Projeto 3

Christopher Alexandre
cc23322@g.unicamp.br

João Felipe Coromberk
cc23319@g.unicamp.br

1 Introdução

Atualmente, os mecanismos de busca utilizam métodos e algoritmos para exibir páginas com diferentes níveis de relevância. Plataformas como o Google não podem apresentar no topo resultados de páginas que são desconhecidas, não confiáveis ou pouco visitadas. Para isso, esses sites utilizam algoritmos que classificam páginas, certificando-se de que as mais relevantes tenham prioridade.

Existem vários algoritmos que atendem à necessidade de classificar sites nos mecanismos de busca. Alguns utilizam palavras-chave para determinar a relevância; outros priorizam o número de visitantes. Além disso, há algoritmos que classificam os sites com base na quantidade de links que direcionam os usuários para eles. Um exemplo é o PageRank, criado e utilizado pelo Google, que se baseia na quantidade e qualidade dos links que apontam para uma página.

2 Pagerank

Criado pelo Google, o **PageRank** é um algoritmo que mede a importância de uma página da web com base na quantidade e na qualidade dos links que apontam para ela. A eficácia

do PageRank reside no fato de que ele considera não apenas a quantidade de links recebidos, mas também a **qualidade** desses links. Se uma página recebe muitas referências de sites com baixo PageRank, seu próprio PageRank não será elevado. Por outro lado, se as referências vierem de sites importantes, com alto PageRank, a página em questão será considerada mais relevante e terá um PageRank mais elevado.

Para calcular o PageRank, utiliza-se o modelo do **surfista aleatório**, que simula o comportamento de um usuário navegando aleatoriamente pela web. O surfista percorre os links de uma página para outra, mas, em determinados momentos, pode optar por parar de seguir os links e "teletransportar-se"diretamente para uma nova página aleatória. Esse comportamento é controlado por um parâmetro chamado **damping factor** (fator de amortecimento), que é um valor entre 0 e 1, geralmente definido como 0,85. Isso significa que, em 85% das vezes, o surfista segue um link da página atual, enquanto nos 15% restantes, ele se desloca aleatoriamente para qualquer outra página do conjunto.

Podemos modelar o comportamento do surfista aleatório usando uma **Cadeia de Markov**, onde as páginas são representadas como estados, e as transições entre elas correspondem às mudanças de página. O PageRank de uma página P é calculado pela seguinte fórmula:

$$PR(P) = (1 - d) + d \sum_{i=1}^{N} \frac{PR(P_i)}{C(P_i)} \tag{1}$$

Onde:

- $PR(P)$ é o PageRank da página P,
- d é o damping factor,
- P_i são as páginas que linkam para P,
- $C(P_i)$ é o número de links saindo da página P_i,

- e N é o número total de páginas que linkam para P.

3 Cadeia de Markov

A cadeia de markov é um processo probabilistico aleatorio onde se faz previsões do proximo estado de acordo apenas com o estado atual

4 Descrição do problema

Nesta seção, apresentaremos os problemas propostos no Projeto 3 da disciplina TI327 - Tópicos em Inteligência Artificial, ministrada pelo Prof. Dr. Guilherme Macedo.

Foi dado pelo professor três diretórios corpus com arquivos HTML interligados. O objetivo do projeto é criar um algoritmo que atribua um PageRank a cada página, utilizando o modelo do surfista aleatório. Foi feito então um algoritmo com um número de amostragens que o surfista percorrerá, adicionara um ao PageRank desta pagina sempre que o surfista acabar nela e calcula a probabilidade de chegar nos outros sites de acordo com o site atual, essa conta é feito com o auxilio do damping factor. Ficando então a probabilidade para todos os sites como:

$$\frac{(1-d)}{n}$$

E para os sites que recebem links:

$$\frac{(1-d)}{n} + \frac{d}{L(Pi)}$$

onde Pi é a pagina atual do surfista, $L(Pi)$ é o número de links que a pagina atual referencia, d é o damping factor e n o numero total de paginas no corpus.

Caso não exista nenhum link na pagina atual a probabilidade de todos os sites se torna:

$$\frac{1}{n}$$

Também foi feito o algoritmo iterativo, onde, em vez de simular o comportamento do surfista com amostragens, o PageRank de cada página é atualizado em cada iteração até que o valor converja. Esse processo baseia-se na seguinte fórmula de atualização:

$$PR(P) = (1 - d) + d \sum_{i=1}^{N} \frac{PR(P_i)}{C(P_i)} \tag{2}$$

Este algoritmo iterativo continua atualizando os valores de PageRank até que a diferença entre os valores da iteração anterior e da atual seja insignificante, garantindo que o PageRank de cada página esteja corretamente calculado.

5 Experimentos computacionais

Todos os experimentos computacionais foram realizados em máquinas distintas.

