

Google PageRank: matemática básica e métodos numéricos

O PageRank tem entrado progressivamente no nosso dia-a-dia através do motor de busca mais usado actualmente: o Google.

Mas, ...

O que significa?

É baseado em algum modelo matemático?

Como se calcula?

Qual é o PageRank da minha página?
e deste centro de investigação?

Com este trabalho pretende-se dar, de forma simples, algumas respostas a estas questões.

Paulo Vasconcelos - CMUP

Sumário

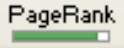

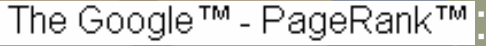



- PageRank
 - Motivação
 - O que é?
 - Como posso saber o PageRank de uma página?
 - Porque se trabalha e investiga sobre o PageRank?
 - O algoritmo original
 - Exemplos
 - Como simular ...
 - O algoritmo e Álgebra Linear Numérica
 - Sistema de equações lineares
 - Problema de valores próprios
 - Matriz Google
- Algumas referências

Motivação

- Uma pesquisa através do Google (www.google.pt) retorna uma quantidade incrível de páginas.
 - Por exemplo, em 4/1/06, uma procura com
 - “cmup” retornou 59.600 páginas e com
 - “porto” 58.700.000 páginas.
- No entanto, as páginas mais relevantes geralmente surgem entre as 10 ou 20 primeiras.
 - Por exemplo, da procura com “porto” as 3 primeiras páginas foram:
 - FC Porto: <http://www.fcporto.pt/>
 - CM Porto: <http://www.cm-porto.pt/>
 - UP - Universidade do Porto: <http://www.up.pt/>
 - por esta ordem ...; com o reencaminhamento do endereço www.up.pt para http://sigarra.up.pt/up/web_page.inicial a UP perdeu em posição relativa. Com o tempo estas posições podem ser alteradas
- Como é que o motor de pesquisa “sabe” quais as páginas mais importantes?
- O Google atribui um número a cada página por forma a reflectir essa importância. Esse número é o PageRank.



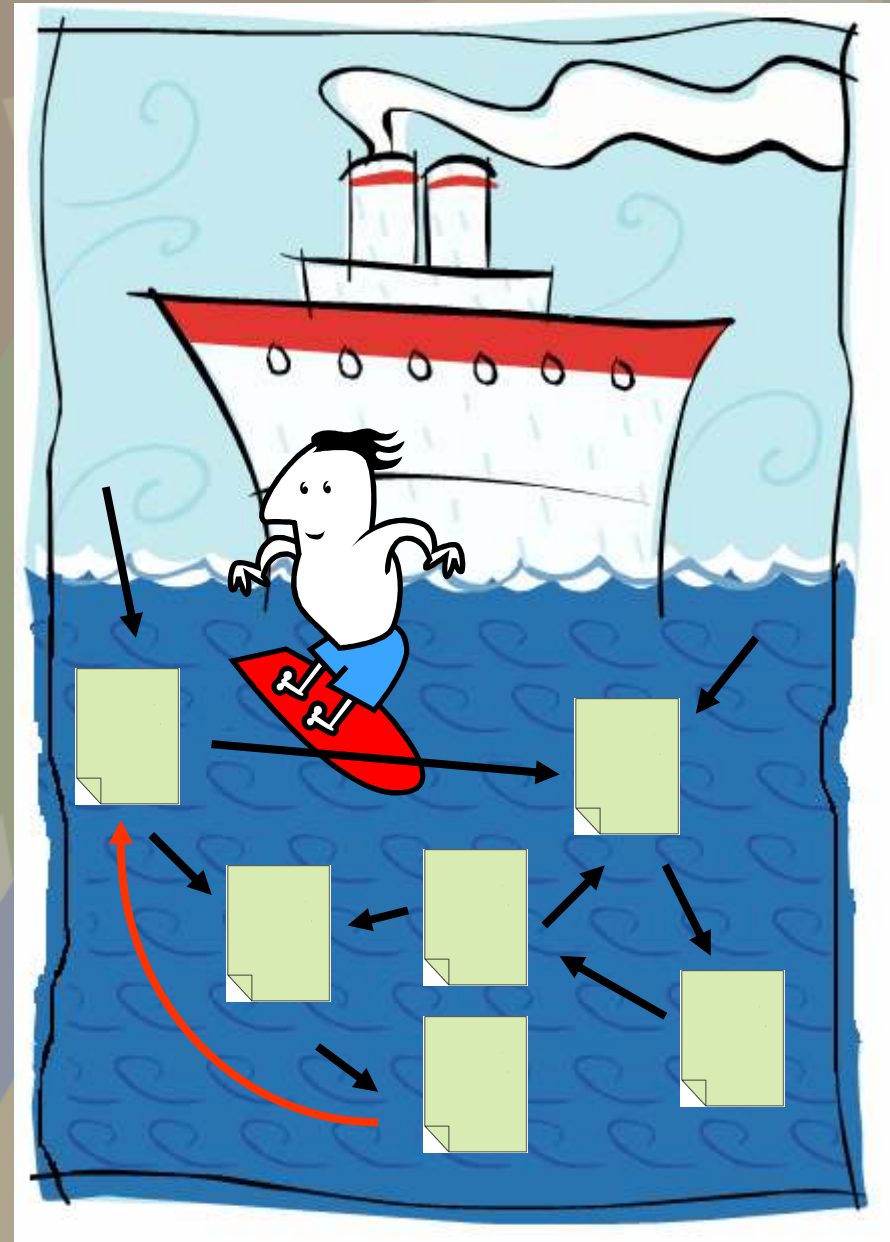
Como posso saber o PageRank (PR) de uma página? O Toolbar PageRank

