de Incentivo à Matemática - PRIMA, pelo incentivo e apoio, espero contar sempre com sua amizade.

A minha família, pelo amor, incentivo e apoio incondicional. Sem vocês a realização deste sonho não seria possível.

Aos amigos e colegas, pelo convívio, incentivo e apoio constantes.

E a todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigado.

“Eu tentei 99 vezes e falhei, mas na centésima tentativa eu consegui, nunca desista de seus objetivos mesmo que esses pareçam impossíveis, a próxima tentativa pode ser a vitoriosa.”

(Albert Einstein)

Resumo

O presente trabalho é direcionado aqueles que estudam Matemática Aplicada e tem como objetivo principal estudar os fundamentos matemáticos do mecanismo de busca das páginas da internet, em especial sobre o método de ordenação de páginas utilizando Cadeia de Markov. O PageRank (denominação dada pelo programa de busca Google), ordena as páginas da internet por meio de um vetor estacionário de uma matriz de transição utilizada mediante o Método das Potências. Nesse sentido, uma prova conceitual, a partir de um modelo de rede, mostrará como é obtida a ordenação de páginas da internet.

Palavras-chaves: Grafos, Método das Potências, Ordenação de Páginas, Cadeias de Markov.

Abstract

The present work is directed to those who study Applied Mathematics and its main objective is to study the mathematical fundamentals of the search engine of web pages, especially about the method of ordering pages using Markov Chain. PageRank (the name given by the Google search program) sorts the web pages by means of a stationary vector of a transition matrix used by the Power Method. In this sense, a conceptual proof, from a network model, will show how the ordering of Internet pages is obtained.

Key-words: Graphs, Power Method, Sorting Pages, Markov chains.

Lista de ilustrações

Figura 1 – Grafo representativo da web.	24
Figura 2 – Grafo representativo quando a página C é acessada.	30

Sumário

	Introdução	9
1	OBJETIVOS	11
2	FUNDAMENTAÇÃO TEÓRICA	12
2.1	Teoria dos Grafos	12
2.2	Teoria Espectral	13
2.2.1	Autovalores e Autovetores	13
2.3	Matrizes Não-Negativas	16
2.3.1	Matrizes Redutíveis e Irredutíveis	16
2.4	O Teorema de Perron-Frobenius	17
3	CADEIAS DE MARKOV	18
3.1	Processos Estocásticos	18
3.2	Processos Markovianos	18
3.3	Matriz Estocástica	19
3.4	Matriz de Transição	19
3.5	Autovalores e Autovetores de Matrizes de Transição	20
3.6	Comportamento de P^n quando $n \rightarrow \infty$	21
3.7	Comportamento de $P^n v$ quando $n \rightarrow \infty$	22
3.8	Vetor de Probabilidade Estacionário	23
4	APLICAÇÃO	24
4.1	Cálculo do PageRank utilizando o Método das Potências, caso Markoviano	29
4.2	Cálculo do PageRank através do Vetor de Probabilidade Estacionário (q)	31
5	CONCLUSÃO	34
	REFERÊNCIAS	35

Introdução

Esse trabalho é direcionado aqueles que estudam Matemática Aplicada e possuem interesse nos fundamentos matemáticos da teoria utilizada para desenvolver algoritmos PageRank (denominação dada pelo programa de busca Google) desenvolvido por Larry Page e Sergey Brin, para executar serviços de busca na internet. Na literatura, os algoritmos que envolvem PageRank são conhecidos geralmente por serem utilizados na resolução de problemas de busca, caracterização de redes sociais e recuperação da informação (Langville e Meyer, 2011).

Nos últimos anos a pesquisa na Web obteve um crescimento exponencial, tornando-se o maior repositório de informações já criado pelo homem e representando uma fonte nova e relevante de informações potencialmente úteis para diversas áreas do conhecimento (Reis, 2013). Devido ao grande volume de dados, algoritmos de busca e recuperação são desenvolvidos para resolver o problema desafiador relacionado à utilização de informações.

O conceito por trás da formulação do PageRank, utilizado nesse trabalho, é com relação a medida de importância de páginas na web, ou seja, da sua importância pelas citações que ela recebe de outras páginas. Assim, uma página não é muito importante se ela apenas cita outras páginas, porém não é citada.

Nesse sentido, o trabalho tem como objetivo principal estudar os fundamentos matemáticos do mecanismo de busca das páginas da internet, em especial sobre o método de ordenação de páginas utilizando Cadeia de Markov.

O algoritmo PageRank é estudado por muitos pesquisadores. O trabalho de Bovoloni et al. (2016) esclarece de forma didática e sucinta, sobre o mecanismo de pesquisa PageRank, trazendo importantes conceitos da Álgebra Linear. Segundo, Oliveira (2014), o algoritmo PageRank do Google, utiliza conceitos de matrizes, sistemas lineares, probabilidades, noções básicas de grafos dirigidos e cadeias de Markov de tempo discreto. Magalhães (2014), apresenta em seu trabalho técnicas de recuperação de informações utilizando uma metodologia de busca e classificação de páginas na Web. Jianya (2012) mede a influência dos membros de redes sociais aplicando o PageRank e Índice W-Entropia. Dessa forma o autor utiliza o algoritmo PageRank para calcular a importância de cada pessoa com base na ligação intrínseca entre os membros de uma rede social. O trabalho de Melo (2009) trás algumas definições e propriedades úteis para compreensão do algoritmo PageRank. No entanto, seu objetivo maior é detalhar por meio de simulações computacionais os diversos fatores que influenciam no tempo de convergência do algoritmo PageRank.

Contribuindo com os trabalhos supracitados, esse trabalho se utiliza do elegante

Teorema de Perron-Frobenius aplicado a grafos fortemente conexos, revelando propriedades matemáticas que descrevem a natureza dos autovalores dominantes, autovetores de valores positivos e matrizes não negativas, essenciais para o desenvolvimento de métodos computacionais que solucionam o problema. Sobre um modelo de rede, mostrar-se-á uma prova conceitual sobre os fundamentos matemáticos abordados na resolução do problema de ordenação de páginas na internet.

1 Objetivos

O trabalho tem por objetivo principal estudar os fundamentos matemáticos do mecanismo de busca das páginas da internet, em especial sobre o método de ordenação de páginas utilizando Cadeia de Markov.

Para alcançar o objetivo principal, os seguintes objetivos foram considerados no desenvolvimento do trabalho.

- (i) Relacionar conceitos importantes de grafos, teoria espectral e probabilidade que são utilizados no desenvolvimento de algoritmos de PageRank.
- (ii) Apresentar o Teorema de Perron-Frobenius garantindo um grafo fortemente conexo.
- (iii) Mostrar a variação do Método das Potências via cadeia de Markov aplicada à matrizes não simétricas .
- (iv) Fazer uma prova de conceito, sobre um modelo de rede da web, utilizando os resultados teóricos que fundamentam o PageRank.

2 Fundamentação Teórica

Nesse capítulo são apresentados os resultados matemáticos que são utilizados no desenvolvimento de algoritmos de busca na internet. Inicialmente são apresentados conceitos da Teoria dos Grafos como passeio, conexidade, matriz de adjacência entre outros. Em seguida, é abordada a Teoria Espectral, ramo da Álgebra Linear, cujas definições conduzirão ao Teorema de Perron-Frobenius aplicado a grafos fortemente conexos.

2.1 Teoria dos Grafos

Definição 2.1.1. Grafo não orientado: Um grafo é um par $G = (V, E)$, onde V é um conjunto não vazio, finito de vértices, e E é um subconjunto finito de arestas de dois elementos do conjunto V .

Uma aresta $e \in E(G)$ é representada por $e = \{u, v\}$ sempre que interliga dois vértices u e v de V . Dois vértices ligados por uma mesma aresta são denominados adjacentes e pode-se dizer que uma aresta e é incidente em u , se u for uma extremidade de e .

Definição 2.1.2. Grau: Dado um grafo $G(V, E)$, o grau de um vértice $v \in V$, denotado por $d(v)$, é igual ao número de arestas que incidem em v .

