Classificação de Conteúdo Web
Estudo e Aplicações

Caio de Moraes Braz & Gustavo Perez Katague

Supervisor: José Coelho de Pina

Instituto de Matemática e Estatística - Universidade de São Paulo

13 de novembro de 2012

A Internet

A Internet

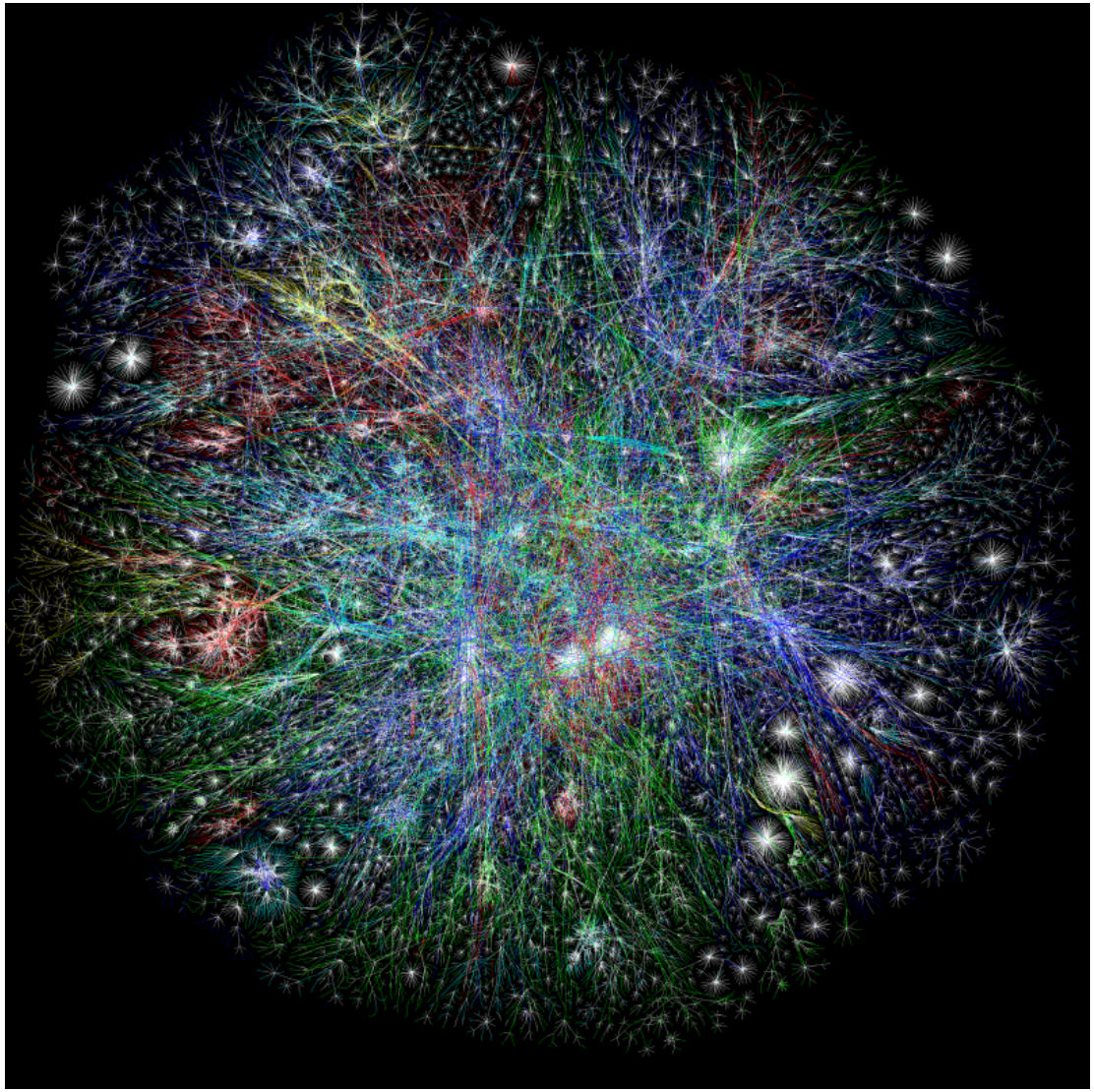


Imagem: www.opte.org/maps/

Contexto

- Internet

  Muita informação, pouca organização

- Mecanismos de busca

  Surgiram junto com o crescimento da internet

  Ajudam a procurar conteúdo relevante

PageRank

- Quem? Onde? Quando?
- Como funciona?

