GSI024 - Organização e recuperação de informação

Prof. Dr. Rodrigo Sanches Miani (FACOM/UFU)

Última atualização - Agosto/2022

Análise de Links - PageRank

Agenda

“Análise de Links - PageRank”

Motivação

História

Ideia básica

Algoritmo simplificado

Cálculo do PageRank

TP-4

TP-4

- Começaram a ler as referências?
- Estão conseguindo fazer?

Motivação

Motivação

Suponha que um modelo clássico, como o vetorial, foi empregado para recuperar informações na Internet. Quais resultados seriam mostrados pela máquina de busca?

Motivação

Suponha que um modelo clássico, como o vetorial, foi empregado para recuperar informações na Internet. Quais resultados seriam mostrados pela máquina de busca?

1. Essa estrutura seria aceitável?
2. Quais os problemas?
3. O que fazer para melhorar?

Desafios causados pela Internet - Busca

- Dados distribuídos;
- Alto percentual de dados voláteis;
- Grande volume de dados (http://www.worldwidewebsize.com/);
- Dados redundantes;
- Qualidade dos dados;
- Dados heterogêneos.

Usuário x Sistema de busca

- Usuário:
  - Conceber uma boa consulta a ser submetida ao sistema de busca.

- Sistema de busca
  - Realizar uma busca rápida e devolver respostas relevantes, mesmo para consultas mal formuladas.

História

Um pouco de história

- Encontrar o site correto, que atenda ao interesse do usuário da máquina de busca Web é algo muito valioso;

- Até 1997/1998 a principal tecnologia para esse fim era fornecida pelo portal Altavista.

Um pouco de história - Altavista

Pilares da busca Web:

1. Uso de programas (webcrawlers) para percorrer os sites da Internet, cadastrando todas as palavras e os links existentes;

2. Cadastramento e organização das palavras em índices, ligando essas palavras ao endereço dos sites;

3. Realização de um ranking da importância ou popularidade relativa de cada site que contém as pesquisa, apresentando ao interessado os sites na ordem do ranking obtido.

Um pouco de história

- O Altavista avançou bastante nos 2 primeiros pilares;
- A solução proposta para o terceiro pilar não era a mais adequada;
- Em pouco tempo, diversos mecanismos foram criados para iludir o processo de elaboração do ranking, diminuido com isso o valor das sugestões de sites oferecidos pelo Altavista.

Um pouco de história

- A solução para o terceiro pilar dos mecanismos de busca veio em 1997/1998 com a proposta do algoritmo PageRank;
- Algoritmo desenvolvido por dois estudantes de doutorado de Stanford;
- Fundadores de uma, até então, pequena empresa chamada Google.
- Grande parte do sucesso da companhia se deveu à qualidade de seu mecanismo para estabelecer o ranking de importância de sites.

Um pouco de história

- O algoritmo usa princípios de Álgebra Linear (autovalores e autovetores) e Probabilidade (cadeias de Markov) para encontrar o rank de uma página;

- Título do artigo original onde o Google e o PageRank foram propostos:
  - "The anatomy of a large-scale hypertextual Web search engine“
  - http://www.sciencedirect.com/science/article/pii/S016975529800110X

Ideia básica

Ideia básica de RI na Web

1. Criar um índice das páginas e de seus conteúdos;
   - Web Crawler – indexar as páginas e recuperar conteúdo;
   - O índice deve ser atualizado periodicamente.

2. Processamento da consulta do usuário:
   - Encontra milhares de páginas relevantes usando sinais de conteúdo (BM25, Vetorial);
   - Com o auxílio do PageRank (sinal estrutural dos links), ordena e exibe as páginas.

PageRank - Motivação

- A ideia principal do algoritmo é estabelecer um índice de popularidade (rank) para cada site da Internet;

- O índice é criado à partir dos sites com links apontando para um determinado site, ponderado pelo índice de popularidade desses sites, e assim recursivamente;

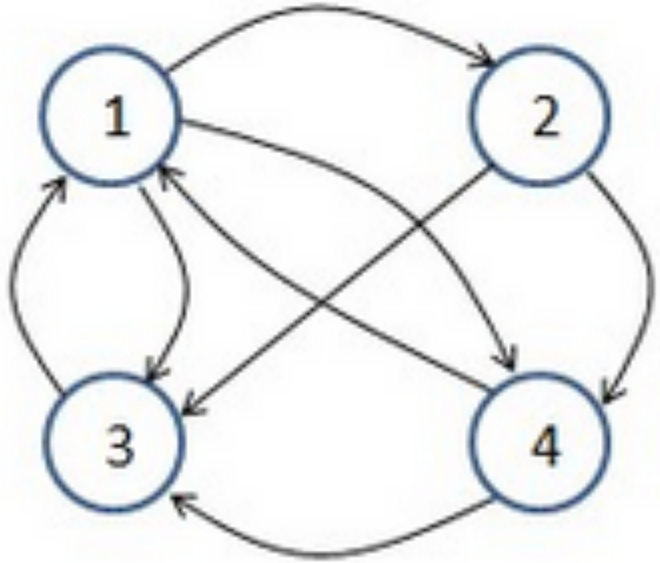
- Importante: o PageRank **não** depende de nenhuma consulta!

