UFMS DIGITAL

Estrutura de Dados

Prof. Dr. Gedson Faria

Prof.ª Dr.ª Graziela Santos de Araújo

Prof. Dr. Jonathan de Andrade Silva

UFMS DIGITAL

Módulo 4 - PageRank: grafos

Unidade 2 - Algoritmo PageRank

PageRank

- Algoritmo proposto pelos fundadores do Google (Larry Page e Sergey Brin), em 1998, para acelerar busca por páginas na internet e ranquear as páginas mais relevantes;
- O PageRank funciona como uma métrica de avaliação da relevância das páginas web de acordo com a quantidade e qualidade dos links em cada página.

PageRank

- Combina conceitos de Álgebra Linear, Cadeias de Markov e Grafos.
  - Utiliza a representação das páginas web por meio de um grafo direcionado.

PageRank

- A ideia central é que as páginas importantes que recebem mais links de outras páginas importantes terão um valor de PageRank mais alto.
  - Processo iterativo para determinar essa importância relativa entre as páginas.

PageRank

- Se temos muitas páginas recomendando uma página da web, podemos dizer que esta é uma página importante.
  - Se temos muitas páginas importantes recomendando outra página web, então essa página deve ser uma página ainda mais importante.

PageRank - Representação

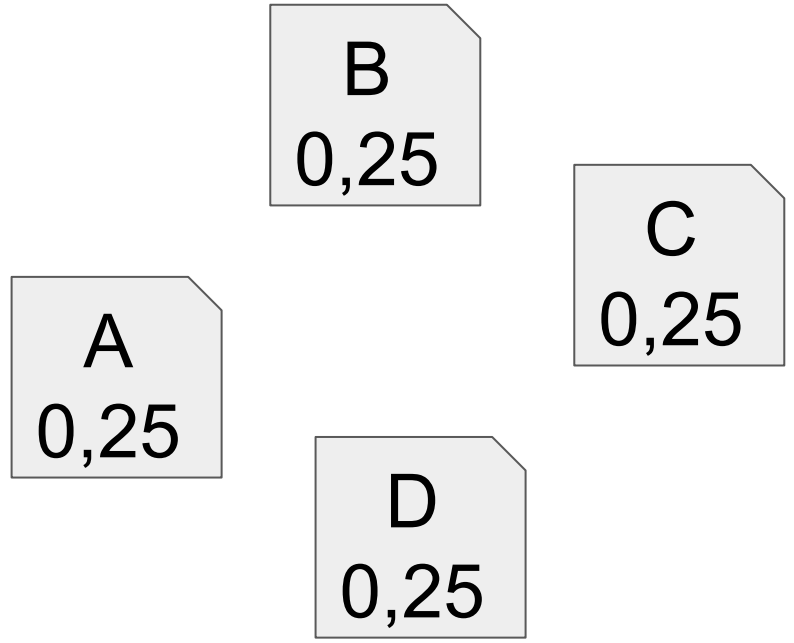
- Podemos utilizar a matriz de adjacências para representar a estrutura de conexões de um grafo de páginas web;
- Cada página é um vértice desse grafo e as arestas são os links que levam a uma outra página (grafo direcionado);

PageRank - Representação

- Cada aresta tem uma porcentagem (peso) do valor de PageRank da página que o recomenda (grafo ponderado).
  - Uma página X com valor de PageRank em 1 terá esse valor distribuído em suas arestas (geralmente de maneira uniforme), indicando o peso das conexões (qualidade da página).