PageRank

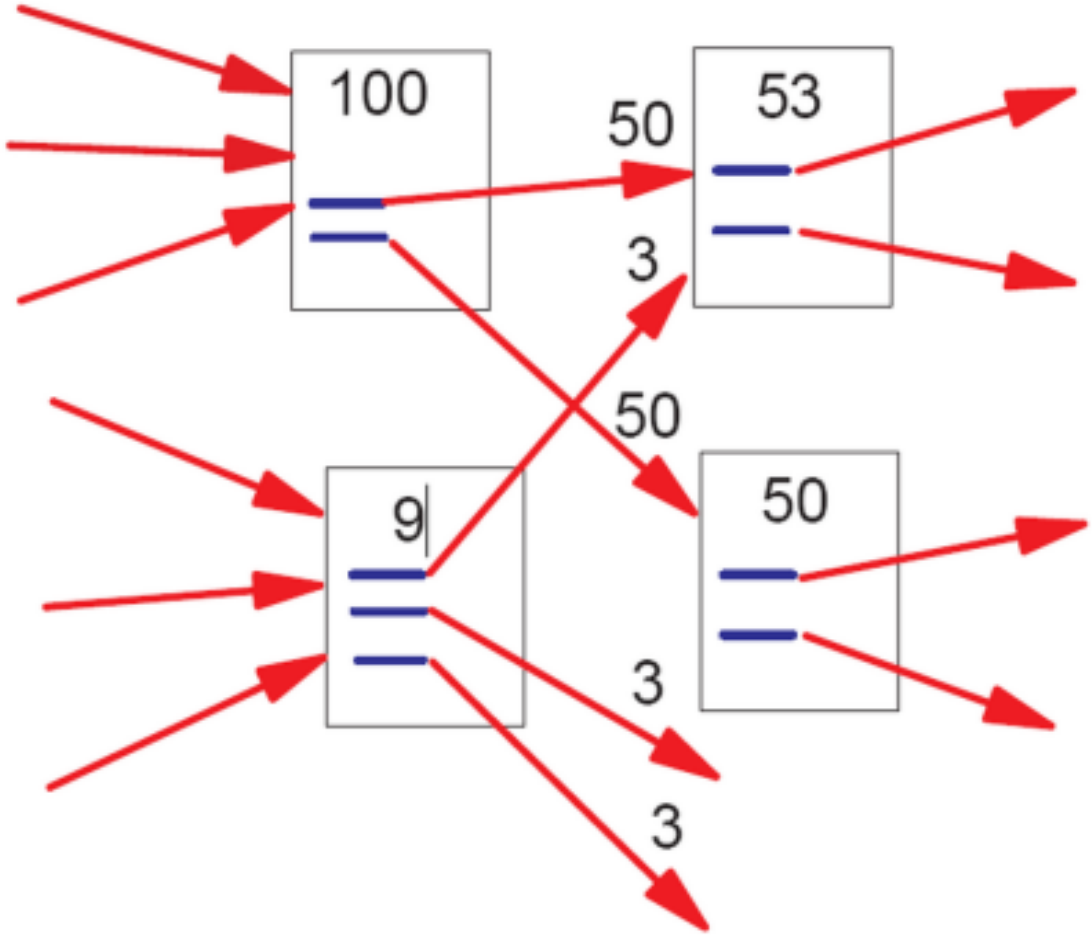


Imagem: http://www.seobook.com/images/pagerank-flow.png