Definição 2.1.3. Passeio: Um passeio em um grafo G é uma sequência de vértices, em que cada vértice é adjacente ao seguinte, isto é, $w = (v_0, v_1, \dots, v_l)$. Nessa sequência é permitido repetir vértices. O comprimento desse passeio é l . Observe que a enumeração dos passeios é a partir do índice 0 e que existe $l + 1$ vértices no passeio.

Definição 2.1.4. Caminho: O caminho é uma sequência de vértices adjacentes onde nessa sequência não é permitido repetir vértices.

Definição 2.1.5. Conexo: Um grafo G é conexo se, e somente se, existir pelo menos um passeio entre todos os pares de vértices do grafo G . Caso algum vértice de G não tenha passeio com algum outro vértice de G , chama-se esse grafo de desconexo.

Definição 2.1.6. Laço: Um laço é uma aresta que possui as duas extremidades em um mesmo vértice de G .

Definição 2.1.7. Grafo orientado (dirigido): Um grafo orientado (dirigido) $G = (V, E)$ consiste de um conjunto de vértices V não vazio e de um conjunto de arestas orientadas. Cada aresta está associada a um par ordenado de vértices. Uma aresta orientada associada a um par ordenado (u, v) começa em u e termina em v .

Quando se descreve um grafo orientado, usa-se uma flecha apontando de u para v , indicando que começa em u e termina em v .

Definição 2.1.8. Grau de entrada e saída: Denomina-se de **grau de entrada** (ou semigrau exterior) de um vértice v , $d^-(v)$, a quantidade de arestas que chegam até v . Denomina-se ainda de **grau de saída** (ou semigrau interior) de um vértice v , $d^+(v)$, a quantidade de arestas que partem de v .

Definição 2.1.9. Fonte e Sumidouro: Uma **fonte** de um grafo orientado é um vértice cujo grau de entrada é igual zero. Um **sumidouro** (ou poço) de um grafo orientado é um vértice cujo grau de saída é igual a zero.

Definição 2.1.10. Fortemente conexo: Um grafo G orientado é **fortemente conexo** se, para cada par de vértices distintos u e v , existe um caminho de u para v e um caminho de v para u .

Definição 2.1.11. Matriz de adjacência : Seja $G = G(V, E)$ um grafo com n vértices. A **matriz de adjacência** $A(G)$ de G é a matriz quadrada de ordem n cujas entradas são:

$$a_{ij} = \begin{cases} 1, & \text{se e somente se, existe um par ordenado } (v_i, v_j) \in E \\ 0, & \text{nos outros casos} \end{cases}$$

Isto significa que $a_{ij} = 1$ quando os vértices v_i e v_j são adjacentes, e $a_{ij} = 0$ em caso contrário. Cabe salientar que, as matrizes de adjacência para grafos não orientado são simétricas.

Observação 2.1.1. Na literatura a definição de grafos não está totalmente padronizada. A denominação **multigrafos** é utilizada para grafos dirigidos ou não dirigidos que possuam arestas paralelas e ou laços.

2.2 Teoria Espectral

2.2.1 Autovalores e Autovetores

Se A for uma matriz $n \times n$, então um vetor não nulo x em $\mathbb{R}^n$ é denominado **autovetor** de A se Ax for um múltiplo escalar de x , isto é, $Ax = \lambda x$ com algum escalar λ . O escalar λ é denominado **autovalor** de A , e diz-se que x é um autovetor associado a λ .

A equação matricial $Ax = \lambda x$ pode ser reescrita como:

$$(A - \lambda I_n)x = 0 \tag{2.1}$$

Uma vez que x é diferente de zero, a equação 2.1 implica que λ é um autovalor de A se e somente se $A - \lambda I_n$ é uma matriz singular; o que equivale a dizer que λ é um autovalor se e só se:

$$\det(A - \lambda I_n) = 0 \quad (2.2)$$

Segue da definição de determinantes que se A é uma matriz de ordem n , então $p(\lambda) = \det(A - \lambda I_n)$ é um polinômio em λ de grau n , chamado de polinômio característico de A .

Definição 2.2.1. Autoespaço: Chama-se de autoespaço de λ e denota-se por W_λ o conjunto dos autovetores associados ao autovalor λ .

Para cada λ de A , associam-se dois números inteiros chamados de multiplicidade algébrica e multiplicidade geométrica, definidos a seguir:

Definição 2.2.2. Multiplicidade Algébrica: Dado λ um autovalor de A , **multiplicidade algébrica** de λ é o número de vezes que o autovalor aparece como raiz de um polinômio característico.

Definição 2.2.3. Multiplicidade Geométrica: Dado λ um autovalor de A , **multiplicidade geométrica** de λ é a dimensão do autoespaço W_λ associado ao autovalor.

Proposição 2.2.1. Seja $A = [a_{ij}]$ uma matriz real e simétrica de ordem n , então cada um dos seus autovalores são números reais.

Definição 2.2.4. Matriz Diagonalizável: Uma matriz A é diagonalizável, se existe uma matriz M inversível e uma matriz D diagonal, tal que $M^{-1}AM = D$.

Observações:

- i) Uma matriz $A_{n \times n}$ é diagonalizável quando existir n autovetores linearmente independentes de A .
- ii) Se A é uma matriz $n \times n$ com n autovalores distintos entre si, então A é diagonalizável.
- iii) Em uma matriz diagonalizável a multiplicidade algébrica de cada autovalor é igual a multiplicidade geométrica.

Definição 2.2.5. Espectro de uma matriz: O **Espectro de uma matriz** quadrada A , denotado por $\sigma(A)$ é o conjunto formado por todos os autovalores de A .

Definição 2.2.6. Raio Espectral: O **Raio Espectral** de A , denotado por $\rho(A)$ é definido por:

$$\rho(A) = \max\{|\lambda| : \lambda \in \sigma(A)\}$$

Definição 2.2.7. Autovalor dominante: Um autovalor de uma matriz A que é igual ao raio espectral de A é chamado de **autovalor dominante**. O autovetor associado a este autovalor é chamado de **autovetor dominante**.

Teorema 2.2.1. Sejam $\lambda_1, \lambda_2, \dots, \lambda_m$ os m autovalores de uma matriz A de ordem n e sejam $v_1, v_2, \dots, v_m$ os correspondentes autovetores. Supondo que λ_1 seja o autovalor dominante de modo que

$$|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_j| \geq \dots \geq |\lambda_m|$$

Considerando x_0 um vetor do subespaço gerado pelos autovetores, isto é,

$$x_0 = c_1 v_1 + c_2 v_2 + \dots + c_m v_m$$

com $c_1 \neq 0$. Então $x_k = \frac{A^k x_0}{\|A^k x_0\|}$ converge para o autovetor dominante v_1

Demonstração: (Freitas, 2012) Temos:

$$A^k x_0 = c_1 A^k v_1 + c_2 A^k v_2 + \dots + c_m A^k v_m$$

$$A^k x_0 = c_1 \lambda_1^k v_1 + c_2 \lambda_2^k v_2 + \dots + c_m \lambda_m^k v_m$$

$$A^k x_0 = c_1 \lambda_1^k \left[v_1 + \frac{c_2}{c_1} \left(\frac{\lambda_2}{\lambda_1} \right)^k v_2 + \dots + \frac{c_m}{c_1} \left(\frac{\lambda_m}{\lambda_1} \right)^k v_m \right] \quad (2.3)$$

A expressão entre colchetes converge para v_1 pois $\left| \frac{\lambda_j}{\lambda_1} \right| < 1$ para $j > 1$. Então a expressão:

$$x_k = \frac{A^k x_0}{\|A^k x_0\|}$$

converge para o autovetor dominante.

Pode-se observar na equação 2.3, que a razão entre os dois maiores autovalores determina quão rápida é a convergência pois, uma vez que temos λ_2 próximo em magnitude do autovalor dominante λ_1 , torna-se necessário que o valor de k se aproxime de um número muito grande para que $|\lambda_2/\lambda_1|^k$ tenda à zero.