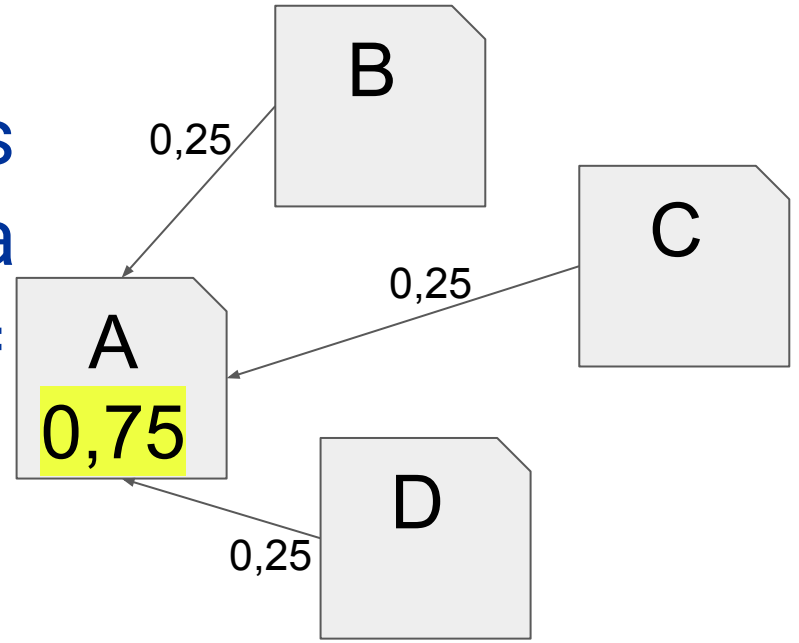
PageRank - Exemplo Simplificado

- Considere 4 páginas web: A, B, C e D.
  - A soma do valor de PageRank (PR) de todas as páginas é 1.
    - 1 = PR(A)+PR(B)+PR(C)+PR(D);
  - Inicialmente vamos distribuir igualmente (0,25 para cada página);
    - PR(A)=0,25;
    - PR(B)=0,25;
    - PR(C)=0,25;
    - PR(D)=0,25;



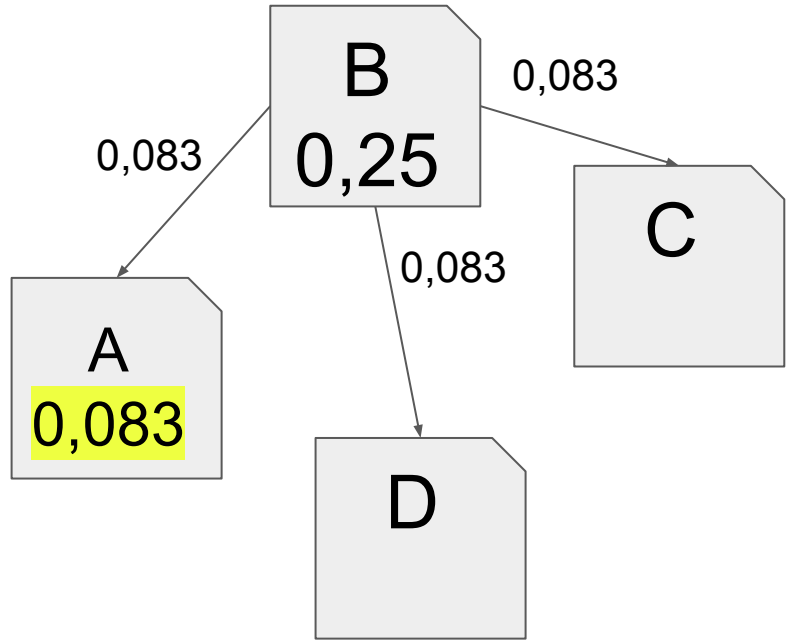
PageRank - Exemplo Simplificado

- Se apenas as páginas B, C e D recomendarem a página A, os seus valores de PageRank serão enviados para a página A, resultando em um novo valor de PR(A) = 0,75.
  - PR(A) = PR(B)+PR(C)+PR(D);
  - PR(A) = 0,25+0,25+0,25;
  - PR(A) = 0,75;
- Indicando que a página A é importante.