PageRank - Modelagem do problema

A Web é vista como uma rede de citações – grafo:

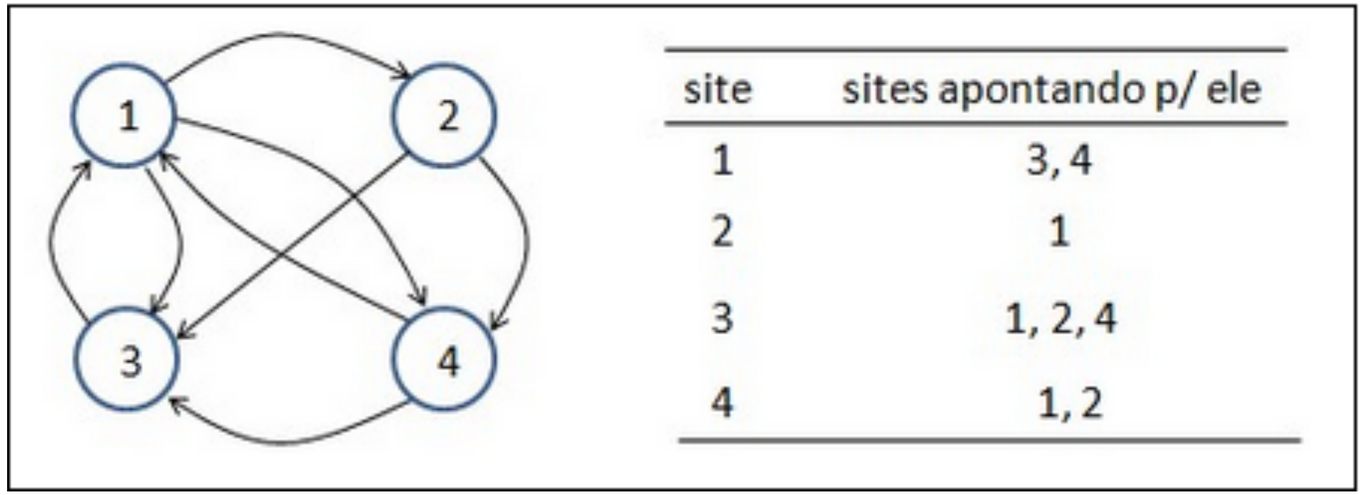
- Cada nó corresponde a uma página e cada ligação (aresta) corresponde a uma referência de uma página para outra (hiperlink).
- O PageRank atribuí um valor a cada nó (página) da rede;
- Um valor maior corresponde a um nó mais importante na rede.

PageRank - Exemplo



site	sites apontando p/ ele
1	3, 4
2	1
3	1, 2, 4
4	1, 2

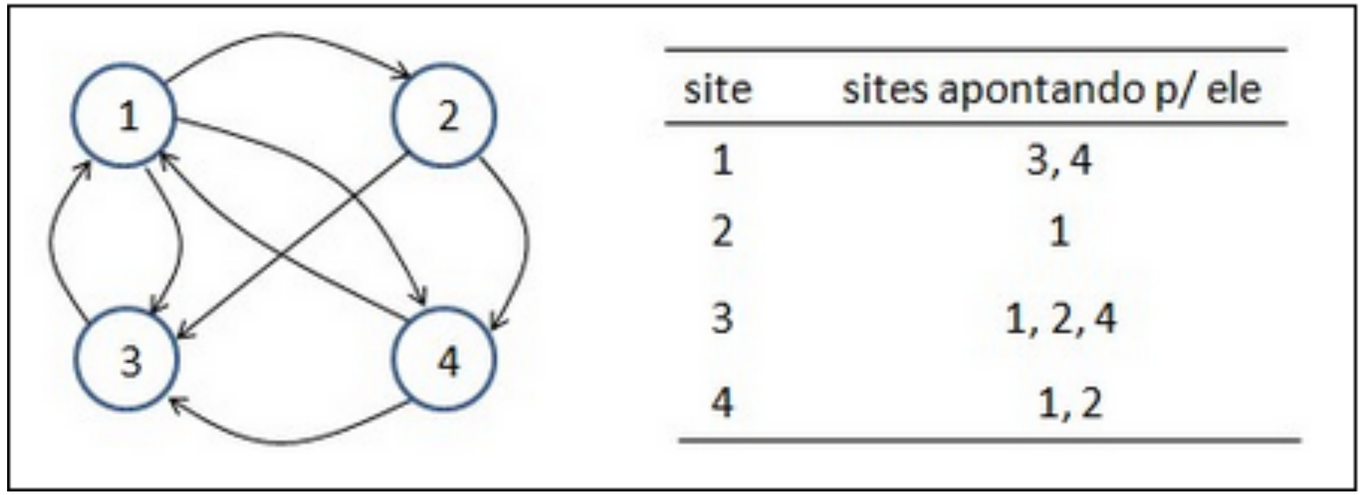
PageRank - Exemplo



site	sites apontando p/ ele
1	3, 4
2	1
3	1, 2, 4
4	1, 2

- O que poderia ser feito para calcular o PageRank dos sites 1, 2, 3 e 4 sabendo que a disposição deles segue o grafo acima? Ou seja quem é o site mais popular?

PageRank - Exemplo



site	sites apontando p/ ele
1	3, 4
2	1
3	1, 2, 4
4	1, 2

- O que poderia ser feito para calcular o PageRank dos sites 1, 2, 3 e 4 sabendo que a disposição deles segue o grafo acima? Ou seja quem é o site mais popular?
- Resposta: o site 3 possui 3 outros sites apontando para ele. É correto dizer que ele é o mais popular?

