Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Bacharelado em Ciências de Computação

Luiz Fernando Rabelo (11796893)

ANÁLISE DE MÉTODOS NUMÉRICOS PARA ENCONTRAR DISTRIBUIÇÕES ESTACIONÁRIAS NO ALGORITMO GOOGLE PAGERANK

Relatório Final do Trabalho Prático da Disciplina SME0300 - Cálculo Numérico, oferecida no primeiro semestre de 2023, sob orientação da professora Dra. Lívia Souza Freire Grion.

São Carlos
Julho / 2023

1 INTRODUÇÃO

Desenvolvido por Larry Page e Sergey Brin, os fundadores do Google, o algoritmo PageRank é um componente primordial do mecanismo de buscas na Internet que revolucionou a forma como as páginas da web são avaliadas e influenciou significativamente o campo da otimização de mecanismos de busca (SEO).

O objetivo algoritmo é fornecer aos usuários resultados de pesquisa mais relevantes, levando em consideração a qualidade e a popularidade das páginas. Ao contrário de abordagens anteriores, que se baseavam principalmente em critérios superficiais, como palavras-chave, o PageRank introduziu um novo conceito: a importância dos links.

O algoritmo considera a Web como um imenso grafo de páginas interconectadas. Cada página é tratada como um nó no grafo, e os links entre as páginas são representados por arestas direcionadas. O PageRank parte do pressuposto de que uma página é mais importante quanto mais links de outras páginas importantes ela recebe. Essa lógica se baseia na ideia de que, se muitas páginas relevantes estão apontando para uma página específica, ela deve ser valiosa e merecedora de uma classificação mais alta.

Nesse sentido, o grafo representante da Web é modelado como uma Cadeia de Markov e o algoritmo procura encontrar a distribuição estacionária da cadeia, a qual produzirá as probabilidades de longo prazo de um usuário estar em cada página. Essas probabilidades de longo prazo são interpretadas como a importância relativa de cada página na Web, pois páginas com maior pontuação de PageRank terão uma probabilidade maior de serem alcançadas por um usuário aleatório.

Considerando P como a matriz de adjacências (ponderada com probabilidades de transição entre páginas conectadas na Web), α como a probabilidade de um usuário mover-se aleatoriamente para qualquer página (sem necessariamente existir uma ligação da página de destino com a página de origem), N o número de páginas da Web e U como uma matriz de 1's $N \times N$, a Matriz Google G, representante de uma Cadeia de Markov ergódica, pode ser definida como:

$$G = (1 - \alpha)P + \frac{\alpha}{N} U$$

Para encontrar a distribuição estacionária da matriz G, podem ser utilizados diferentes métodos numéricos, como a Estimação de Monte Carlo, que obtém uma solução aproximada para a distribuição, além da Decomposição QR e do Método das Potências, sobre os quais a distribuição é extraída de um ponto de vista algébrico-analítico. Todos esses métodos citados foram explorados ao longo do presente trabalho.

2 METODOLOGIA

Para os testes dos cálculos das distribuições estacionárias, foram considerados dois grafos:

1. *Rede fictícia com apenas 6 nós:* uma rede sem conexão com a realidade que simula relacionamentos em uma rede com poucas páginas. A escolha de uma rede pequena se fez necessária para visualização dos resultados encontrados e para a avaliação da corretude dos algoritmos. A representação visual da rede considerada é dada por:

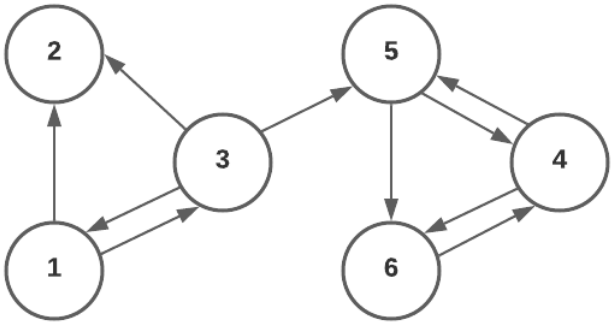

Figura 1 – Representação Visual da Rede Simplificada

2. *Rede abstraída do "Free On-Line Dictionary of Computing (FOLDOC)":* uma rede de referências cruzadas entre 13.356 conceitos definidos no dicionário FOLDOC. Cada nó representa um termo e cada aresta direcionada de um termo A para um termo B com peso m, representa que o termo B é utilizado m vezes na definição do termo A. Embora a semântica da rede não represente propriamente páginas da Web, seu uso é interessante para avaliação de como os métodos aplicados se comportam em redes maiores. A aplicação do PageRank, nesse caso, pode ser interpretada como a ordem de importância dos conceitos definidos no dicionário.