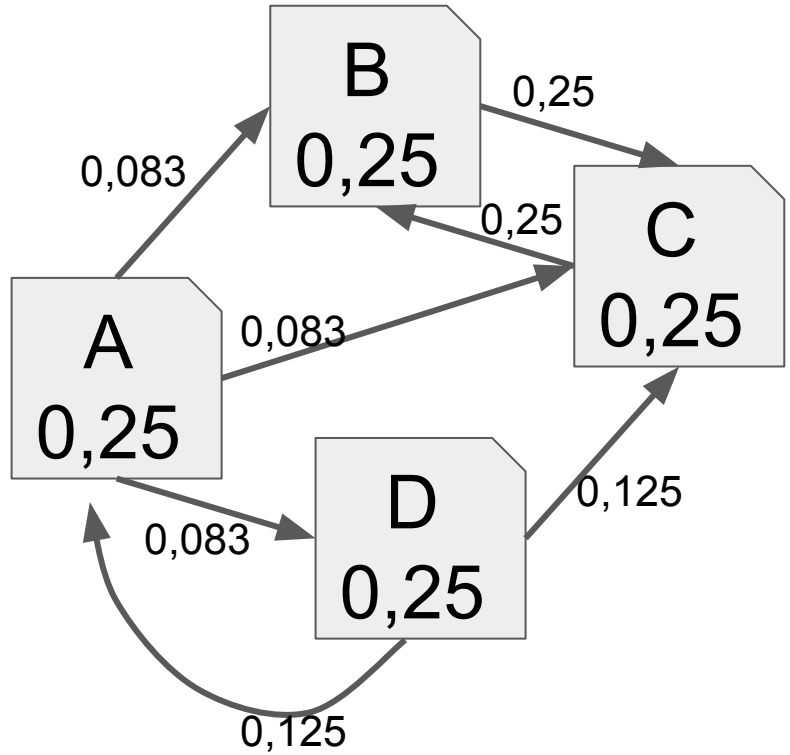
PageRank - Exemplo Simplificado

- Se a página B recomenda todas as páginas, o PageRank da página A será ⅓ do PR(B).
  - PR(A) = ⅓ * PR(B);
  - PR(A) = 0,25/3
  - PR(A) = 0,083;



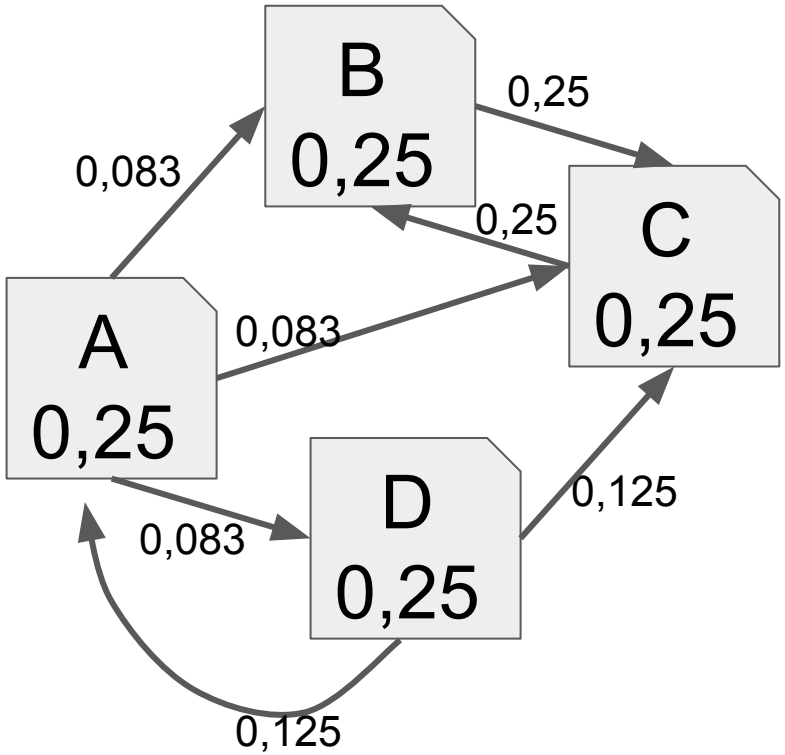
PageRank - Exemplo Simplificado

- Vamos considerar a seguinte ligação entre essas páginas.
  - Página A distribui ⅓ de 0,25 para cada página (0,083 para B, C e D);
  - Página C distribuiu todo o valor de PageRank para página B (0,25), o mesmo para a página B.
  - D dividiu pela metade (½) o seu PageRank e distribuiu para as páginas A e C o valor 0,125.