- O Google toolbar PageRank dá uma medida do PageRank de uma página, em geral numa representação por gráfico de barra (0-10) com este aspecto: 
- Para instalar o Google toolbar PageRank  vá a <http://toolbar.google.com/>
 - além da medida do PageRank faculta outras funcionalidades
 - mas atenção: há informação que é enviada para os servidores da Google.
- Pode usar a facilidade da Google para calcular o toolbar PageRank : <http://www.mygooglepagerank.com/pagerank.php>
- Outros “sites” possíveis:
 -  : <http://pr.blogflux.com/index2.php>
 -  : http://www.prchecker.info/check_page_rank.php
 -  : http://www.search-this.com/pagerank_decoder/
- O Google toolbar PageRank traduz o PageRank numa escala logarítmica, pelo que aumentar o PR de uma página de 5 para 6 é muito mais difícil que passar de 2 para 3.

O que é o PageRank?

- Como já sabemos, quando é colocada uma pergunta (“query”), um motor de busca procura as páginas na “web” que satisfazem a “query” e apresenta-as ordenadas pelo seu PageRank
- Quando o Google foi lançado já existiam vários motores de busca, sendo o seu êxito devido ao PageRank (responsável por essa ordenação)
- O Google PageRank™ é um método que classifica documentos da “web” por um número
- Foi desenvolvido por
 - Larry Page (<http://www.google.com/corporate/execs.html#larry>) e
 - Sergey Brin (<http://www.google.com/corporate/execs.html#sergey>) enquanto alunos na Univ. Stanford com 24 e 23 anos resp.; foram os fundadores da Google
- O PageRank é baseado na estrutura de ligação da “web”; a classificação dada a um documento é dada pela classificação dos documentos que a ele ligam, sendo pois obtida recursivamente pelo PageRank desses documentos
- O Google explica o PageRank como um processo democrático que interpreta um “link” da página A para a B como um voto

- A ideia do PageRank é muitas vezes apresentada como a de um surfista a navegar na “web” ...
- O surfista vai de página em página escolhendo aleatoriamente um “link” de saída
- Pode, no entanto, numa página não encontrar saída por esta não existir (“dangling links”) ou por entrar em ciclo num conjunto de páginas interligadas
- Para evitar isto, é necessário criar um mecanismo para saltar para uma página sem seguir um “link” existente nessa página



Porque se trabalha e investiga sobre o PageRank?

- O Page Rank pode ser observado de 2 perspectivas:
 - a do utilizador:
 - o que se pode fazer para melhorar o PageRank de páginas?
 - importante para a visibilidade de uma instituição, fundamental para o trabalho de um “webmaster”
 - a do Google:
 - como calcular o PageRank?
- Para se trabalhar ambas as perspectivas é necessário modelar matematicamente as ideias subjacentes ao PageRank e desenvolver algoritmos para o calcular

O algoritmo PageRank

- O PageRank (PR) original descreve-se por

$$PR(P_0) = p \times \sum_{i=1}^n \frac{PR(P_i)}{c(P_i)} + \frac{1-p}{n} \quad (1)$$