PageRank - Exemplo

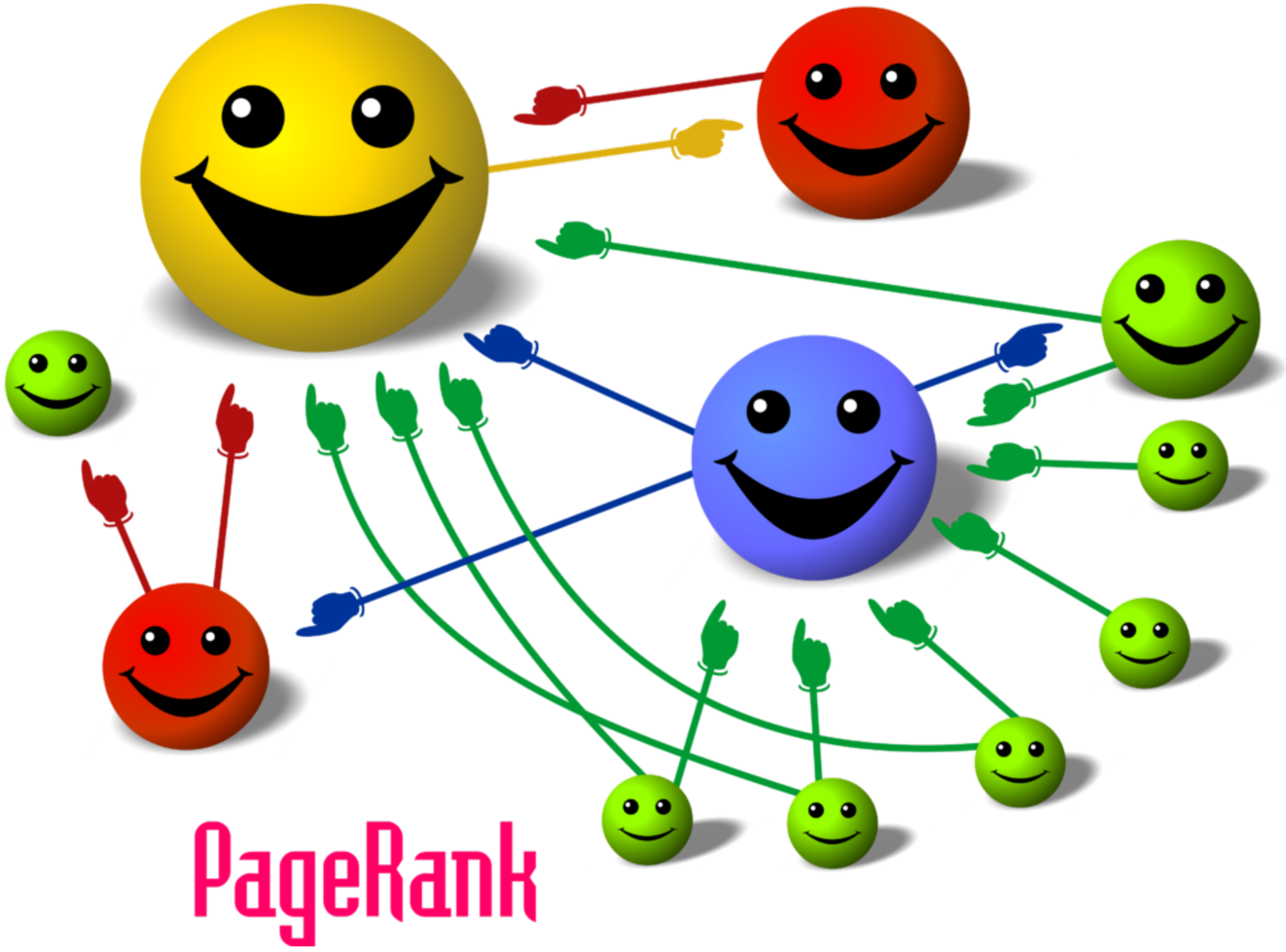
- Usar **somente** o número de links de forma isolada como critério de popularidade é um problema;
- Não é considerada a **popularidade** dos sites que estão apontando para o site de interesse;
- A simples criação de muitos sites, sem expressão alguma, apontando para o site de interesse, poderia facilmente “inflar” a popularidade desse site.

PageRank - Ideia Básica

Uma página A tem um valor mais alto de PageRank se:

1. Existem muitas páginas que apontam para A;
2. Existem algumas páginas com valor de PageRank alto que apontam para A.

PageRank - Ideia Básica



PageRank - Proporcionalidade dos votos

- Entender cada link dentro de um site A, apontando para outro site B, como sendo um **voto** proporcional do site A para o site B com respeito à popularidade;
- Nessa interpretação “democrática” da Internet, cada site teria direito a um **único voto**, de forma que se apontasse para *s* sites, os votos computados para cada um desses sites seria de fato somente *1/s*;
- E a popularidade do site no processo de voto?

PageRank - Proporcionalidade dos votos

Cada voto (proporcional) seria **ponderado** pelo **índice de popularidade** do site, visando estabelecer um processo democrático com um componente meritocrático, fundamentado na popularidade.