- **Notebook**
  - **Processador:** Intel Celeron 1.80GHz
  - **Memória RAM:** 12 GB
- **Notebook**
  - **Processador:** Intel Core i5 1.00GHz
  - **Memória RAM:** 8 GB

- **Desktop**
  - **Processador:** AMD Ryzen 5 5500 3.6GHz
  - **Memória RAM:** RAM: 32 GB

Os experimentos foram desenvolvidos com Python 3.12.2.

6 Resultados

6.1 Resultados de corpus0

Nesta seção, aparesentamos os resultados obtidos da aplicação do métodos em um conjunto de páginas web da pasta corpus1.

Página	PageRank Iterativo	PageRank Amostral
1.html	0.2230	0.2202
2.html	0.4297	0.4289
3.html	0.2167	0.2202
4.html	0.1306	0.1307

Tabela 1: Resultados de PageRank no corpus0

6.1.1 Análise dos Resultados

A página 2.html apresenta o maior valor de PageRank, com 0.4297 no método iterativo e 0.4289 no método amostral. Isso sugere que esta página é a mais relevante dentro do corpus0, possivelmente devido à sua conexão com várias outras páginas ou sua vinculação a páginas importantes.

As páginas 1.html e 3.html têm valores de PageRank relativamente próximos. 1.html possui 0.2230 no PageRank iterativo e 0.2202 no PageRank amostral, enquanto 3.html apresenta 0.2167 e 0.2202, respectivamente. Esse equilíbrio sugere que ambas têm um nível de importância intermediário, estando moderadamente conectadas no corpus.

A página 4.html tem o menor valor de PageRank, com 0.1306 no método iterativo e 0.1307 no método amostral, indicando que essa página é a menos relevante ou conectada entre as quatro.

6.2 Resultados de corpus1

Nesta seção, são apresentados os resultados da aplicação dos método de pagerank iterativo e pagerank amostral sobre o conjustos de páginas web na pasta corpus1.

Página	**PageRank Iterativo**	**PageRank Amostral**
bfs.html	0.1200	0.1151
dfs.html	0.0781	0.0806
games.html	0.2269	0.2272
minesweeper.html	0.1177	0.1183
minimax.html	0.1294	0.1305
search.html	0.2122	0.2100
tictactoe.html	0.1157	0.1183

Tabela 2: Resultados de PageRank no corpus1

6.2.1 Análise dos Resultados

A página `games.html` apresenta o maior valor de PageRank tanto no método Iterativo quanto no Amostral, com valores muito próximos (Iterativo: 0.2269, Amostral: 0.2272). Isso

indica que esta página tem grande importância e interconexão dentro do corpus1.

A página `dfs.html` tem o menor valor de PageRank em ambos os métodos (Iterativo: 0.0781, Amostral: 0.0806), sugerindo que essa página tem menos relevância ou menos ligações com outras páginas no corpus.

As páginas `search.html` e `minimax.html` também se destacam com valores relativamente elevados, sugerindo que essas páginas são bem interligadas e possuem uma boa relevância no contexto do corpus.

De maneira geral, os valores do PageRank Iterativo e Amostral são muito próximos, com pequenas diferenças que podem ser atribuídas às naturezas distintas dos métodos: o algoritmo iterativo refina os valores ao longo das iterações, enquanto o amostral faz uma aproximação estatística.

6.3 Resultados de corpus2

Nesta seção, apresentamos os resultados obtidos pela aplicação dos métodos de PageRank Interativo e PageRank Amostral em um conjuto de páginas web da pasta corpus2. O PageRank Iterativo é uma implementação tradicional do algoritmo do pagerank, enquanto pagerank amostral usa técnicas em amostragem para estimar os mesmo valores. A **Tabela 3** abaixo resume os valores de PageRank cada página.

Página	PageRank Iterativo	PageRank Amostral
ai.html	0.1877	0.1884
algorithms.html	0.1263	0.1067
c.html	0.1315	0.1243
inference.html	0.1315	0.1291
logic.html	0.0270	0.0264
programming.html	0.2329	0.2293
python.html	0.1234	0.1243
recursion.html	0.0681	0.0716

Tabela 3: Resultados de PageRank no corpus2

6.3.1 Análise dos Resultados

As páginas `programming.html` e `ai.html` apresentam os maiores valores de PageRank em ambos os métodos. Isso indica que são as páginas mais relevantes ou interconectadas dentro do corpus analisado.

Em contraste, a página `logic.html` tem o menor valor de PageRank, sugerindo uma menor influência ou menor número de interligações em comparação com outras páginas.

Observamos que, na maioria dos casos, os resultados entre os dois métodos são bastante próximos, o que valida a consistência das duas abordagens. Pequenas discrepâncias podem ser explicadas pela natureza diferente dos cálculos: o método iterativo refina os valores a cada iteração, enquanto o amostral faz uma estimativa com base em uma amostragem de links.