- $PR(P_0)$ é o PageRank da página P_0
- $PR(P_i)$ é o PageRank da página P_i que liga à página P_0
- $c(P_i)$ é o número de ligações de saída (“outbound links” ou “out-degree”) na página P_i ; $c(P_i)$ deve ser $\neq 0$
- n é o número de páginas
- p é um factor (“dumping”), $0 < p < 1$; tanto quanto se julga saber o Google usa $p=0.85$

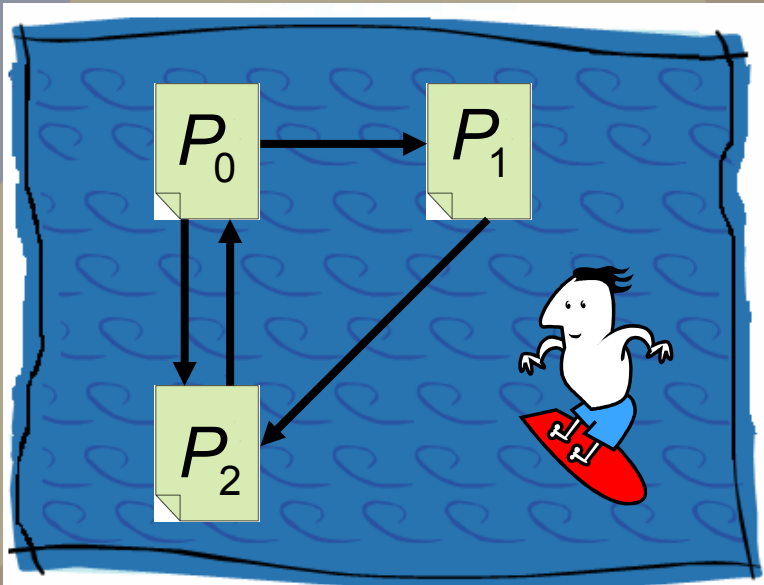
- O PageRank forma uma distribuição de probabilidade sobre as páginas “web”, sendo que a soma dos PageRank de todas as páginas da web é 1
- Na verdade no seu paper original “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, Lawrence Page e Sergey Brin definiram PageRank por

$$PR(P_0) = p \times \sum_{i=1}^n \frac{PR(P_i)}{c(P_i)} + (1 - p) \quad (2)$$

- Nesta versão a probabilidade de um passeio aleatório alcançar a página é ponderada pelo número total de páginas “web”
- Neste caso, a soma dos PageRank de todas as páginas da web é n
- Esta variante é muitas vezes usada para ilustrar o modo de funcionamento do algoritmo pois não necessita do valor de n ; usando a fórmula (2) vamos ver 2 exemplos:

Exemplo 1

- Seja uma pequena “web” formada por 3 páginas



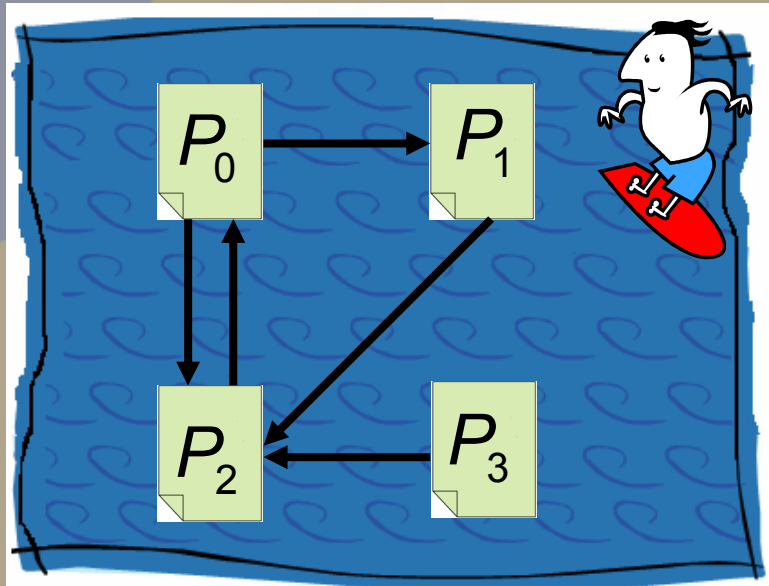
$$\begin{cases} PR(P_0) = 0.85 \cdot PR(P_2) + 0.15 \\ PR(P_1) = 0.85 \cdot PR(P_0 / 2) + 0.15 \\ PR(P_2) = 0.85 \cdot (PR(P_0 / 2) + PR(P_1)) + 0.15 \end{cases}$$

$$\begin{cases} PR(P_0) = 1.1634 \\ PR(P_1) = 0.6444 \\ PR(P_2) = 1.1922 \end{cases}$$