O teorema acima fornece um algoritmo para aproximar o autovalor do um autovetor unitário associado a uma matriz simétrica A , desde que o autovalor dominante seja positivo. Esse algoritmo é denominado Método das Potências com mudança de Escala Euclidiana e é descrito abaixo.

Algoritmo 2.2.1. O Método das Potências com Mudança de Escala Euclidiana

- Passo 1.** Escolha o vetor não-nulo qualquer e normalize, se necessário para obter um vetor unitário x_0 .
- Passo 2.** Calcule Ax_0 e normalize para obter a primeira aproximação x_1 de um autovetor dominante unitário. Calcule $Ax_1 \cdot x_1$ para obter a primeira aproximação do autovalor dominante.
- Passo 3.** Calcule Ax_1 e normalize para obter a segunda aproximação x_2 de um autovetor dominante unitário. Calcule $Ax_2 \cdot x_2$ para obter a segunda aproximação do autovalor dominante.
- Passo 4.** Calcule Ax_2 e normalize para obter a terceira aproximação x_3 de um autovetor dominante unitário. Calcule $Ax_3 \cdot x_3$ para obter a terceira aproximação do autovalor dominante.

Continuando assim, em geral, obtém-se uma sequência de aproximações cada vez melhores do autovalor dominante e de um autovetor unitário associado.

O teorema 2.2.1 enunciado anteriormente para matrizes simétricas, também é válido para qualquer matriz A de tamanho $n \times n$ que tenha n -autovetores linearmente independentes e um autovalor dominante. Mais detalhes pode ser encontrado em Anton e Rorres (2012).

2.3 Matrizes Não-Negativas

Uma matriz é chamada não-negativa se todas as suas entradas são maiores ou iguais a zero. Chama-se uma matriz de positiva quando suas entradas são maiores que zero.

2.3.1 Matrizes Redutíveis e Irredutíveis

Definição 2.3.1. Matrizes Redutíveis e Irredutíveis: Uma matriz quadrada de ordem n é **irredutível** se a matriz de adjacência $\mathcal{A}$ está associada a um grafo fortemente conexo, caso contrário diz-se que a matriz é **redutível**.

Corolário 2.3.1. Se A é uma matriz de ordem n , não-negativa e irredutível, então todo autovetor não-negativo de A deve ser positivo.

2.4 O Teorema de Perron-Frobenius

Em muitas aplicações da teoria de matrizes, o autovalor de interesse é o autovalor positivo e dominante igual ao raio espectral (ver definição 2.2.6), que geometricamente é o raio do menor círculo centrado na origem que contém o espectro de A , chamado círculo espectral de A , e o autovetor correspondente com componentes positivas. Daí surge a importância do teorema de Perron-Frobenius para uma classe numerosa de matrizes. A primeira versão deste teorema foi provado para matrizes positivas pelo matemático alemão Oskar Perron (1880-1975), chamado Teorema de Perron e publicado em 1907 em um artigo sobre frações contínuas (Poole, 2004). Frobenius, a partir do resultado de Perron conseguiu provar um teorema mais geral, cuja formulação baseia-se no conceito de matrizes irredutíveis.

O Teorema de Perron-Frobenius está demonstrando no trabalho de Freitas (2012) e está enunciado a seguir:

Teorema 2.4.1. (Perron-Frobenius) Seja A uma matriz quadrada de ordem n , irredutível e não-negativa, então:

- (i) A tem um autovalor positivo r , igual ao raio espectral de A ;
- (ii) Existe um autovetor positivo associado a r ;
- (iii) O autovalor r tem multiplicidade algébrica igual a 1.

No próximo capítulo, serão apresentados conceitos sobre Cadeias de Markov e o Teorema de Perron-Frobenius no caso Markoviano, a fim de apresentar a variação do Método das Potências.

3 Cadeias de Markov

Esse capítulo trata em especial do modelo de probabilidade para os Processos Estocásticos Markovianos, que evoluem no tempo, em espaço de estado discreto. Resultados importantes sobre convergência em Cadeias de Markov, bem como do Teorema de Perron-Frobenius no caso Markoviano também são apresentados.

3.1 Processos Estocásticos

Um Processo Estocástico é definido como uma coleção de variáveis aleatórias (X_n) indexados por um parâmetro n pertencente a um conjunto T . Frequentemente T é tomado para ser um conjunto dos inteiros não negativos (porém, outros conjuntos são perfeitamente possíveis) e (X_n) representa uma característica mensurável de interesse no passo n . Exemplificando, (X_n) pode representar as páginas da web acessadas no passo n .

Processos Estocásticos são de interesse para descrever o procedimento do sistema operando sobre algum passo n (ou período de tempo t), com isso, em termos formais a variável aleatória (X_n) representa o estado do sistema no passo n (ou tempo t). Portanto, pode-se afirmar que (X_n) é definido como um espaço denominado Espaço de Estados.

Em relação ao Estado, o Processo Estocástico pode ser: Estado Discreto (cadeia) quando, (X_n) é definido sobre um conjunto enumerável ou finito, ou Estado Contínuo (sequência) caso contrário.

Em relação ao Passo n (ou tempo t), tem-se o tempo discreto quando n é finito ou enumerável ou então é tempo contínuo caso contrário.

3.2 Processos Markovianos

Um Processo Estocástico é dito ser um Processo Markoviano se:

$$\begin{aligned} P(X_n = x_n | X_{n-1} = x_{n-1}) &= P(X_n = x_n | X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \dots, X_0 = x_0) \\ &= P(X_n = x_n | X_{n-1} = x_{n-1}) \end{aligned} \quad (3.1)$$

A expressão 3.1 pode ser traduzida pela probabilidade condicional de qualquer evento futuro, dado que qualquer evento passado e o estado presente X_{n-1} , é independente do evento passado e depende somente do estado presente. Em outras palavras o estado futuro depende apenas do estado presente e não dos estados passados.

Definição 3.2.1. Seja $\{X_n, n = 0, 1, 2, 3, \dots\}$ um processo aleatório, tomando seus valores num conjunto $T = \{A, B, C, \dots\}$. Diz-se que $\{X_n\}$ é uma cadeia de Markov se a

probabilidade $P(X_n)$, $X_n \in T$, depender somente do valor do processo no passo anterior, X_{n-1} , e não em qualquer dos passos anteriores $X_{n-2}, X_{n-3}, \dots, X_0$.

3.3 Matriz Estocástica

Definição 3.3.1. Matriz Estocástica: Seja $A_{m \times n} = (a_{ij})$ ($a_{ij} \in \mathbb{R}$) tal que $a_{ij} \geq 0$, $\forall i, j$, com $\sum_{i=1}^n a_{ij} = 1$, para todo j . Isto é, A é uma matriz não-negativa (todos elementos são não negativos), onde a soma de cada uma de suas colunas é igual a 1. Então uma matriz de probabilidade é uma **matriz estocástica**.

Se A for uma matriz coluna então A é um **vetor de probabilidade**.

Observação 3.3.1. Pode-se obter a soma dos elementos das colunas de $A_{m \times n}$ através do produto $J \cdot A$, onde $J \in M_{1 \times m}$ possui todas suas entradas iguais a 1.

Corolário 3.3.1. O produto de duas matrizes estocásticas é uma matriz estocástica.

Teorema 3.3.1. O produto de uma matriz positiva por uma matriz estocástica é uma matriz positiva.

Proposição 3.3.1. Se $P_{n \times n}$ é uma matriz estocástica e $v \in \mathbb{R}^n$ é um vetor de probabilidade, então Pv é um vetor de probabilidade.

3.4 Matriz de Transição

Definição 3.4.1. Matriz de Transição: Se uma cadeia de Markov tiver k estados possíveis, identificados por $1, 2, \dots, k$, então a probabilidade do sistema estar no estado i (passo atual) dado que estava no estado j (passo anterior), é denotada por p_{ij} . Onde p_{ij} é a **probabilidade de transição** do estado j ao estado i . Assim, a matriz quadrada $P = [p_{ij}]$, é dita **matriz de transição** de uma Cadeia de Markov, se e somente se, estiver definida por $P(X_n = x_n | X_{n-1} = x_{n-1}) = P(X_n = x_n | X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \dots, X_0 = x_0) = P(X_n = x_n | X_{n-1} = x_{n-1}) = p_{ij}$, com $p_{ij} \in [0, 1]$ e $\sum_{i \in T} p_{ij} = 1$ para todo $i, j \in T$.