Algoritmo simplificado

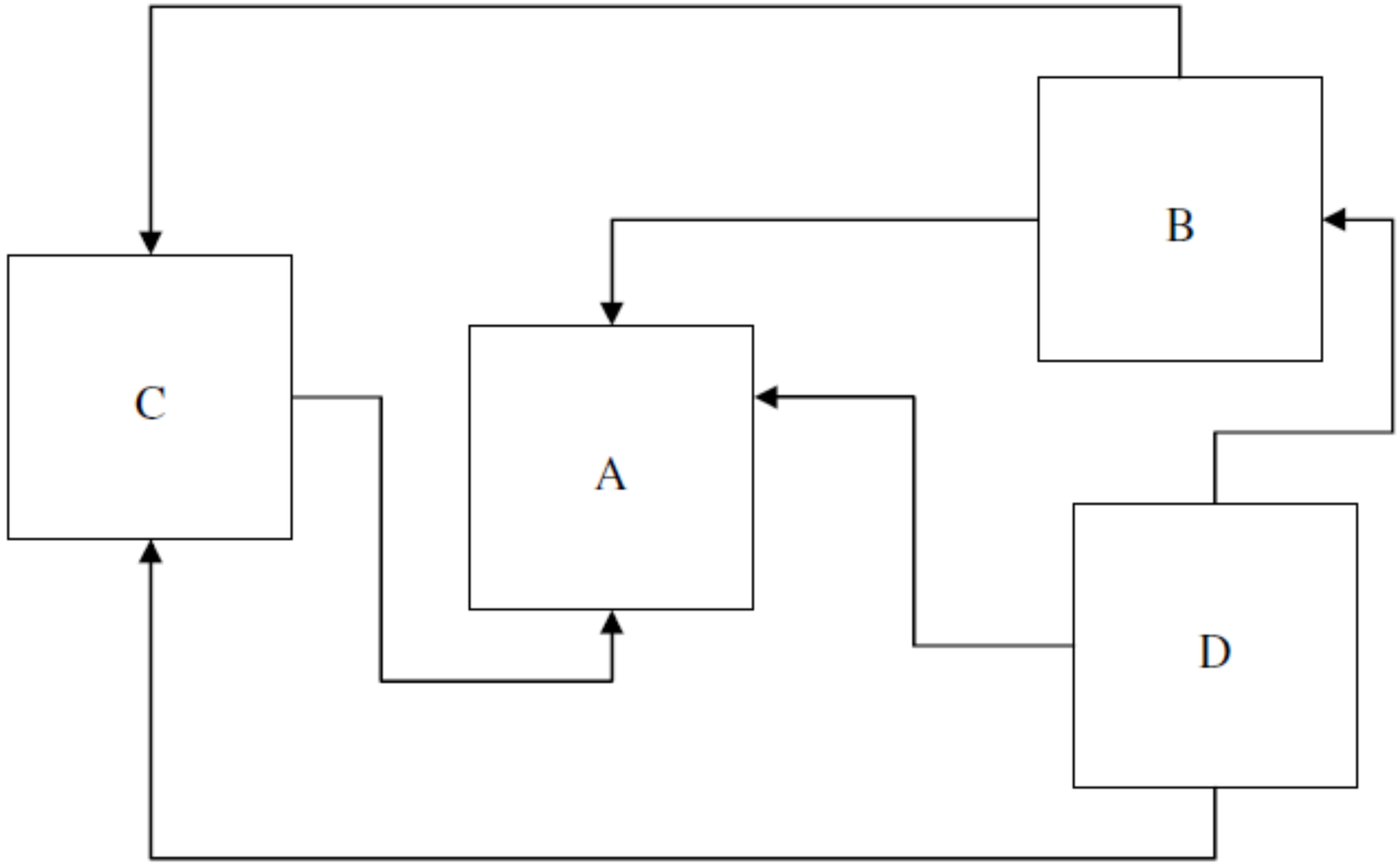
PageRank – Algoritmo simplificado

- Imagine uma "Internet" formada por apenas 4 páginas A, B, C e D.

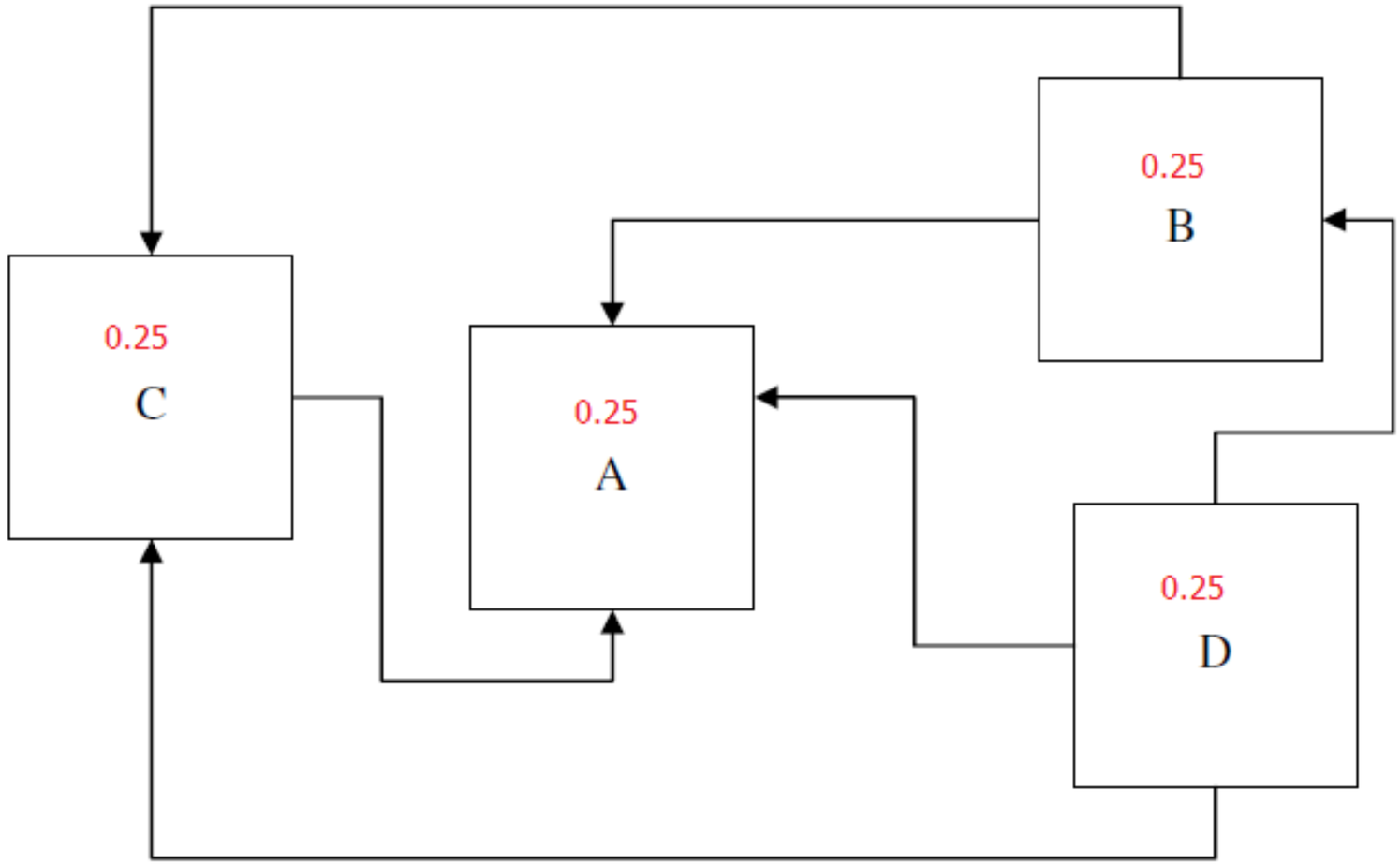
- No primeiro passo do processo de cálculo iterativo, todas as páginas têm o mesmo valor de PageRank;