Para determinar a distribuição ambas as redes, foram utilizados os seguintes métodos:

- *Simulação de Monte Carlo:* o método pode ser utilizado para estimar a distribuição estacionária de uma Cadeia de Markov através da simulação estocástica. O método baseia-se na ideia de que, ao se transitar na Cadeia de Markov por um número suficientemente grande de passos, a distribuição das amostras convergirá para a distribuição estacionária. Nesse sentido, a partir de um estado inicial, é gerada uma sequência de visitação de estados na cadeia, em que as transições entre os estados ocorrem de acordo com a matriz de

probabilidade de transição. O processo é repetido por um número suficientemente grande de iterações. Ao final, contam-se as ocorrências de visitação dos estados e normaliza-se o resultado, o qual representará uma estimativa da probabilidade de transição de cada estado na distribuição estacionária.

- *Decomposição QR:* O Método QR é um algoritmo que pode ser utilizado para encontrar os autovetores de uma matriz. Como, para a distribuição estacionária π_∞, temos que $\pi_\infty G = \pi_\infty$, podemos interpretar a equação sob a ótica de uma Transformação Linear: a distribuição estacionária está diretamente relacionada à determinação do autovetor π associado ao autovalor 1. Assim, empregando o método QR com os devidos ajustes (transposição da matriz G e consideração dos vetores em colunas) e calculando todos os autovetores, é possível selecionar tal autovetor relacionado ao autovalor 1 que, quando normalizado, corresponde à distribuição estacionária da matriz G.

- *Método das Potências:* O Método das Potências utiliza a mesma abordagem da Decomposição QR para a determinação da distribuição estacionária: encontrar o autovetor associado ao autovalor 1 e normalizá-lo. Entretanto, ao contrário da Decomposição QR, que calcula todos os autovetores, o Método das Potências é mais direto, uma vez que ele determina apenas o autovetor associado ao autovalor dominante em uma Transformação Linear. E no escopo da matriz G, tal autovetor é justamente aquele correspondente ao autovalor unitário, bastando normalizá-lo.

3 RESULTADOS

Aplicando os métodos supracitados nos grafos considerados, foram obtidos os seguintes tempos de execução (média de 10 experimentos):

Monte Carlo	QR	Potências
0.29924	6.22258	0.00107

Tabela 1 - Tempos de Execução (s) para Rede Simplificada

Monte Carlo	QR	Potências
5.59789	> 3600	6.67728

Tabela 2 - Tempos de Execução (s) para Rede FOLDOC

4 CONCLUSÕES

Ao comparar os resultados obtidos para as duas cadeias de Markov, podemos tirar algumas conclusões relacionadas à complexidade dos métodos e determinar o melhor algoritmo

para cada caso. Para a primeira rede, com 6 nós, por a matriz gerada ser relativamente pequena, os tempos de execução são razoavelmente curtos para todos os métodos. O Método das Potências se destaca como o algoritmo mais eficiente, levando, na média, apenas uma fração de segundo para encontrar a distribuição estacionária. A Decomposição QR, por realizar cálculos além do necessário, para o escopo do problema, é o método mais lento, mas ainda assim aceitável para uma cadeia de tamanho pequeno. Para a rede FOLDOC, a matriz gerada é consideravelmente maior e mais complexa em relação à primeira. Podemos ver que a Simulação de Monte Carlo obteve o melhor tempo, seguido do Método das Potências. A Decomposição QR foi a mais lenta, novamente, com execuções que tomaram mais de 1 hora.

Com base nesses resultados, podemos concluir que a escolha do melhor algoritmo depende do tamanho e da complexidade da cadeia da Matriz Google. O método da Decomposição QR, embora seja capaz de encontrar a distribuição estacionária de maneira exata, possui baixa viabilidade de utilização no cenário considerado, pelos cálculos dos demais autovetores desnecessários. A escolha entre o Método das Potências e a Simulação de Monte Carlo depende da precisão requerida e do tamanho da cadeia. Para cadeias menores, o Método das Potências é o mais indicado, pois oferece uma solução rápida e precisa. Já para cadeias maiores, o Método das Potências, mesmo sendo mais rápido que a Decomposição QR, possui uma complexidade de tempo considerável e, para matrizes de ordem grande, o Monte Carlo pode ser uma opção viável em cenários que não necessitam de uma precisão tão acurada. Por fim, vale apontar que a proposição original do algoritmo PageRank adotou o Método das Potências como solução padrão, em virtude da priorização da precisão no resultado.

5 REFERÊNCIAS

- Quarteroni, Alfio. (2006). **Scientific Computing with MATLAB and Octave**.
- Page, L. (1999). **The PageRank citation ranking: Bringing order to the web. Stanford InfoLab**.
- Rodrigues, F. A. (2020). **Cadeias de Markov: O Algoritmo PageRank**. Disponível em https://www.youtube.com/watch?v=iwraRlbubeQ