Para simplificar a notação, sem perda de generalidade faz-se: $x_n = i$ e $x_{n-1} = j$. Logo $P(X_n = i | X_{n-1} = j) = p_{ij}$

Definição 3.4.2. Vetor Estado: O **vetor estado** de uma observação de uma Cadeia de Markov com k estados é um vetor coluna v_n cujo i -ésimo componente v_i é a probabilidade do sistema estar, naquela observação, no i -ésimo estado.

As entradas em qualquer vetor estado de uma cadeia de Markov são não negativas e têm soma 1 (vetor de probabilidade). Note que o vetor de estado v_0 denota o vetor na observação inicial de uma cadeia de Markov. O teorema seguinte permitirá determinar os vetores estado $v_1, v_2, \dots, v_n, \dots$, nas observações subsequentes.

Teorema 3.4.1. (Método das Potências, caso Markoviano) Se P é a matriz de transição de uma Cadeia de Markov e v_n é um Vetor estado na n -ésima observação, então $v_{n+1} = Pv_n$

De modo iterativo, obtém-se

$$\begin{aligned} v_1 &= Pv_0 \\ v_2 &= Pv_1 = P^2v_0 \\ v_3 &= Pv_2 = P^3v_0 \\ &\vdots \\ v_n &= Pv_{n-1} = P^n v_0 \end{aligned}$$

Dessa forma, o Vetor estado inicial v_0 e a matriz de transição P determinam v_n para $n = 1, 2, \dots$. Logo, (p_{ij}^n) é a probabilidade de se passar do estado j ao estado i em n transições. Assim a sequência de vetores de probabilidade $v_1, v_2, \dots, v_n$ juntamente com a matriz de transição P forma uma Cadeia de Markov.

Definição 3.4.3. Vetor de Probabilidade Estacionário: Em uma cadeia de Markov, o vetor v_{n_0} é **estacionário** se $P \cdot v_{n_0} = v_{n_0}$. Isto é, v_{n_0} é estacionário se para todo $n > n_0$, $v_n = v_{n_0}$.

Definição 3.4.4. Matriz Regular: Uma matriz é dita **regular** se alguma de suas potências for positiva.

Corolário 3.4.1. Se P é a matriz de transição de uma cadeia de Markov então, P^n ($n \in \mathbb{N}$) é uma matriz estocástica.

Corolário 3.4.2. Seja P é uma matriz de transição positiva, qualquer potência de P será positiva.

Corolário 3.4.3. Se P é uma matriz de transição regular, digamos com P^n positiva, então P^{n+w} , $\forall w \in \mathbb{N}$, será positiva.

3.5 Autovalores e Autovetores de Matrizes de Transição

Teorema 3.5.1. Se P é uma matriz de transição de ordem n de uma cadeia de Markov, então 1 é autovalor de P .

Teorema 3.5.2. Seja P uma matriz de transição de ordem n de uma cadeia de Markov, com autovalor λ então $|\lambda| \leq 1$. E ainda se P é regular, $\lambda \neq -1$.

Teorema 3.5.3. Seja P uma matriz de transição regular então $\exists n \in \mathbb{N}$, tal que o autovalor $\lambda = 1$ de $(P^{n+w})^t$, $w = 0, 1, 2, \dots$, tem multiplicidade geométrica igual a 1.

Teorema 3.5.4. Seja P uma matriz de transição regular diagonalizável de uma cadeia de Markov então $\exists n \in \mathbb{N}$ tal que o autovalor $\lambda = 1$ de P^{n+w} ($\forall w \in \{1, 2, \dots\}$) tem multiplicidade geométrica igual a 1.

Observação 3.5.1. Esse teorema equivale a afirmar que: “Existe $n \in \mathbb{N}$ tal que todos os autovetores de P^{n+w} (para cada $w = 0, 1, \dots$) associados ao autovalor $\lambda = 1$ são linearmente dependentes.

Teorema 3.5.5. (Teorema de Perron-Frobenius, caso Markoviano) Seja P uma matriz de transição de uma cadeia de Markov, então

- (i) Se λ é autovalor de P , então $|\lambda| \leq 1$;
- (ii) $\lambda = 1$ é autovalor de P .
- (iii) $\lambda = 1$ tem multiplicidade algébrica e geométrica igual a 1.

Uma Cadeia de Markov que é regida por uma matriz de transição regular é chamada de **Cadeia de Markov Regular**. A seguir será visto que qualquer Cadeia de Markov Regular possui um vetor estacionário q tal que, para qualquer escolha de v_0 , o vetor $P^n v_0$ converge para q quando n aumenta. O cálculo de q se dá simplesmente pela solução do sistema linear homogêneo $(P - \lambda I)q = 0$, com $\lambda = 1$.

3.6 Comportamento de P^n quando $n \rightarrow \infty$

Teorema 3.6.1. Seja P uma matriz de transição regular de uma cadeia de Markov diagonalizável, então existe o limite de P^n . Logo:

$$P \text{ regular diagonalizável} \Rightarrow \exists \lim_{n \rightarrow \infty} P^n$$

Teorema 3.6.2. Se P é uma matriz de transição regular de ordem n de uma cadeia de Markov então

$$P^n \rightarrow \begin{bmatrix} q_1 & q_1 & \cdots & q_1 \\ q_2 & q_2 & \cdots & q_2 \\ \vdots & \vdots & \ddots & \vdots \\ q_k & q_k & \cdots & q_k \end{bmatrix}$$

quando $n \rightarrow \infty$, onde q_i ($i = 1, 2, \dots$) são números positivos tais que $q_1 + q_2 + \dots + q_k = 1$.

Demonstração:

Considera-se que P é diagonalizável, sabe-se pelo teorema 3.6.1 que existe $Q = \lim_{n \rightarrow \infty} P^n$.

Pelo corolário 3.4.1, P^n é estocástica ($\forall n \in \mathbb{N}$). De acordo com a observação 3.3.1, tem-se que:

$$J \cdot Q = J \cdot \lim_{n \rightarrow \infty} P^n = \lim_{n \rightarrow \infty} J \cdot P^n = \lim_{n \rightarrow \infty} J = J$$

o que implica que Q é estocástica.

Sabe-se pelo teorema 3.5.4 que P^{n_1+w} possui um autovalor $\lambda = 1$ de multiplicidade geométrica igual a 1. Sendo P regular, pelo corolário 3.4.3 tem-se que P^{n_2+w} é positiva. Assim, se $n_0 = \max\{n_1, n_2\}$, então P^{n_0} é positiva, e seu autovalor $\lambda = 1$ tem multiplicidade geométrica igual a 1.

Observe que

$$P^{n_0} \cdot Q = P^{n_0} \lim_{n \rightarrow \infty} P^n = \lim_{n \rightarrow \infty} P^{n+n_0} = Q$$

De $P^{n_0} \cdot Q = Q$ pode-se concluir que:

- Q é positiva pelo teorema 3.3.1, já que P^{n_0} é positiva e Q é estocástica.
- As colunas de Q são linearmente dependentes, pois as colunas de Q são autovetores de P^{n_0} correspondentes ao autovalor $\lambda = 1$.

Portanto as colunas de Q tem a mesma soma e são múltiplas uma das outras. Logo as colunas são todas iguais. O fato de $q_i > 0$ segue de Q ser positiva.

3.7 Comportamento de $P^n v$ quando $n \rightarrow \infty$

Teorema 3.7.1. Seja

$$Q = \begin{bmatrix} q_1 & q_1 & \cdots & q_1 \\ q_2 & q_2 & \cdots & q_2 \\ \vdots & \vdots & \ddots & \vdots \\ q_k & q_k & \cdots & q_k \end{bmatrix} \quad \text{e} \quad q = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_k \end{bmatrix},$$

onde $Q = \lim_{n \rightarrow \infty} P^n$. Se P é uma matriz de transição regular e v é um vetor de probabilidade, então quando $n \rightarrow \infty$, $P^n v \rightarrow q$. O vetor q é estacionário e independe de n , possui todas as entradas positivas.