- Nas iterações seguintes cada página "transfere" valores de PageRank para as páginas que elas possuem links.

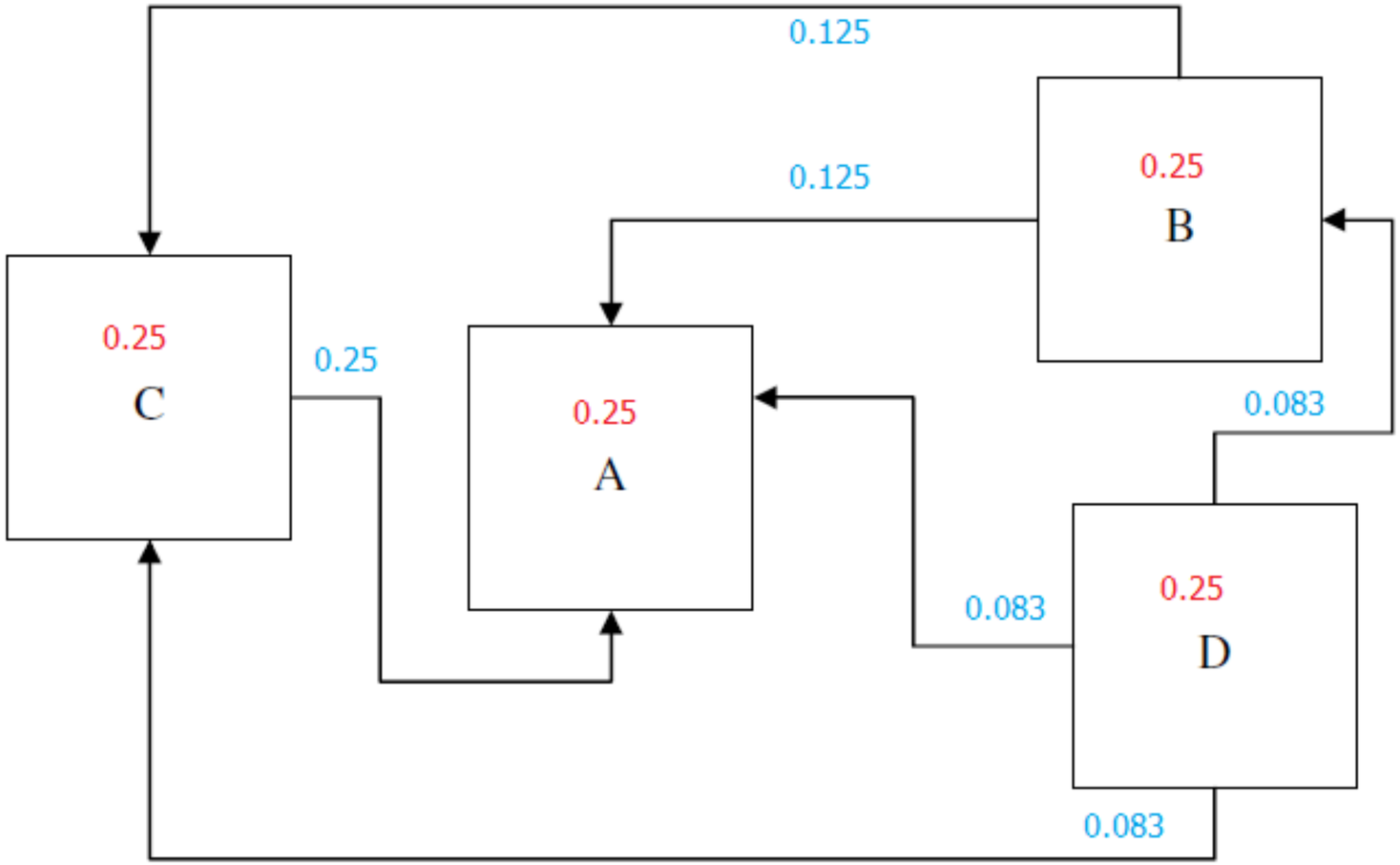
PageRank – Algoritmo simplificado



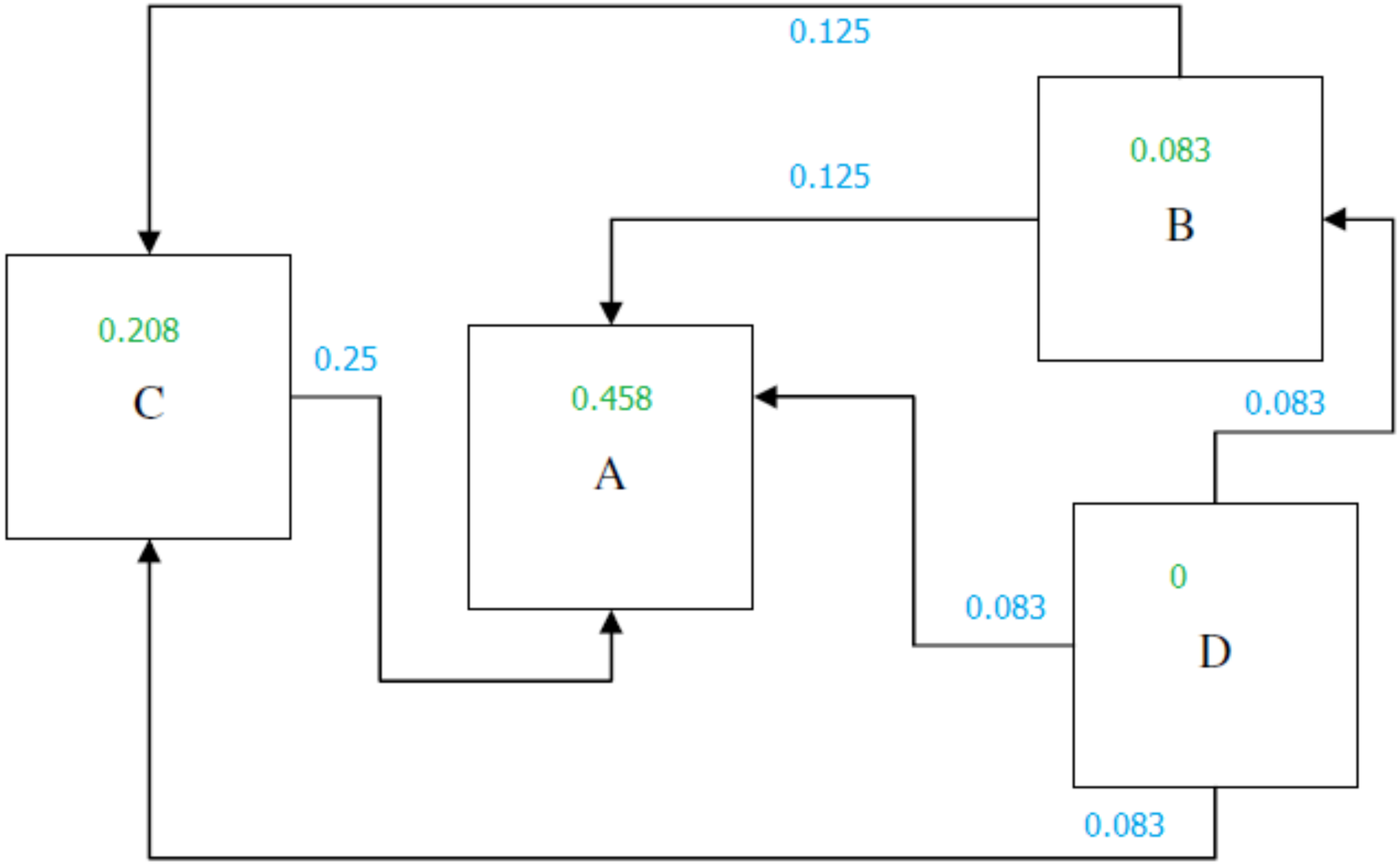
Algoritmo simplificado – Iteração 1



Algoritmo simplificado – Iteração 2



Algoritmo simplificado – Iteração 3



PageRank – Algoritmo simplificado

