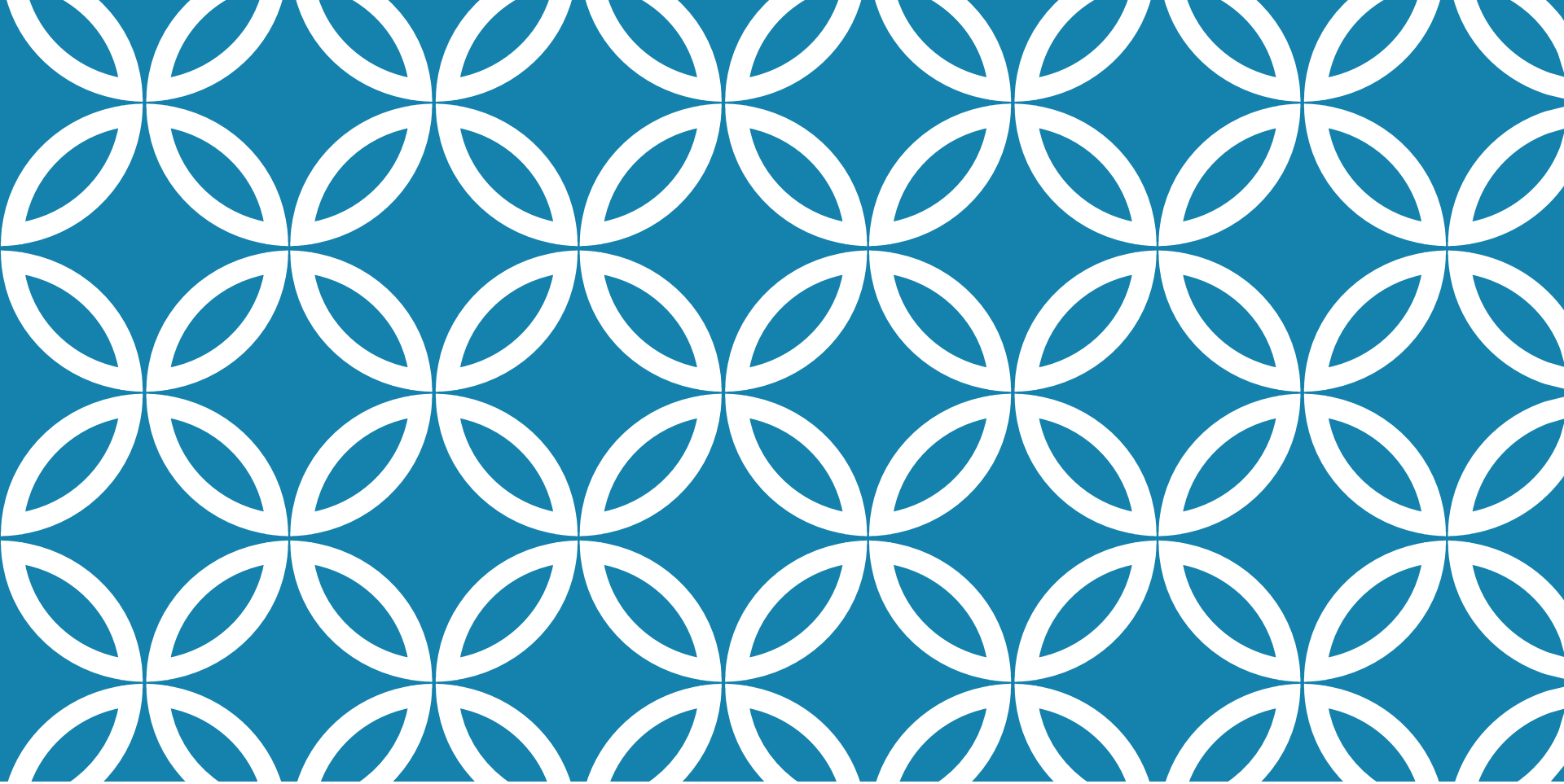


ANÁLISE DE LINKS E BUSCA NA WEB

Redes Sociais e Econômicas

Prof. André Vignatti

The PageRank Citation Ranking: Bringing Order to the Web

January 29, 1998

Abstract

The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them.