PageRank - Exemplo Simplificado

- Podemos representar esse grafo com os valores de PR de cada vértice no vetor V^0 e matriz de transições T e matriz pesos P^0:



0,25
0,25
0,25
0,25

V^0

0	⅓	⅓	⅓
0	0	1	0
0	1	0	0
½	0	½	0

T

0	0,083	0,083	0,083
0	0	0,25	0
0	0,25	0	0
0,125	0	0,125	0

P^0

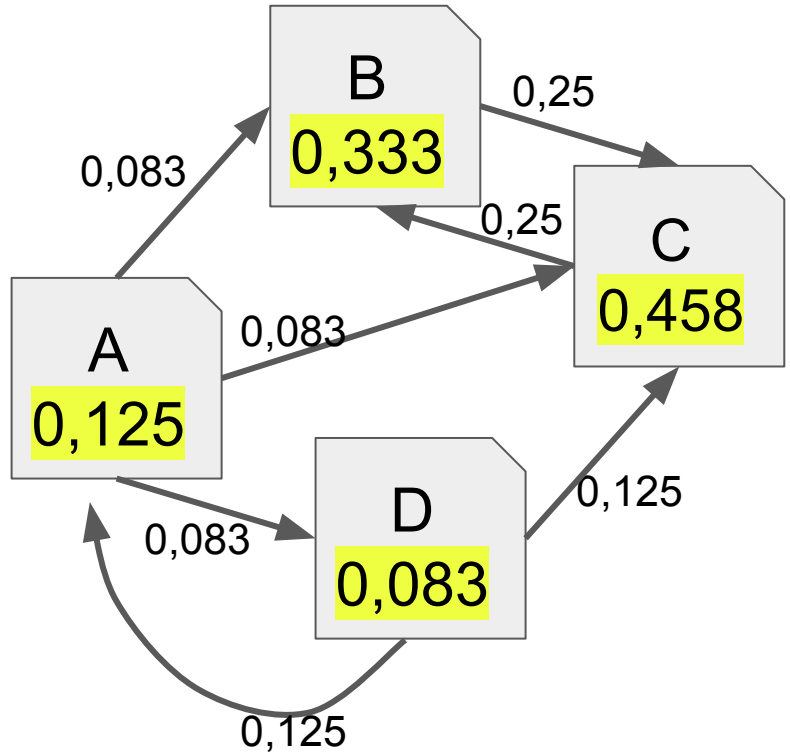
PageRank - Exemplo Simplificado

- Podemos observar então que o valor de PageRank da página B é:

  PR(B) = PR(A)/Nº links saindo de A +

  PR(C)/Nº links saindo de C

- PR(B) = 0,083+0,25 = 0,333; PR(A) = 0,125; PR(C)=0,458 e PR(D)=0,083.