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}$$

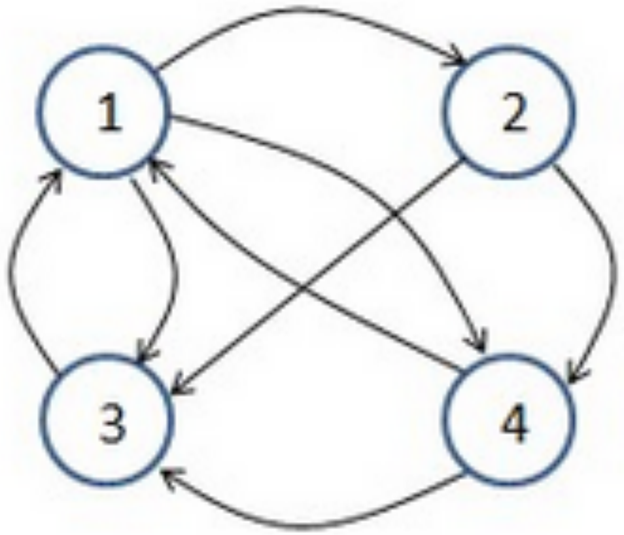
$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}$$

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}$$

PageRank – Algoritmo simplificado

- PR(A) = PageRank da página A;
- L = função que retorna o número de links em uma página;
- B(u) = conjunto de todas as páginas que referenciam a página u.