Demonstração:

Como

$$P \cdot Q = P \cdot \lim_{n \rightarrow \infty} P^n = \lim_{n \rightarrow \infty} P^{n+1} = Q$$

onde cada coluna de Q é um autovetor de P associado a $\lambda = 1$.

Então, o produto da matriz Q pelo vetor de probabilidade v é dado por:

$$Qv = \begin{bmatrix} q_1 & q_1 & \cdots & q_1 \\ q_2 & q_2 & \cdots & q_2 \\ \vdots & \vdots & \ddots & \vdots \\ q_k & q_k & \cdots & q_k \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{bmatrix} = \begin{bmatrix} q_1 v_1 + q_1 v_2 + \cdots + q_1 v_k \\ q_2 v_1 + q_2 v_2 + \cdots + q_2 v_k \\ \vdots \\ q_k v_1 + q_k v_2 + \cdots + q_k v_k \end{bmatrix} = (v_1 + v_2 + \cdots + v_k) \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_k \end{bmatrix} = (1)q = q$$

Esse teorema é de fato válido, pois do teorema 3.6.2 tem-se que quando $n \rightarrow \infty$, $P^n \rightarrow Q$. Então $P^n v \rightarrow Qv = q$.

Corolário 3.7.1. Segue diretamente do teorema que a convergência de uma sequência (v_n) de uma cadeia de Markov independe do vetor de probabilidade inicial.

3.8 Vetor de Probabilidade Estacionário

Teorema 3.8.1. O vetor de probabilidade estacionário q de uma cadeia de Markov que tem uma matriz de transição regular P é o único vetor de probabilidade que satisfaz a equação $Pq = q$.

Demonstração:

Considera-se a identidade matricial $P \cdot P^n = P^{n+1}$. Pelo teorema 3.6.2 tem-se $PQ = Q$. Qualquer uma das colunas dessa equação matricial dá $Pq = q$. Para provar a unicidade de q , suponha que r seja um outro vetor de probabilidade tal que $P \cdot r = r$. Então $P^n \cdot r = r$. Pelo teorema 3.7.1 $P^n \cdot r = q$. Resulta $r = q$.

Uma vez provado a unicidade de q , pode-se escrever o sistema linear $Pq = q$ como o sistema linear homogêneo

$$Pq = q \Rightarrow Pq - Iq = Iq - Iq \Rightarrow (P - I)q = 0$$

Assim, o Teorema 3.8.1 também pode ser expresso como:

$$(P - I)q = 0 \tag{3.2}$$

Nesse caso, a resolução pelo sistema homogêneo torna-se mais prática pois não é necessário o desenvolvimento da sequência (v_n) .

Todas as demonstrações e resultados mostrados nesse capítulo, podem ser encontrados com mais detalhes no trabalho de Silva (2005).

No próximo capítulo será mostrado, sobre uma estrutura de rede, uma aplicação dos resultados matemáticos apresentados ao longo desse trabalho.

4 Aplicação

A prova conceitual é realizada sobre um problema de ordenação de páginas da internet, onde a importância de uma página é medida em relação a quantidade de referências (links) que esta página recebe em toda rede. Esse problema então, pode ser modelado como um grafo dirigido (figura 1) que representa uma rede mundial de computadores da web. Nessa rede, os vértices representam as páginas da web e as arestas os links entre as páginas.

O modelo adotado na aplicação, pode ser interpretado como um passeio aleatório e um grafo da web. Por exemplo, um internauta começa a navegar na web a partir de uma página qualquer, e segue a navegação por um dos links da página atual escolhido uniformemente, depois de muito tempo nessa tarefa as páginas começam a repetir e o internauta entediado pára o processo e recomeça a partir de uma outra página.

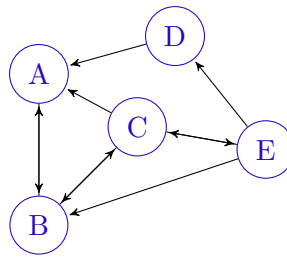


Figura 1 – Grafo representativo da web.

O conjunto $T = \{A, B, C, D, E\}$ representa o universo de páginas e a matriz de adjacência correspondente ao grafo dirigido tem a forma:

$$\mathcal{A} = \begin{matrix} & \begin{matrix} A & B & C & D & E \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix} & \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix} \end{matrix}$$

Observe pela matriz de adjacência do grafo que modela a rede que, os vértices tem os seguintes graus de saída (ver definição 2.1.8) e de entrada (ver definição 2.1.8):

Graus de saída: $d^+(A) = 1$; $d^+(B) = 2$; $d^+(C) = 3$; $d^+(D) = 1$; $d^+(E) = 3$

Graus de entrada: $d^-(A) = 3$; $d^-(B) = 3$; $d^-(C) = 2$; $d^-(D) = 1$; $d^-(E) = 1$

A matriz de adjacência permite calcular o número de passeios aleatórios em um grafo da web. Por exemplo, para calcular o número de passeios de comprimento 2, faz-se o cálculo de $\mathcal{A} \cdot \mathcal{A} = \mathcal{A}^2$. Cada elemento $a_{ij}^2 \in \mathcal{A}^2$ representa o número de passeios de comprimento 2 com início na página i e término na página j . Assim tem-se:

$$\mathcal{A}^2 = \begin{matrix} & \begin{matrix} A & B & C & D & E \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix} & \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 & 1 \\ 1 & \boxed{2} & 2 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 3 & 1 & 1 & 0 & 1 \end{bmatrix} \end{matrix}$$

O elemento $a_{CB}^2 = 2$ mostra que existem 2 passeios de comprimento 2. Esse passeio inicia no vértice C e termina no B . Generalizando, tem-se que $\mathcal{A}^n$ é a matriz dos elementos $a_{ij}^n \in \mathcal{A}^n$ de passeios de comprimento n .

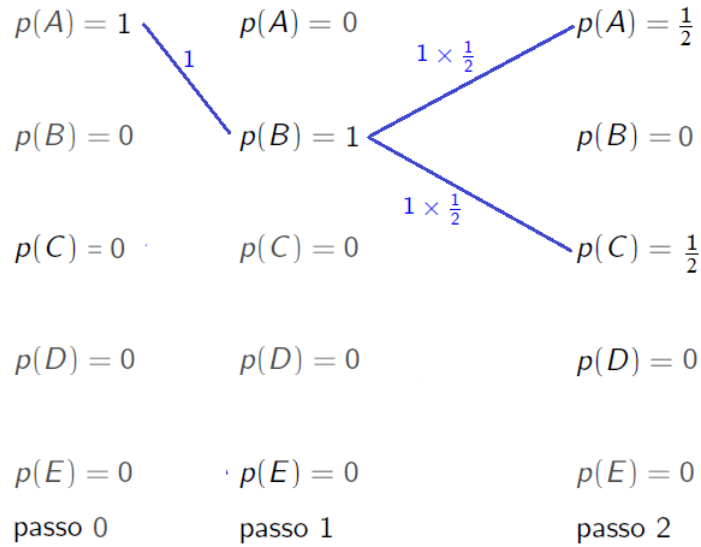
Pode-se obter as probabilidades de transição p_{ij} em cada passo de forma matricial, representada da seguinte forma para uma matriz de transição 5×5 :

$$P = \begin{bmatrix} p_{AA} & p_{AB} & p_{AC} & p_{AD} & p_{AE} \\ p_{BA} & p_{BB} & p_{BC} & p_{BD} & p_{BE} \\ p_{CA} & p_{CB} & p_{CC} & p_{CD} & p_{CE} \\ p_{DA} & p_{DB} & p_{DC} & p_{DD} & p_{DE} \\ p_{EA} & p_{EB} & p_{EC} & p_{ED} & p_{EE} \end{bmatrix}$$

Onde a probabilidade $p_{ij} = \frac{1}{d^+(j)}$, sendo $d^+(j)$ o número de links que partem de j .