We compare PageRank to an idealized random Web surfer. We show how to efficiently compute PageRank for large numbers of pages. And, we show how to apply PageRank to search and to user navigation.

1 Introduction and Motivation

PAGERANK

o **PageRank** é uma espécie de "fluido" que circula pela rede

para uma rede com n nós, o **PageRank** é calculado da seguinte forma:

- atribuir a todos os nós o mesmo **PageRank** inicial, $1/n$
- escolha um número de passos k
- faça k atualizações dos valores **PageRank** usando a seguinte regra:
- **regra de atualização básica do PR**: cada página divide igualmente seu **PageRank** atual em suas arestas de saída e passa essas partes iguais para as páginas que ela aponta (se a página não tem arestas de saída, ela passa todo seu **PageRank** para si mesma)

cada página atualiza seu novo **PageRank** como sendo a **soma das partes que recebe**

PAGERANK

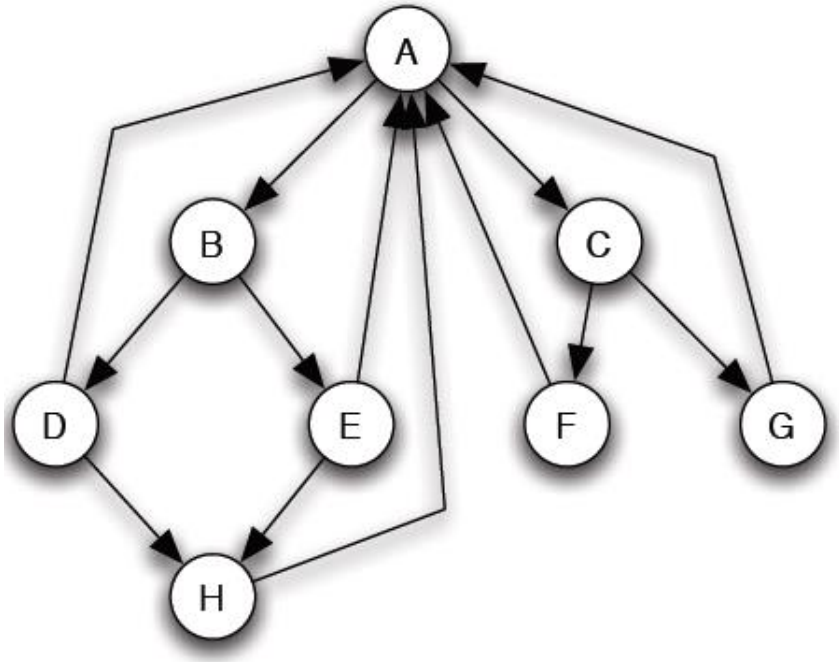
note que PageRank **não é criado nem destruído**, só se move de nó em nó

assim o PageRank total da rede sempre permance o mesmo

desta forma, **não é necessário um passo de normalização** (como nos valores de Hubs e Autoridades)

PAGERANK - EXEMPLO

todas as páginas começam com **PageRank** de $1/8$



após as primeiras duas atualizações fica:

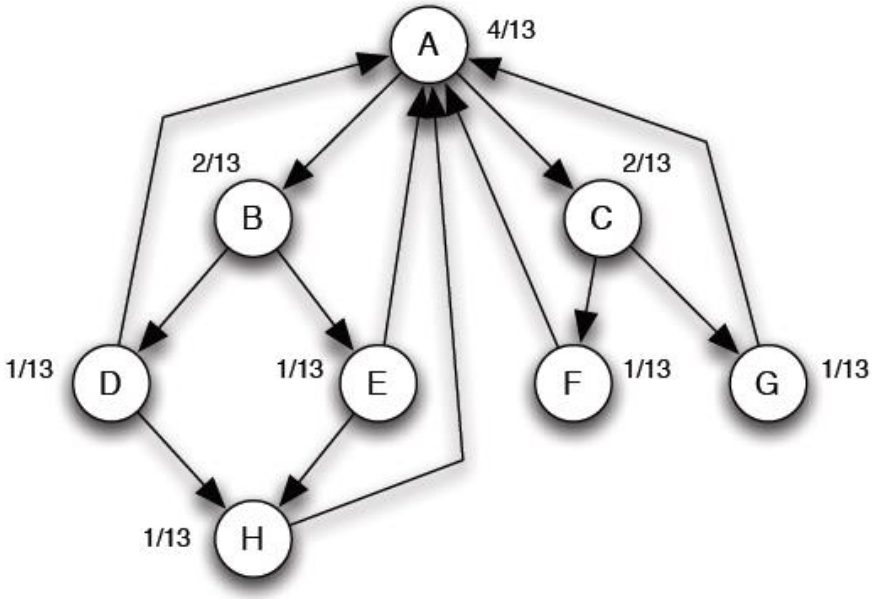
Step	A	B	C	D	E	F	G	H
1	1/2	1/16	1/16	1/16	1/16	1/16	1/16	1/8
2	3/16	1/4	1/4	1/32	1/32	1/32	1/32	1/16