PageRank – Valor do PageRank para cada site?



site	sites apontando p/ ele
1	3, 4
2	1
3	1, 2, 4
4	1, 2

Cálculo do PageRank

Modelagem

- O modelo do PageRank é baseado na suposição de um usuário navegando na Web aleatoriamente;
- O usuário clica em links sucessivos de forma aleatória;
- Eventualmente ele pode ficar entediado e pular a página que está vendo e visitar outra página;

Modelagem do algoritmo completo

- Para modelar a probabilidade do usuário "ficar entediado", é usado um fator de amortização *d* que assume valores entre 0 e 1;
- Normalmente, o valor *d* é fixo em 0,85;
- Este valor representa uma probabilidade de 85% do usuário clicar em um link na página que está vendo, contra 15% de probabilidade de escolher outra página aleatoriamente para começar a navegar novamente.

PageRank + fator de amortização

Dois componentes:

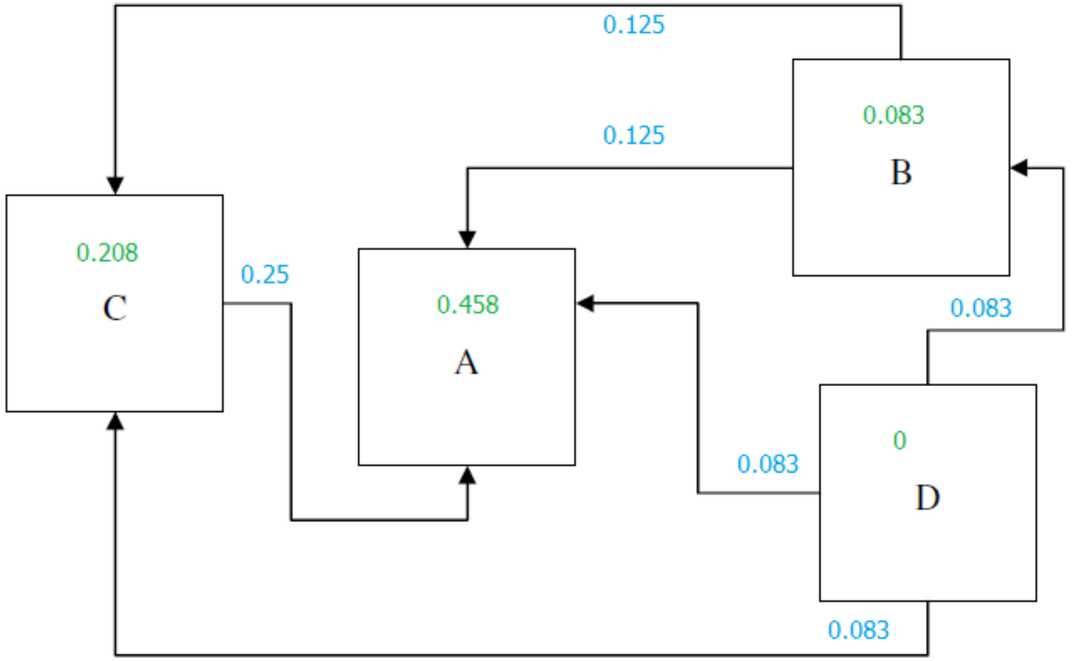
- Contribuição das páginas que apontam para A, ponderado pela probabilidade **d** do usuário seguir os links das páginas;
- Usuário ter selecionado a página aleatoriamente, ponderado pela probabilidade de o usuário não seguir os links das páginas.

$$PR(A) = \frac{1-d}{N} + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right)$$

PageRank + fator de amortização

Essa modelagem permite que páginas que não foram citadas por ninguém (páginas sem links) também possua um valor de PageRank diferente de zero.

$$PR(A) = \frac{1-d}{N} + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right)$$



PageRank + fator de amortização

- Uma página, pelo simples fato de existir, tem uma probabilidade igual a todas as outras de ser selecionada aleatoriamente pelo usuário;

- Uma página que não tenha ligações está ligada a todas as páginas da rede.

PageRank – Implementação iterativa

De um modo geral, para calcularmos os valores do PageRank de um conjunto de *n* páginas ligadas entre si, teremos que resolver um sistema de *n* equações (as equações do PageRank de cada uma das páginas) com *n* incógnitas (o valor do PageRank de cada uma das páginas).

- Se o valor de *n* for pequeno, não há problema. No entanto, a base de dados do Google é formada por bilhões de páginas, o que torna pouco prático o cálculo do valor exato do PageRank.

- Neste caso, o que se faz é um cálculo iterativo de valores aproximados do PageRank, sendo estes valores tanto mais próximos dos valores exatos quanto maior for o número de iterações.