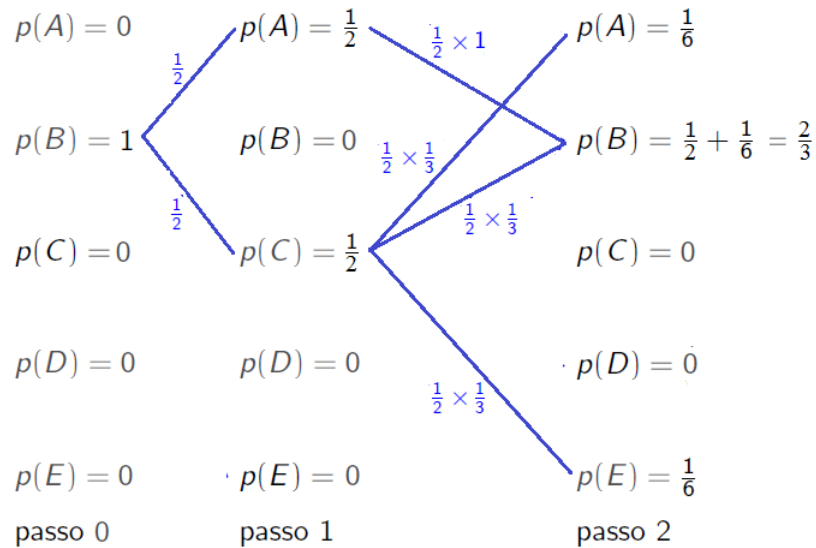
É possível mostrar o cálculo das probabilidades condicionais e a enumeração dos passeios aleatório no grafo da web a partir das árvores de probabilidades. Seguem os resultados a partir da construção das árvores probabilidade acessando as páginas da internet desde o passo 0 até o passo 2:

Cálculo das probabilidades para as variáveis aleatórias X_1 e X_2 dado que $X_0 = A$



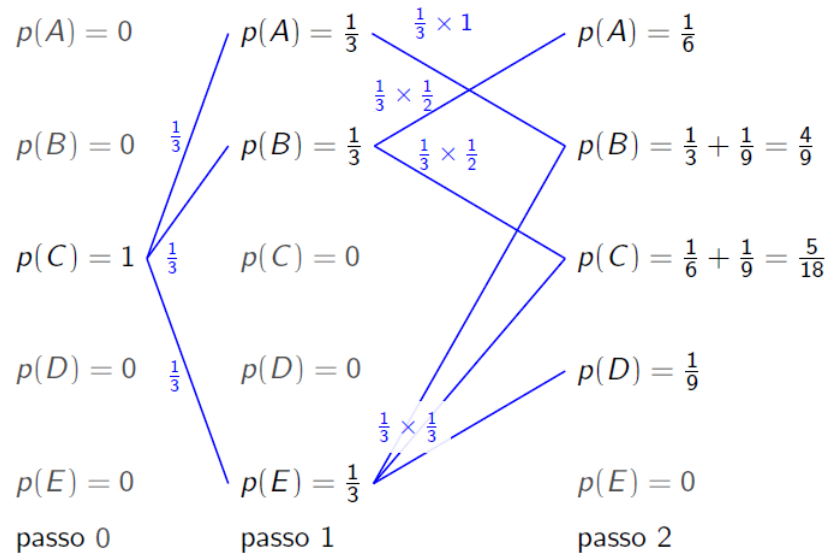
Através da árvore obtém-se o passeio $(A \rightarrow B \rightarrow A)$ e o passeio $(A \rightarrow B \rightarrow C)$, com o total de 2 passeios.

Cálculo das probabilidades para as variáveis aleatórias X_1 e X_2 dado que $X_0 = B$



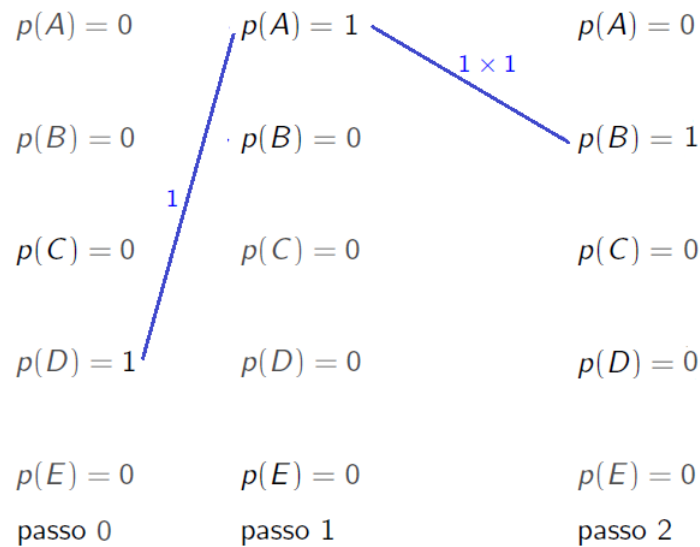
Através da árvore obtém-se o passeio $(B \rightarrow C \rightarrow A)$, os passeios $(B \rightarrow A \rightarrow B)$ ou $B \rightarrow C \rightarrow B)$ e o passeio $(B \rightarrow C \rightarrow E)$ com o total de 4 passeios.

Cálculo das probabilidades para as variáveis aleatórias X_1 e X_2 dado que $X_0 = C$



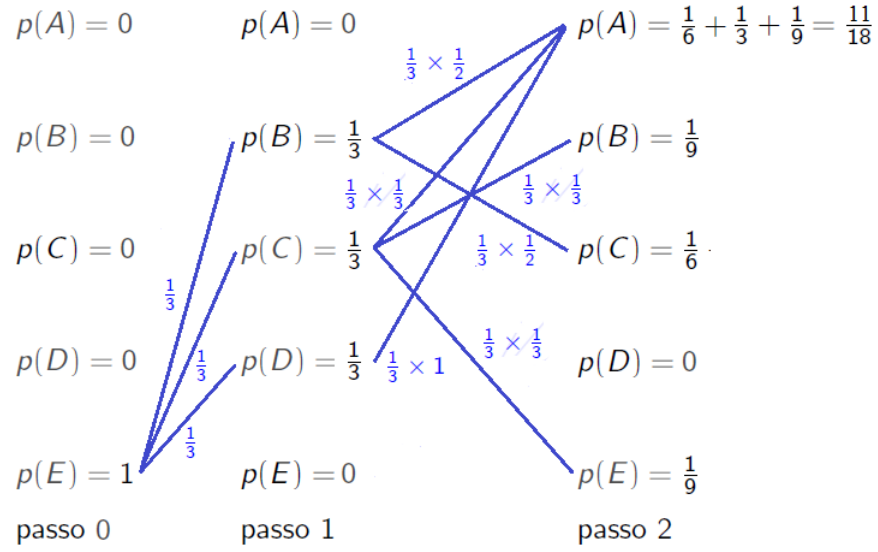
Através da árvore obtém-se o passeio $(C \rightarrow B \rightarrow A)$, os passeios $(C \rightarrow A \rightarrow B$ ou $C \rightarrow E \rightarrow B)$, os passeio $(C \rightarrow B \rightarrow C$ ou $C \rightarrow E \rightarrow C)$ e o passeio $(C \rightarrow E \rightarrow D)$ com o total de 6 passeios.

Cálculo das probabilidades para as variáveis aleatórias X_1 e X_2 dado que $X_0 = D$



Através da árvore obtém-se o passeio $(D \rightarrow A \rightarrow B)$.

Cálculo das probabilidades para as variáveis aleatórias X_1 e X_2 dado que $X_0 = E$



Através da árvore obtém-se o passeio ($E \rightarrow B \rightarrow A$, $E \rightarrow C \rightarrow A$ ou $E \rightarrow D \rightarrow A$), o passeio ($E \rightarrow C \rightarrow B$), o passeio ($E \rightarrow B \rightarrow C$) e o passeio ($E \rightarrow C \rightarrow E$) com o total de 6 passeios.