PageRank

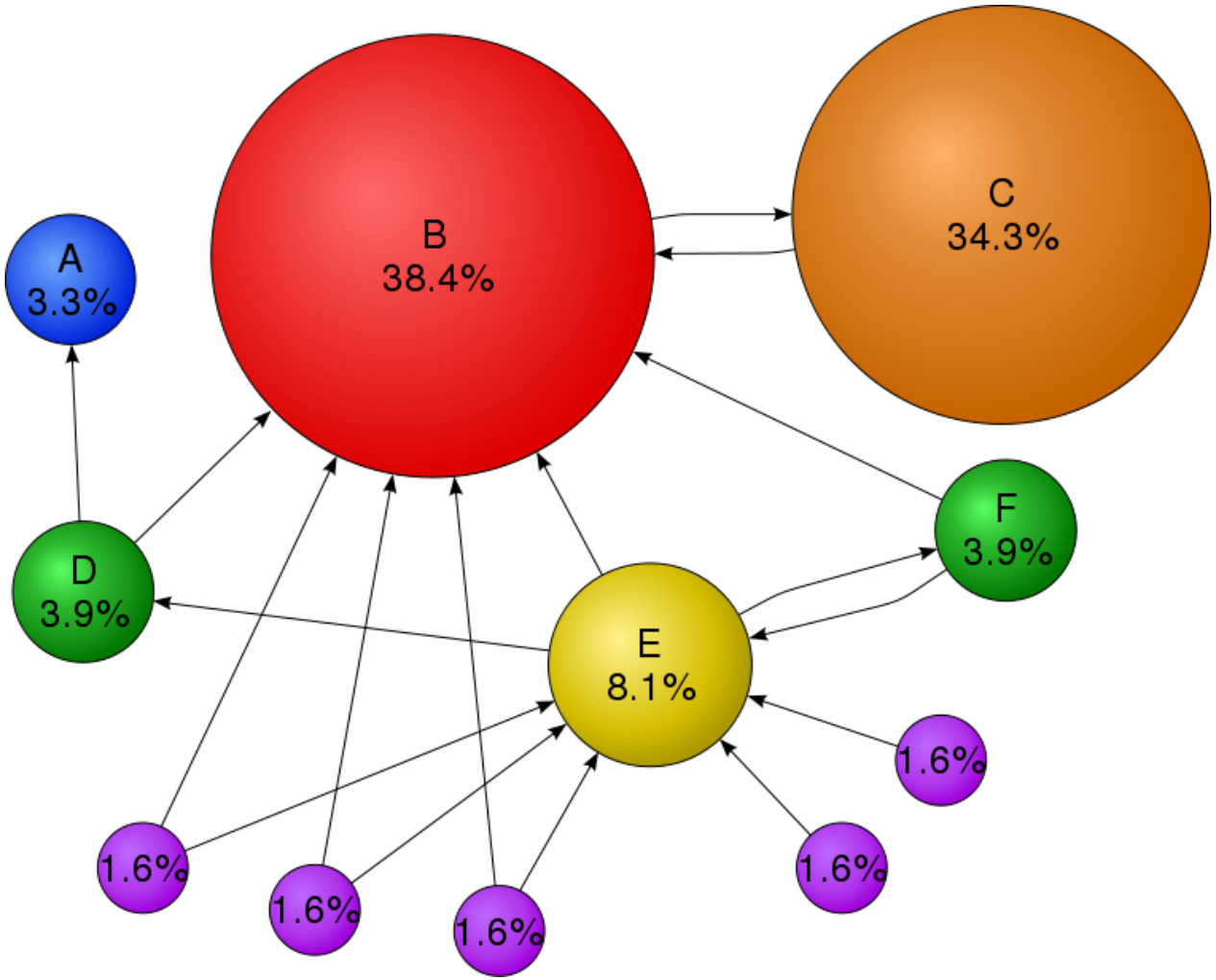
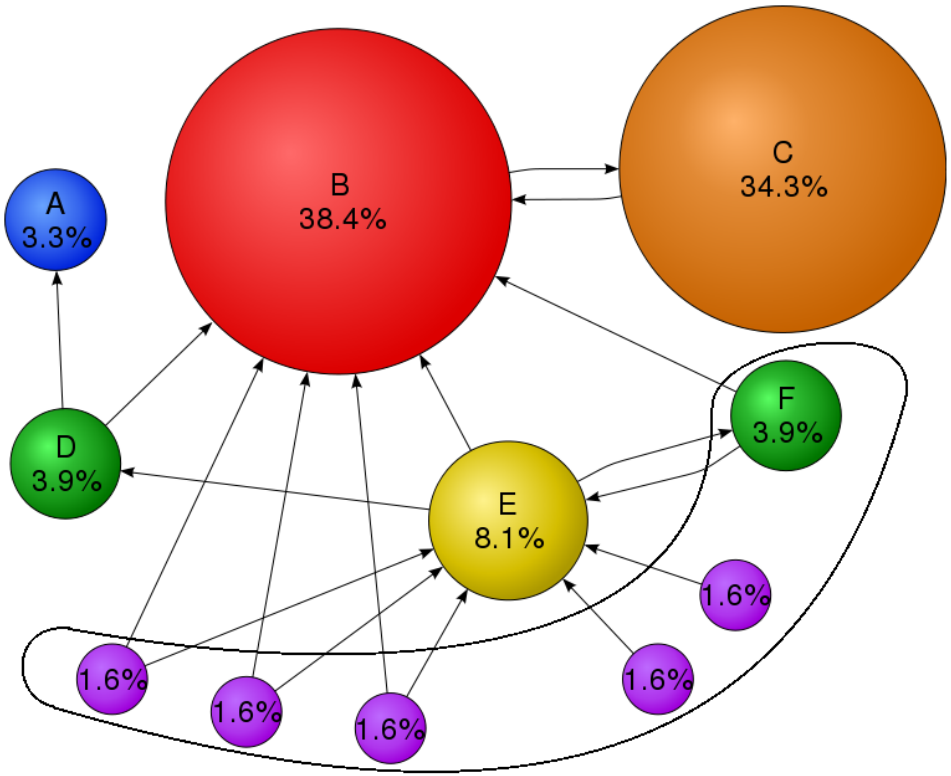