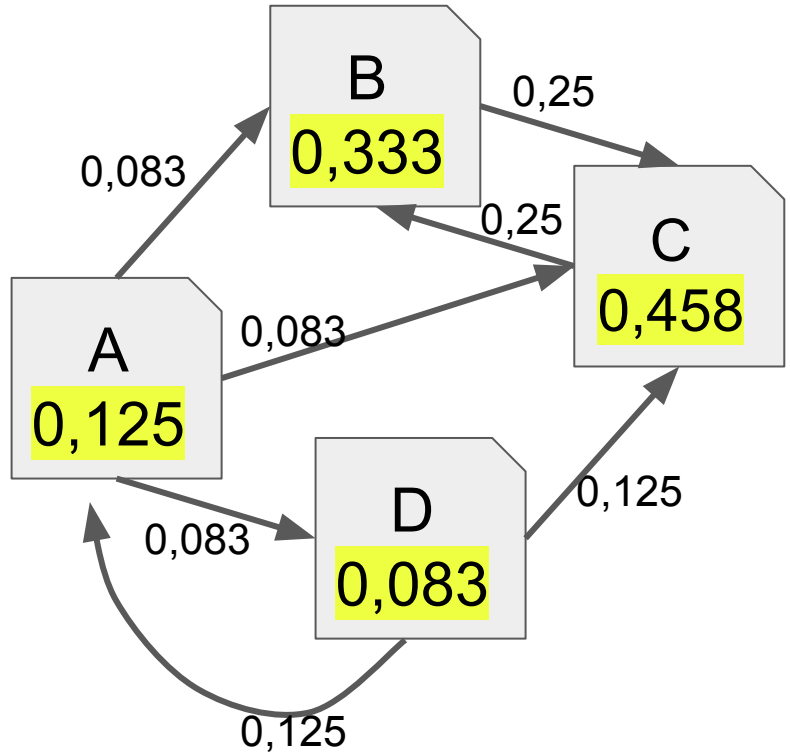
PageRank - Exemplo Simplificado

- Com os pesos das arestas P^0, vamos recalcular o PR de cada página para obter o novo V^1.

P^0

0	0,083	0,083	0,083
0	0	0,25	0
0	0,25	0	0
0,125	0	0,125	0

$\sum =$ | 0,125 | 0,333 | 0,458 | 0,083 | V^1



PageRank - Exemplo Simplificado

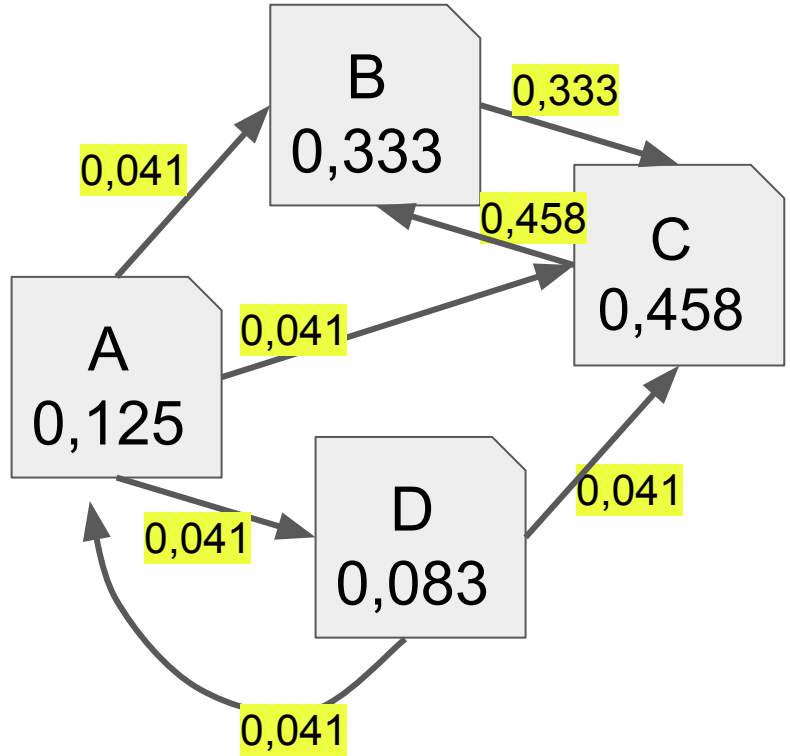
- Em um nova iteração vamos recalcular os pesos das arestas P^1.