Observe que a matriz de transição P , representativa do passo 1, referente as árvores de probabilidade é:

$$P = \begin{matrix} & \begin{matrix} A & B & C & D & E \end{matrix} \\ \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 1 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & 0 \end{bmatrix} & \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix} \end{matrix}$$

Por exemplo, a probabilidade de transição da página C para a página B , após um passo é:

$$p_{BC}^1 = P(X_1 = B \mid X_0 = C) = p_{BC} = \frac{1}{3}.$$

E assim sucessivamente.

Observe que a matriz $P = [p_{ij}]$ é tal que $p_{ij} = 0$ se não houver link de j para i e $p_{ij} = \frac{1}{d^+(j)}$ se houver link de j para i , sendo $d^+(j)$ o número de links que partem de j . Por exemplo:

$$p_{DE}^1 = P(X_1 = D \mid X_0 = E) = p_{DE} = \frac{1}{d^+(E)} = \frac{1}{3}$$

Observe que a matriz de transição P^2 , representativa do passo 2, referente as

árvores de probabilidade é:

$$P^2 = P \cdot P = \begin{bmatrix} p_{AA} & p_{AB} & p_{AC} & p_{AD} & p_{AE} \\ p_{BA} & p_{BB} & p_{BC} & p_{BD} & p_{BE} \\ p_{CA} & p_{CB} & p_{CC} & p_{CD} & p_{CE} \\ p_{DA} & p_{DB} & p_{DC} & p_{DD} & p_{DE} \\ p_{EA} & p_{EB} & p_{EC} & p_{ED} & p_{EE} \end{bmatrix} \cdot \begin{bmatrix} p_{AA} & p_{AB} & p_{AC} & p_{AD} & p_{AE} \\ p_{BA} & p_{BB} & p_{BC} & p_{BD} & p_{BE} \\ p_{CA} & p_{CB} & p_{CC} & p_{CD} & p_{CE} \\ p_{DA} & p_{DB} & p_{DC} & p_{DD} & p_{DE} \\ p_{EA} & p_{EB} & p_{EC} & p_{ED} & p_{EE} \end{bmatrix}$$

$$P^2 = P \cdot P = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 1 & 0 \\ \boxed{1} & \boxed{0} & \boxed{\frac{1}{3}} & \boxed{0} & \boxed{\frac{1}{3}} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & \frac{1}{2} & \boxed{\frac{1}{3}} & 1 & 0 \\ 1 & 0 & \boxed{\frac{1}{3}} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & \boxed{0} & 0 & \frac{1}{3} \\ 0 & 0 & \boxed{0} & 0 & \frac{1}{3} \\ 0 & 0 & \boxed{\frac{1}{3}} & 0 & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & 0 & \frac{11}{18} \\ 0 & \frac{2}{3} & \boxed{\frac{4}{9}} & 1 & \frac{1}{9} \\ \frac{1}{2} & 0 & \frac{5}{18} & 0 & \frac{1}{6} \\ 0 & 0 & \frac{1}{9} & 0 & 0 \\ 0 & \frac{1}{6} & 0 & 0 & \frac{1}{9} \end{bmatrix}$$

Por exemplo, para encontrarmos a probabilidade de transição da página C para a página B em 2 dois passo é:

$$\begin{aligned} p_{BC}^2 &= P(X_2 = B \mid X_0 = C) = p_{BA} \cdot p_{AC} + p_{BB} \cdot p_{BC} + p_{BC} \cdot p_{CC} + p_{BD} \cdot p_{DC} + p_{BE} \cdot p_{EC} \\ &= 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + \frac{1}{3} \cdot 0 + 0 \cdot 0 + \frac{1}{3} \cdot \frac{1}{3} = \frac{4}{9} \end{aligned}$$

Os passeios são $(C \rightarrow A \rightarrow B$ ou $C \rightarrow E \rightarrow B)$.

Observação 4.0.1. A probabilidade $p_{BC}^2 = \frac{4}{9}$ corresponde ao elemento $a_{32}^2 = a_{CB}^2 = 2$ na matriz de passeio de comprimento 2.

Assim pode-se concluir que o enésimo-passo da matriz de transição P é a enésima potência de P .

Nas próximas seções são mostradas duas formas de cálculo PageRank. Primeiramente utilizando o Método das Potências descrito no teorema 3.4.1 e posteriormente utilizando o vetor de probabilidade estacionário (q).

4.1 Cálculo do PageRank utilizando o Método das Potências, caso Markoviano

Embora o número de páginas da Web seja imenso, ainda assim é finito. Quando a estrutura de links da rede mundial de computadores (World Wide Web) é modelada por uma Cadeia de Markov, cada página na rede é um estado dessa cadeia. Essa subseção mostra uma aplicação do estudo das Cadeias de Markov na estrutura dos links da Web,

focando nas cadeias que estacionam após um número finito de iterações partindo de um estado inicial (“página”).

A figura 2 representa um grafo quando uma página da internet é acessada.

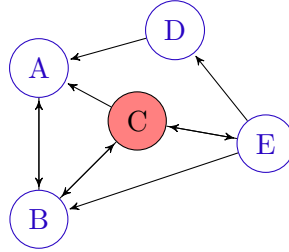


Figura 2 – Grafo representativo quando a página C é acessada.

Cuja matriz de transição P é dada por:

$$P = \begin{array}{ccccc} & A & B & C & D & E \\ \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 1 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & 0 \end{bmatrix} & A \\ & B \\ & C \\ & D \\ & E \end{array}$$

As Cadeias de Markov calculam o PageRank utilizando o processo iterativo descrito no teorema 3.4.1.

A página C é escolhida aleatoriamente para iniciar a navegação do vetor de estado inicial v_0 .

$$v_0 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\text{Início: } v_0 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\text{Passo 1: } v_1 = Pv_0 = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 1 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ 0 \\ 0 \\ \frac{1}{3} \end{bmatrix}$$

$$\text{Passo 2: } v_2 = Pv_1 = P^2v_0 = \begin{bmatrix} \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & 0 & \frac{11}{58} \\ 0 & \frac{2}{3} & \frac{4}{9} & 1 & \frac{1}{9} \\ \frac{1}{2} & 0 & \frac{5}{18} & 0 & \frac{1}{6} \\ 0 & 0 & \frac{1}{9} & 0 & 0 \\ 0 & \frac{1}{6} & 0 & 0 & \frac{1}{9} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{6} \\ \frac{4}{9} \\ \frac{5}{18} \\ \frac{1}{9} \\ 0 \end{bmatrix}$$

Utilizando um software livre para os cálculos, obtem-se

$$v_{20} = Pv_{19} = P^{20}v_0 = \begin{bmatrix} 10873758553 \\ 37192366944 \\ 21798409087 \\ 55788550416 \\ 12233707681 \\ 55788550416 \\ 2724497393 \\ 111577100832 \\ 25207081 \\ 344373768 \end{bmatrix} \approx \begin{bmatrix} 0.29236532779353 \\ 0.39073266690844 \\ 0.21928706857906 \\ 0.02441806941284 \\ 0.07319686730610 \end{bmatrix}$$

A convergência do método é observada a partir de $n > 20$. Para $n > 20$, tem-se que todos os vetores estado são aproximadamente iguais a v_{20} . Assim os vetores convergem a um vetor fixo (vetor estado estacionário) à medida que n cresce.

As coordenadas do vetor v_{20} são a pontuação PageRank das páginas. A importância de uma página na rede é medida pelo tamanho relativo da coordenada correspondente no Vetor Estado Estacionário.

Logo, a página mais importante de acordo é a página B , que corresponde à maior coordenada de v_{20} . A ordenação completa é B ; A ; C ; E e D .

4.2 Cálculo do PageRank através do Vetor de Probabilidade Estacionário (q)

Viu-se que uma rede mundial de computadores pode ser modelada como um grafo direcionado. Como a matriz de transição P para essa Cadeia de Markov é regular, o Teorema 3.6.2 garante que existe um Vetor Estado Estacionário q para a cadeia, onde as coordenadas de q são a pontuação PageRank das páginas.