Imagem: upload.wikimedia.org/wikipedia/commons/f/fb/PageRanks-Example.svg

PageRank - Definição

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|P_j|}$$



Modelagem da Web

Páginas P_i e P_j

Matriz de adjacência L

$$L_{ij} = \begin{cases} 1, & \text{se existe um } \textit{link} \text{ de } P_i \text{ a } P_j \\ 0, & \text{caso contrário} \end{cases}$$

PageRank - Abstração

Random Surfer

- Navega *aleatoriamente* pela web

PageRank - Modelagem - H

Primeira representação matricial: H

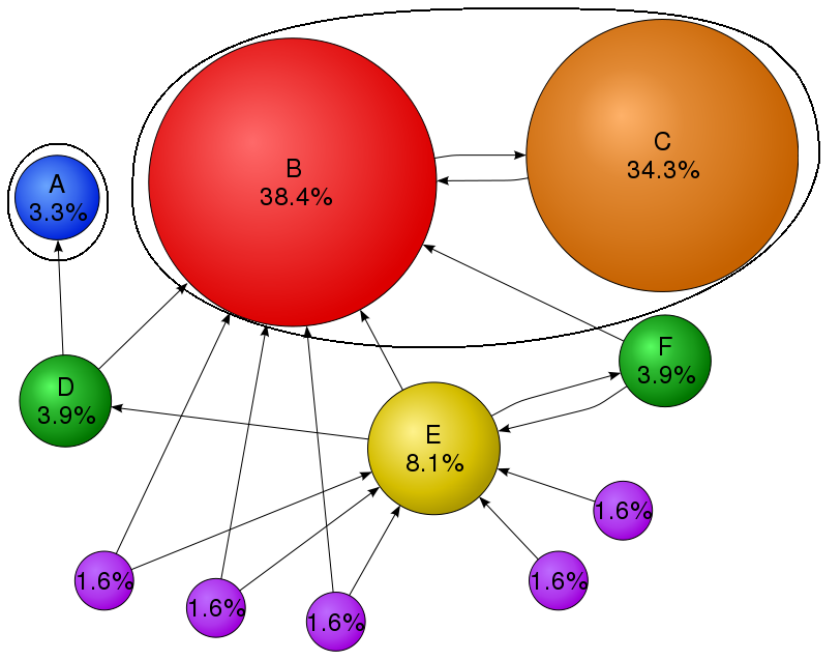
$$H_{ij} = \begin{cases} 1/|P_i|, & \text{se existe um arco ligando } i \text{ a } j \\ 0, & \text{caso contrário} \end{cases}$$

