
PROJETO 3

Marcelo Henrique Morello Manzo
cc23326@g.unicamp.br

Vitor Henrique Gírio Paes
cc23340@g.unicamp.br

1 Introdução

O algoritmo PageRank é utilizado para avaliar a relevância de páginas na internet, com base nas ligações entre elas. Criado por Larry Page e Sergey Brin, fundadores do Google, esse método proporciona uma classificação mais eficiente em comparação com contagens simples de links. No projeto, foram implementadas duas abordagens para o cálculo do PageRank: o Modelo do Navegador Aleatório e o Algoritmo Iterativo.

2 Objetivos

O principal objetivo deste projeto é implementar e comparar duas abordagens do algoritmo PageRank. Pretendemos:

- Compreender o funcionamento do algoritmo PageRank e suas aplicações.
- Implementar o Modelo do Navegador Aleatório e o Algoritmo Iterativo em Python.
- Analisar e comparar a eficiência de ambos os métodos na determinação do PageRank.

3 Descrição

O PageRank estabelece a relevância de uma página web com base na quantidade e na qualidade dos links que a citam. Este sistema foi concebido para reduzir a fraude e manipulação, pois, ao invés de simplesmente contabilizar links, também considera a autoridade dos sites que os utilizam. Para compreender o funcionamento deste algoritmo, pense em um indivíduo navegando na internet clicando em links aleatórios - eles não seguirão a mesma ordem de cliques em todas as situações. Depois de alguns links, as chances sugerem que o usuário se encontra em um ciclo sem fim. Para prevenir este cenário, a chance de o usuário acessar uma página da internet aleatória é estabelecida como uma opção, diversificando suas interações.

As duas principais abordagens para calcular o PageRank são:

- **Modelo do Navegador Aleatório:** Um navegador teórico segue links de forma aleatória. A probabilidade de estar em uma página específica reflete seu PageRank. Esse modelo pode ser descrito por meio de uma Cadeia de Markov.
- **Algoritmo Iterativo:** Nesta abordagem, as páginas são iteradas, ajustando continuamente suas pontuações. Inicialmente, considera-se que todas as páginas têm o mesmo PageRank. Em cada iteração, os valores são atualizados com base nas pontuações das páginas que fazem referência a ela.

A fórmula para a iteração é expressa como:

$$PR(p) = \frac{1-d}{N} + d \sum_{i \in M(p)} \frac{PR(i)}{L(i)}$$

Onde:

- $PR(p)$ representa o valor do PageRank da página p .
- d é o fator de amortecimento (geralmente fixado em 0.85).
- N é o número total de páginas no conjunto.
- $M(p)$ é o conjunto de páginas que direcionam para a página p .
- $PR(i)$ é o valor do PageRank da página i que aponta para p .
- $L(i)$ é o total de links que saem da página i .

4 Estrutura do Projeto

O projeto é organizado em três partes principais:

1. **Conjunto de Corpus:** Representa a rede de links que será analisada, onde cada um dos três corpus contém páginas HTML.
2. **Implementações em Python dos algoritmos de PageRank:** Inclui a execução dos métodos de amostragem e iterativo para calcular o PageRank.
3. **Resultados e Análise:** Compara os resultados obtidos pelos diferentes métodos, discutindo sua precisão e convergência.

4.1 Implementação dos Algoritmos em Python

4.1.1 Modelo do Navegador Aleatório

O **Modelo do Navegador Aleatório** simula um navegador que, ao acessar uma página, segue um link aleatoriamente com uma probabilidade d ou escolhe outra página da rede com probabilidade $1 - d$.

4.1.2 Modelo do Algoritmo Iterativo

A função a seguir foi criada para aplicar o **modelo iterativo** em Python. Essa função calcula os valores de PageRank de acordo com a fórmula anteriormente citada.

5 Resultados

Os resultados foram obtidos após a execução dos dois métodos de cálculo do PageRank. A seguir, apresentamos uma tabela comparativa entre os valores de PageRank obtidos por ambos os métodos para os quatro campi, cada um contendo quatro páginas HTML.

5.1 Resultados do Corpus 0

Página	Modelo do Navegador Aleatório	Algoritmo Iterativo
Página A1	21%	22%
Página A2	39%	42%
Página A3	22%	22%
Página A4	16%	13%

5.2 Resultados do Corpus 1

Página	Modelo do Navegador Aleatório	Algoritmo Iterativo
bfs	13%	11%
dfs	10%	08%
games	18%	22%
minesweeper	11%	11%
minimax	14%	13%
search	19%	20%
tictactoe	12%	11%

5.3 Resultados do Corpus 2

Página	Modelo do Navegador Aleatório	Algoritmo Iterativo
ai	15%	10%
algorithms	11%	12%
c	12%	12%
interference	14%	02%
logic	06%	22%
programming	18%	12%
python	11%	12%
recursion	09%	07%

6 Conclusão

A implementação do PageRank com o Modelo do Navegador Aleatório e o Algoritmo Iterativo demonstrou a eficácia de ambos os métodos. Apesar de os resultados finais serem semelhantes, variam em termos de tempo de execução e precisão, com o modelo iterativo mostrando maior estabilidade conforme o número de iterações aumenta. Este estudo denota a relevância do algoritmo na análise de redes e sua aplicação em sistemas práticos, como motores de busca.