VALORES DE EQUILÍBRIO DO PAGERANK

os valores **PageRank** de todos os nós convergem para limites à medida que ***k*** tende ao infinito (exceto em certos casos degenerados)

os valores-limite apresentam o seguinte tipo de **equilíbrio:**

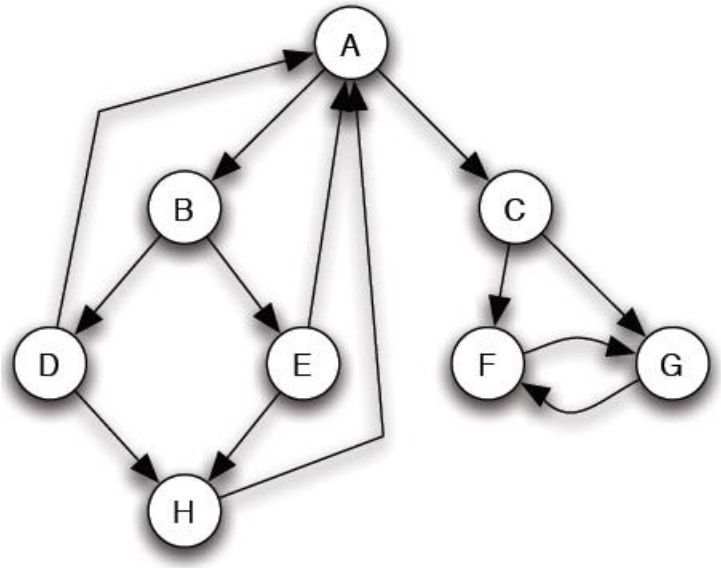
- a soma de todos os valores de **PageRank** = ***1***
- ao aplicar a **atualização básica de PageRank,** chega-se nos mesmos valores



PAGERANK: UM PROBLEMA

na maioria das redes reais o **PageRank** pode **"vazar"**

- **PageRank** que flui de ***C*** para ***F*** e ***G*** não podem circular de volta para o resto da rede
- **PageRank** de ***F*** e ***G*** ficam ***1/2*** cada, o resto fica com ***0***



claramente, isso não reflete a importância das páginas ***F*** e ***G***

PAGERANK: UM PROBLEMA

se houver um pequeno conjunto de nós que pode ser alcançado a partir do resto do grafo, mas não têm caminhos de volta, então **PageRank** se **acumulará lá** (ideia: componentes fortemente conexos)

felizmente, há uma maneira simples natural de resolver isso

ESCALANDO A DEFINIÇÃO DE PAGERANK

por que toda a água do planeta não se concentra somente nos pontos mais baixos?

porque **a água evapora**, e **chove** nas partes elevadas novamente!

vamos usar essa ideia…

ESCALANDO A DEFINIÇÃO DE PAGERANK

escolhemos um **fator de escala** s entre *0* e *1*

substituímos a regra de atualização básica por esta:

regra de atualização escalada do PageRank:

- primeiro aplique a **regra básica de atualização PageRank**
- em seguida, **reduza** todos os valores de **PageRank** por um **fator de** s
- o **PageRank** total na rede **diminuiu** de 1 para s
- dividimos a quantidade residual $1 - s$ de **PageRank** igualmente por todos nós , dando $(1 - s)/n$ para cada **(fizemos chover!)**