- A soma dos PageRank é 3
- O maior PageRank é da página P_2 pois é a única apontada por 2 páginas (a menos da ponderação)
- Entre P_0 e P_1 o maior PageRank é de P_0 pois é apontada pela página com maior PageRank

Exemplo 2

- Acrescentemos uma 4ª página



$$\begin{cases} PR(P_0) = 0.85 \cdot PR(P_2) + 0.15 \\ PR(P_1) = 0.85 \cdot PR(P_0 / 2) + 0.15 \\ PR(P_2) = 0.85 \cdot (PR(P_0 / 2) + PR(P_1) + PR(P_3)) + 0.15 \\ PR(P_3) = 0.85 \cdot (0) + 0.15 \end{cases}$$

$$\begin{cases} PR(P_0) = 1.4901 \\ PR(P_1) = 0.7833 \\ PR(P_2) = 1.5766 \\ PR(P_3) = 0.1500 \end{cases}$$

- A soma dos PageRank é agora 4
- Os PageRank aumentaram por influência de P_3 mas mantiveram a posição relativa
- O PageRank de P_3 é 0.15, i.e, a probabilidade de ser escolhido ao acaso dado que não tem ligações para si

Como simular ...

- Os sistemas lineares anteriores podem ser resolvidos iterativamente dado valores iniciais aos $PR(P_i)$, $i=1,\dots,n$
- Para simular os casos anteriores ou para criar novas situações pode recorrer a esta calculadora de PageRank



Google's PageRank - Calculator

http://www.webworkshop.net/pagerank_calculator.php

- a matriz de adjacências W neste “site” é a transposta da matriz a definir em breve neste documento
- pode também calcular o PageRank de páginas concretas: para isso basta seleccionar “name pages” no canto inferior esquerdo da calculadora

O algoritmo e Álgebra Linear Numérica

- Vamos supor que existem n páginas e que W , $n \times n$, representa a matriz de adjacências correspondente ao grafo dirigido: $w(i,j)=1$ se existe ligação para a página i a partir da página j e $w(i,j)=0$ caso contrário
- Por (1) o PageRank da página P_i , agora a designar por $x(i)$, é a componente i do vector x dado por

$$x = pWDx + \frac{1-p}{n}e \quad (3)$$

- x , $n \times 1$, é o vector PageRank (normalizado por $e^T x = 1$)
- D , $n \times n$, é a matriz diagonal dos inversos dos “out-degree”
- e , $n \times 1$, é um vector de uns ($e = (1, \dots, 1)^T$)
- Se o surfista seguir um “link” com prob. p e saltar para uma página aleatória com probabilidade $1-p$, então x_i pode ser interpretado como a probabilidade do surfista estar na página P_i

Sistema de equações lineares

- O sistema (3) é equivalente a

$$(I - pWD)x = \frac{1-p}{n}e \quad (4)$$

- A solução de (4) pode ser obtida
 - por um método directo:
 - eliminação de Gauss
 - por um método iterativo:
 - Jacobi
 - Gauss-Seidel
 - métodos baseados em subespaços de Krylov (GMRES, BiCGstab, ...)

Problema de valores próprios

- O sistema (3) é equivalente a

$$x = Gx, \quad G = \left(pWD + (1-p)\frac{ee^T}{n} \right), \quad e^T x = 1 \quad (5)$$

- A matriz $\frac{ee^T}{n}$ é designada de “teleportation”
- A solução de (5), x , pode ser obtida pelo método da potência, dado que se procura o vector próprio associado ao maior valor próprio em magnitude, valor próprio 1
- O vector próprio x tem todas as entradas não negativas e é o único vector próprio com esta propriedade (tal resulta do teorema de Perron-Frobenius)
- Este método em geral funciona bem pois a sua convergência depende do rácio entre o 2º maior valor próprio em magnitude e o 1º; provou-se que o 2º valor próprio é sempre igual a p , e logo $p=0.85$ é suficientemente afastado de 1 para este método numérico convergir; maiores dificuldades ocorrem para maiores valores de p