0,125
0,333
0,458
0,083

V^1

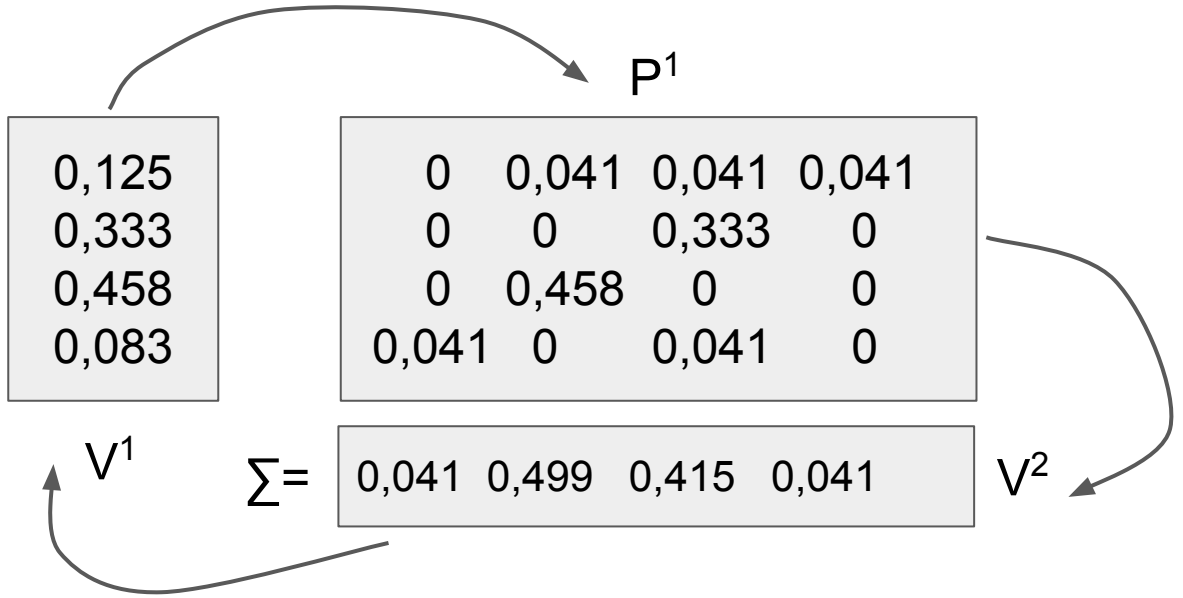
0	0,041	0,041	0,041
0	0	0,333	0
0	0,458	0	0
0,041	0	0,041	0

P^1



PageRank - Exemplo Simplificado

- O algoritmo continua sua execução até que um número específico de iterações seja alcançado ou os valores de V em duas iterações consecutivas não se alterem.
  - Simulação do exemplo.



PageRank - Considerações

- Existem cenários em que podemos ter links “mortos” (*dead links*);
  - Páginas que não recomendam nenhuma outra (sumidouro).
  - Nesses casos, o algoritmo pode não conseguir reajustar os valores de PageRank uma vez que não redistribui.

PageRank - Considerações

- Algumas páginas podem artificialmente induzir páginas importantes (quantidade de links)
  - Páginas com auto ligação (*self loop*);
  - Múltiplos links de uma página para outra página.

PageRank - Considerações

- Uma maneira de amenizar essas situações é a inserção de um fator de amortecimento d (*damping factor*):
  - Na página B do nosso exemplo:
    - Sem *damping factor*:
      - PR(B) = PR(A)/Nº links saindo de A +
      - PR(C)/Nº links saindo de C
    - Com *damping factor*:
      - PR(B) = (1-d)/N° de páginas + d * (PR(A)/Nº links saindo de A + PR(C)/Nº links saindo de C)

PageRank - Considerações

- A ideia do uso de um fator de amortecimento é simular o cenário em que o usuário pode em algum momento deixar de seguir as indicações de páginas e começar a buscar um caminho alternativo.
  - Geralmente assume-se que em 85% do casos o usuário segue os links e em 15% deseja buscar alternativas;
    - d = 0,85
    - PR(B) = (1-0,85)/N° de páginas + 0,85 * (PR(A)/Nº links saindo de A + PR(C)/Nº links saindo de C)

Referências

OLIVEIRA, M. V.; BOVOLONI, J. O.; Leite Filho, E. S.; MENEZES, G. Page Rank: o funcionamento da ferramenta de busca do Google. **Cadernos de Graduação: Ciências exatas e tecnológicas**, Aracaju, v. 3, n. 3, p. 73-84, out. 2016,. Disponível em: https://link.ufms.br/AqveS, Acesso em 30 mai 2023.

CAMARGO, Ariely da Silva; GALVES, Ana Paula Tremura. Abordagem matemática por trás do algoritmo PageRank. **Revista Eletrônica Paulista de Matemática**, v. 21, dez.2021, Disponível em: https://link.ufms.br/VVRSZ . Acesso em 30 mai 2023.

Licenciamento