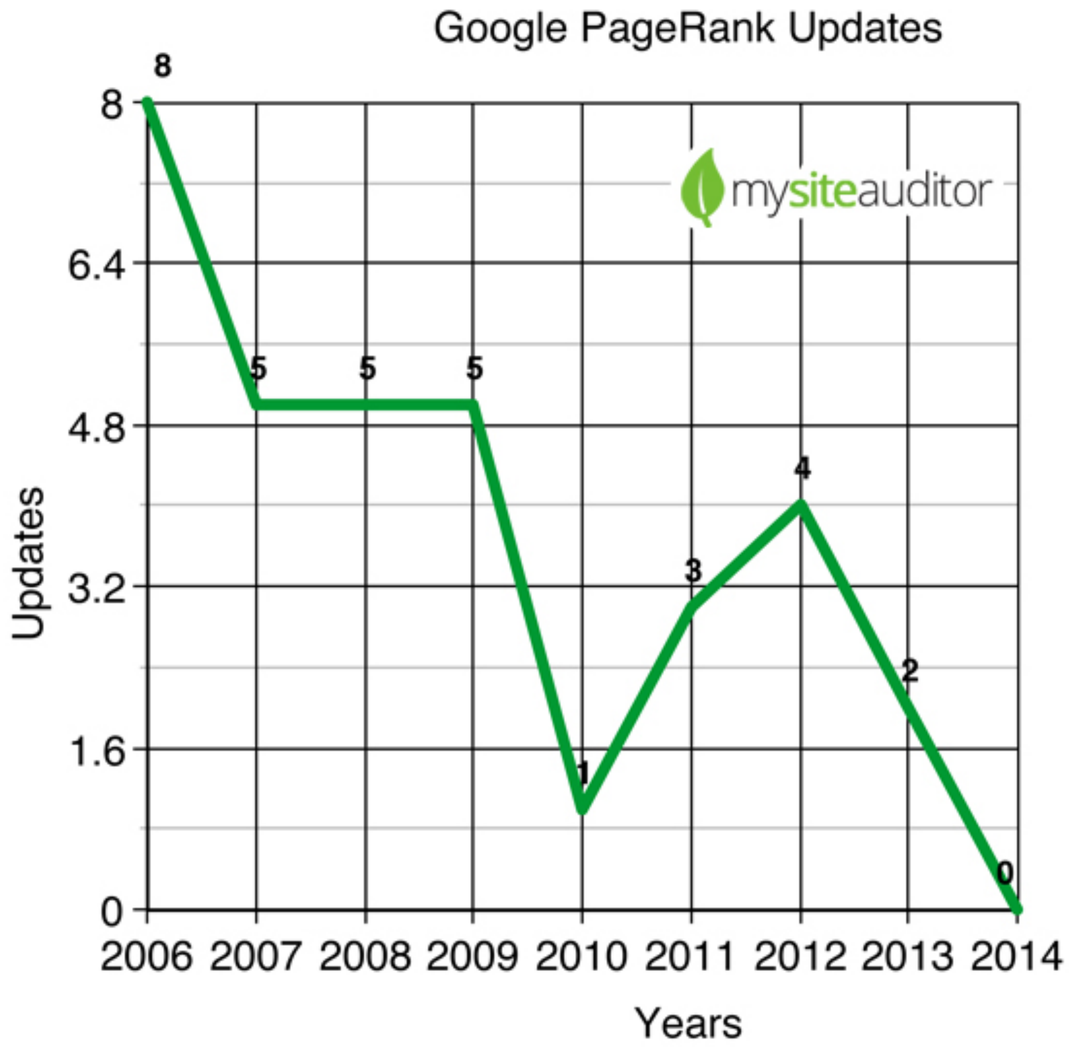
PageRank – Implementação iterativa

- Representação matricial do somatório M – cada linha da matriz representa a contribuição da página A para as outras páginas do conjunto;
- Ajuste do fator de amortecimento;
- A solução é o autovetor da matriz M, correspondente ao autovalor 1 (solução através do método da potência – *power method*).

PageRank – Implementação iterativa

- No caso real do Google, a matriz M tem mais de 5 bilhões de linhas e colunas;
- Os cálculos demoram dias, sendo utilizadas, cerca de 50 a 100 iterações (com algumas otimizações para minimização de cálculos desnecessários);
- Não há motivo para se ter uma precisão absoluta no resultado;
- O processo iterativo usualmente termina na iteração k quando, os últimos 2 vetores obtidos estão suficientemente próximos, de acordo com alguma métrica.

PageRank – Atualizações



PageRank – Atualizações

- O Google não divulga os dados do cálculo do PageRank desde 2014;
- Tais dados eram constantemente utilizados por PageRank Spammers (pessoas vendendo links em páginas com alto valor de PageRank);
- Ou seja, o público não verá mais esses dados, mas o Google ainda o utiliza para a criação do ranking.

Ranqueamento de máquinas de busca Web

Onde o PageRank entra na prática?

Existem diversos tipos de sinais usados para construir as funções de ranqueamento de motores de busca Web:

1. Sinais de conteúdo - relacionados ao texto propriamente dito (modelo vetorial + bag of words + TF-IDF, por exemplo);

2. Sinais estruturais - relacionados à estrutura de links da Web (PageRank!);

3. Uso da Web - Realimentação usando cliques, contexto geográfico do usuário, contexto tecnológico e contexto temporal (histórico de consultas).

Ou seja…

Uma simples consulta no Google envolve TUDO o que vimos durante o curso:

- Identificar os sinais de conteúdo usando os modelos estudados (TF-IDF e modelo vetorial);
- Identificar as páginas mais populares (PageRank);
- Expandir a consulta através da i) identificação de diversos contextos associados ao usuário (localização, navegador, sistema operacional, histórico de consulta) e ii) dados implícitos ou explícitos.

Roteiro de estudo

Referências

- Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca
  - Capítulo 9.5
- http://rpubs.com/adriano/PageRank
- http://en.wikipedia.org/wiki/PageRank
- http://www.inf.ufrgs.br/~lzgallina/files/Pagerank%20para%20Ordenacao%20de%20Resultados%20em%20Ferramenta%20de%20Busca%20na%20Web.pdf
- http://www.teses.usp.br/teses/disponiveis/45/45133/tde-08052009-152811/publico/dissertacaofinalsubmissao.pdf