LIMITES DA REGRA DE ATUALIZAÇÃO ESCALADA DO PAGERANK

note que a regra escalada também **preserva o PageRank total** na rede

é possível provar que aplicar repetidamente (k tendendo ao infinito) a regra escalada de atualização faz os valores de PageRank **convergirem para limites**

LIMITES DA REGRA DE ATUALIZAÇÃO ESCALADA DO PAGERANK

além disso, o **equilíbrio é único** (obviamente, em função de **s**)

- **na prática**, a versão de atualização escalada é usada, com o fator de escala **s** escolhido entre ***0,8*** e ***0,9***

essa versão do PageRank também se mostra **menos sensível** a adição e remoção de links ou nodos

ANÁLISE DE LINKS NA BUSCA WEB ATUAL

os dois algoritmos que vimos desempenham **papel fundamental** nas funções de ranking do Google, Yahoo! e Bing

é difícil dizer como são as funções de ranking usadas **hoje**:

- elas são muito mais complexas
- evoluem o tempo todo
- são mantidas em segredo! (há boas razões para isso, veremos depois)

ANÁLISE DE LINKS NA BUSCA WEB ATUAL

em particular, o **PageRank** sempre foi o **componente central** do **Google**:

- **dizem** que está em decadência
- **dizem** que foi misturado com um método totalmente diferente chamado “Hilltop”
- **dizem** que também foi misturado com ideias do Hubs e Autoridades

COMBINANDO LINKS, TEXTOS E DADOS SOBRE USO

na prática, para obter resultados melhores, existem melhorias

por exemplo, considerar o **conteúdo textual** dos **textos de âncora** pode ajudar

- **ex:** eu sou um estudante da UFPR
- **ex:** dentre vários lugares, esse, esse e esse

a primeira frase tem muito mais "significado" se estamos fazendo uma busca por **"UFPR"** (não basta ter um link para www.ufpr.br)

COMBINANDO LINKS, TEXTOS E DADOS SOBRE USO

estender pageRank e hubs-autoridades: basta **multiplicar por um fator** que indica a qualidade do texto

é possível também usar dados sobre o uso dos links

- **uma busca por "UFPR":** se os usuários clicam muito **mais vezes na 2ª opção retornada,** então devemos considerar isso

SEARCH ENGINE OPTIMIZATION (SEO)

teoria dos Jogos: o mundo reage e se adapta às regras

muitas empresas querem aparecer no topo das buscas

empresas grandes podem perder milhões após o Google atualizar suas regras de ranking

assim, autores de páginas Web **criam suas páginas** Web com a fórmula de ranking do mecanismo de busca em mente

Cliff Linch: "a busca na Web é um novo tipo de aplicação de recuperação de informação em que **os documentos estão a toda hora se comportando mal**"

SEARCH ENGINE OPTIMIZATION (SEO)

surgiu o SEO (search engine optimization): regras e guias gerais de **como projetar páginas que sejam rankeadas bem**

muitas empresas grandes acabaram **contratando experts de SEO**, mas...

voltando à teoria dos jogos: tanta gente estava interessada nas técnicas de SEO, que as **técnicas de SEO** começaram a ser **facilmente encontradas nos próprios mecanismos de busca**!

UM ALVO EM MOVIMENTO

esses avanços tiveram **sérias consequências**:

1) do ponto de vista do motor de busca: a função de ranking perfeita será sempre um **alvo em movimento**

   - **se ela manter um modo de ranking "perfeito" por muito tempo**: as pessoas fazem **engenharia reversa no método** e podem **se posicionar onde quiserem** do ranking

2) as empresas de busca na Web **mantém guardado a 7 chaves** sua função de ranking

   - não somente para empresas concorrentes, **mas principalmente para webdesigners**

UM ALVO EM MOVIMENTO: UM NEGÓCIO LUCRATIVO

as empresas de busca na Web conseguiram **transformar o pesadelo em dinheiro**

criaram **espaços reservados** no resultado da busca onde as empresas podem **pagar para colocar seu link**

as pessoas acham **mais fácil pagar o anúncio** do que recorrer a técnicas avançadas (e possivelmente defasadas) de SEO

desta forma, as empresas de busca na Web:

- **diminuem a prática de SEO** (que é ruim para elas)
- **ganham MUITO dinheiro**