A matriz P é matriz de transição, logo pelo teorema de Perron-Forbenius caso Markoviano (teorema 3.5.5) ela possui um autovalor igual a 1, com multiplicidade algébrica igual a 1. Então pode-se utilizar a equação 3.2 do teorema 3.8.1, na qual q é o

autovetor associado ao autovalor $\lambda = 1$.

Calculo do PageRank através do vetor de probabilidade estacionário (q):

Resolvendo o sistema linear $(P - I)q = 0$ encontra-se o vetor de probabilidade estacionário (q) que é o vetor PageRank do grafo da web.

$$\text{A matriz de transição é } P = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 1 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & 0 \end{bmatrix}$$

$$\text{Considerando } q = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ q_5 \end{bmatrix}$$

de modo que o sistema linear $(P - I)q = 0$ é

$$\begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{3} & -1 & 0 \\ -1 & 1 & -\frac{1}{3} & 0 & -\frac{1}{3} \\ 0 & -\frac{1}{2} & 1 & 0 & -\frac{1}{3} \\ 0 & 0 & 0 & 1 & \frac{1}{3} \\ 0 & 0 & -\frac{1}{3} & 0 & 1 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ q_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

A matriz dos coeficientes na forma escalonada é

$$\begin{bmatrix} 1 & 0 & 0 & 0 & -4 \\ 0 & 1 & 0 & 0 & -\frac{16}{3} \\ 0 & 0 & 1 & 0 & -3 \\ 0 & 0 & 0 & 1 & -\frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

de modo que o sistema linear original é equivalente ao sistema

$$q_1 = 4q_5$$

$$q_2 = \frac{16}{3}q_5$$

$$q_3 = 3q_5$$

$$q_4 = \frac{1}{3}q_5$$

qualquer solução do sistema linear é da forma

$$q = q_5 \begin{bmatrix} 4 \\ \frac{16}{3} \\ 3 \\ \frac{1}{3} \\ 1 \end{bmatrix}$$

Para fazer disso um vetor de probabilidade, coloca-se

$$q_5 \left(4 + \frac{16}{3} + 3 + \frac{1}{3} + 1 \right) = 1 \Rightarrow q_5 = \frac{1}{4 + \frac{16}{3} + 3 + \frac{1}{3} + 1} = \frac{3}{41}$$

Assim, o vetor de estado estacionário que representa o vetor PageRank do grafo da web é

$$q = \frac{3}{41} \begin{bmatrix} 4 \\ \frac{16}{3} \\ 3 \\ \frac{1}{3} \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{12}{41} \\ \frac{16}{41} \\ \frac{9}{41} \\ \frac{1}{41} \\ \frac{3}{41} \end{bmatrix} \cong \begin{bmatrix} 0.292682926 \\ 0.390243902 \\ 0.219512195 \\ 0.024390243 \\ 0.073170731 \end{bmatrix}$$

As coordenadas do vetor q são a pontuação PageRank das páginas. A importância de uma página na rede é medida pelo tamanho relativo da coordenada correspondente no Vetor Estado Estacionário.

Logo, a página mais importante de acordo com o PageRank é a página B , que corresponde à maior coordenada de q . A ordenação completa é B ; A ; C ; E e D .

Finalmente, cabe salientar que, quando uma rede é fortemente conexa, isto é, quando é possível passar de uma página arbitrária para outra qualquer apenas clicando nos links, o conjunto de soluções do sistema linear homogêneo tem dimensão 1.

5 Conclusão

O trabalho é importante na medida em que desperta a pesquisa e o conhecimento sobre teorias matemáticas de Grafos, Álgebra Linear e Cadeias de Markov. Nesse sentido, a identificação de uma Cadeia Markov sobre um passeio aleatório, em um grafo fortemente conexo, permitiu a resolução do problema de ordenação de páginas da internet. A partir de uma prova conceitual foi possível resolver o mesmo problema de rede com metodologias diferentes. Sob o ponto de vista computacional, é importante avaliar as duas metodologias utilizadas na resolução de um problema do mundo real. Dependendo do tamanho da rede, o custo computacional para obter o vetor estacionário pode apresentar diferenças significativas em relação as duas metodologias apresentadas. Em relação a trabalhos futuros seria interessante um estudo sobre o modelo de citação da literatura acadêmica aplicado a web, que geralmente considera a contagem de citações para uma determinada página, onde o grande desafio podem ser os fatores que influenciam na velocidade de convergência do método utilizado.

Referências

- ANTON, H.; RORRES, C. *Álgebra linear com aplicações*. 10. ed. Porto Alegre: Bookman, 2012. Citado na página 16.
- BOVOLONI, J. O. et al. Page rank: O funcionamento da ferramenta de busca do google. *Ciencias da Computação cadernos de graduação ciencias exatas e tecnológicas*, Aracaju, 2016. Citado na página 9.
- FREITAS, C. R. *Teoria Espectral Aplicada a Problemas de Localização*. Dissertação (Mestrado) — Programa de Pós Graduação em Modelagem Computacional da Universidade Federal do Rio Grande, Rio Grande, abril 2012. Citado 2 vezes nas páginas 15 e 17.
- JIANYA, Z. *Modelagem de Inuflência de Sócios das Redes Sociais pelos PageRank e Índice W-Entropia*. Dissertação (Mestrado) — Departamento de Ciência da Computação - UnB, Brasília, março 2012. Disponível em: <http://repositorio.unb.br/bitstream/10482/11061/1/2012_ZhengJianya.pdf>. Acesso em: 30.04.2018. Citado na página 9.
- LANGVILLE, A. N.; MEYER, C. D. *Google's PageRank and Beyond: The Science of Search Engine Ranking*. Princeton and Oxford: Princeton University Press E-book, 2011. Citado na página 9.
- MAGALHÃES, T. M. de. O motor de busca google e o algoritmo pagerank. *Revista Eletrônica - Faculdade Santos Dumont*, 2014. Disponível em: <<http://fsd.edu.br/revistaeletronica/arquivos/6Edicao/artigo35%20TERESINHA.pdf>>. Citado na página 9.
- MELO, M. P. de. *Ordenação das páginas do Google - "Page Rank"*. Dissertação (Mestrado) — Instituto de Matemática e Estatística - USP, São Paulo, maio 2009. Disponível em: <<http://www.teses.usp.br/teses/disponiveis/45/45133/tde-08052009-152811/pt-br.php>>. Acesso em: 30.04.2018. Citado na página 9.
- OLIVEIRA, J. C. F. de. *Noções de grafos dirigidos, cadeias de Markov e as buscas do Google*. Dissertação (Mestrado) — Departamento de Matemática - UFS, São Cristovão, agosto 2014. Disponível em: <https://sca.profmatt-sbm.org.br/sca_v2/get_tcc3.php?id=817>. Acesso em: 02.01.2018. Citado na página 9.
- POOLE, D. *Álgebra Linear*. São Paulo: Thomson, 2004. Citado na página 17.
- REIS, T. *Algoritmo Rastreador Web Especialista Nuclear*. Dissertação (Mestrado) — Instituto de Pesquisas Energéticas e Nucleares. Autarquia Associada à Universidade de São Paulo, São Paulo, novembro 2013. Disponível em: <<http://www.teses.usp.br/teses/disponiveis/85/85133/tde-07012014-134548/pt-br.php>>. Acesso em: 30.04.2018. Citado na página 9.
- SILVA, W. Q. da. *Cadeias de Markov: Interpretação e Método de Solução*. Dissertação (Mestrado) — Especialização em Ensino de Matemática - UFF, Niterói, dezembro 2005. Disponível em: <<https://d577cf68-a-62cb3a1a-s-sites.googlegroups.com/site/>>

[wellersonquintaneiro/publicacoes-e-trabalhos-de-conclusao-de-curso/QUINTANEIRO%20CW.CAdeiasdeMarkov.pdf](#)>. Acesso em: 23.04.2018. Citado na página 23.