- Adaptação da matriz L
- H é semi-estocástica

PageRank - Modelagem - Problemas

Rank sinks

- *Dangling nodes*
- Ciclos



PageRank - Abstração

Random Surfer

- Navega *aleatoriamente* pela web
- Ao chegar em uma *página sem links*, escolhe outra aleatoriamente para continuar o processo

PageRank - Modelagem - S

Segunda representação matricial: S

$$S = H + (1/n)ae'$$

$$a_i = \begin{cases} 1, & \text{se a } i\text{-ésima linha de } H \text{ for nula} \\ 0, & \text{caso contrário} \end{cases}$$

- *Dangling links* apontam indiretamente para todas as outras páginas
- S é estocástica

PageRank - Abstração

Random Surfer

- Navega *aleatoriamente* pela web
- Ao chegar em uma *página sem links*, escolhe outra aleatoriamente para continuar o processo
- A qualquer momento, com probalilidade $1 - \alpha$ ele pode *"cansar"* de seguir a estrutura de links e escolher uma outra página

PageRank - Modelagem - G

Terceira representação matricial: G

$$G = \alpha S + (1 - \alpha)(1/n)ee'$$

- α é a probabilidade de seguir a estrutura original de links
- G é aperiódica e irredutível

PageRank - Modelagem - π

Finalmente:

$$\pi' = \pi' G$$

- Abstraindo G para uma matriz de transição de uma cadeia de Markov
- π é a distribuição estacionária de G
- π_i corresponde ao *rank* da página P_i

PageRank - Implementação

- Método da potência
- Iteramos até que: $\left|\left|\pi_k - \pi_{k-1}\right|\right|_1 \leqslant \epsilon$

PageRank - Implementação

$res \leftarrow 1$

$pi \leftarrow pi_0$

while $res \geqslant \epsilon$ **do**

$prev\_pi \leftarrow pi$

$pi' \leftarrow pi' * G$

$res \leftarrow ||pi - prev\_pi||_1$

end while

PageRank - Desafios

- Estruturas de dados
- Problemas de escala
- Problemas de precisão numérica

PageRank - Sensibilidade

- Parâmetro α

  Influencia diretamente a convergência

- Matriz H

  Quanto mais próximo α fica de 1, mais sensível

  π fica a pertubações em H

PageRank - Número de Iterações

α	0,5	0,6	0,7	0,8	0,9	0,95
V = 2293 A = 9644	14	19	27	43	88	173
V = 4298 A = 21956	15	20	28	42	83	167
V = 4334 A = 17424	15	20	29	46	95	192
V = 5354 A = 24389	15	21	29	46	97	196
V = 7399 A = 36121	16	21	30	48	100	203
V = 8011 A = 34672	15	19	27	43	90	183

PageRank - Número de Iterações

α	0,99	0,999	0,9999	0,99999
V = 2293 A = 9644	800	8.017	80.202	802.052
V = 4298 A = 21956	802	8.007	80.104	801.068
V = 4334 A = 17424	971	9.745	97.494	974.975
V = 5354 A = 24389	993	9.966	99.705	997.092
V = 7399 A = 36121	1.032	10.362	103.660	1.036.641
V = 8011 A = 34672	910	9.135	91.390	913.940

OBRIGADO!