Matriz Google

- Em 4/1/06 a matriz G representava a estrutura de ligações da “web” com cerca de 8×10^9 páginas
 - A matriz G tem pois enormes dimensões: $8 \times 10^9 \times 8 \times 10^9$ e é uma matriz densa
 - Por oposição WD é esparsa
- O seu armazenamento e processamento requer máquinas com grande capacidade de memória e com elevada capacidade computacional - supercomputadores- máquinas de processamento paralelo
- Isto tem implicação também nos métodos numéricos a usar; por exemplo, a sua dimensão inviabiliza o uso de métodos directos e aconselha métodos iterativos (até porque não se necessita de resultados com grande precisão)

Matriz Google

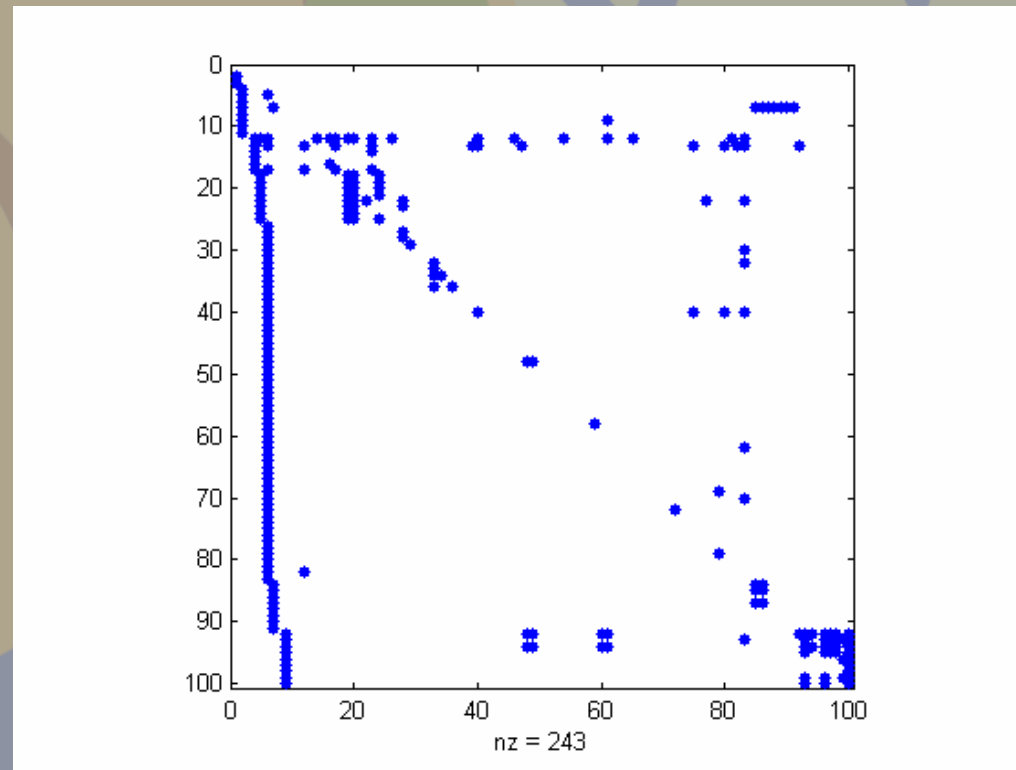
- A Google actualiza o PageRank das páginas 1 vez por mês
- Muitos trabalhos têm sido publicados sobre como calcular o vector PageRank
- Uma solução possível e simples de explicar é:
 - Pensemos num método iterativo.
 - Neste tipos de métodos não temos de conhecer a matriz Google mas apenas o resultado da sua acção sobre vectores: multiplicação matriz-vector
 - De (3) note-se que:
 - 1ª parcela: o produto $p(W(Dx))$ é barato de calcular pois quer W quer D são matrizes esparsas (pode poupar-se em armazenamento e em custo computacional operando só as entradas não nulas)
 - 2ª parcela: basta calcular uma componente do vector $(1-p)e/n$ pois as restantes são iguais; só é necessário calcular uma vez durante o processo iterativo atendendo à normalização $e^T x = 1$

Como gerar amostras “web”?

- Para melhor perceber o PageRank, em vez de imaginar uma pequena “web” podemos trabalhar sobre uma amostra de toda a “web”
- Por exemplo usando o MATLAB, podemos fazer apelo às funções:
 - **function [U,W] = surfer(root,n)**
cria o grafo de adjacências de uma porção da Web com n páginas a partir do URL root
 - **x = pagerank(U,W,p)**
usa os URLs e a matriz de adjacência dados pela função anterior para calcular o PageRank, $p=0.85$
 - **spy(W)**
para visualizar o padrão de esparsidade da matriz de adjacências

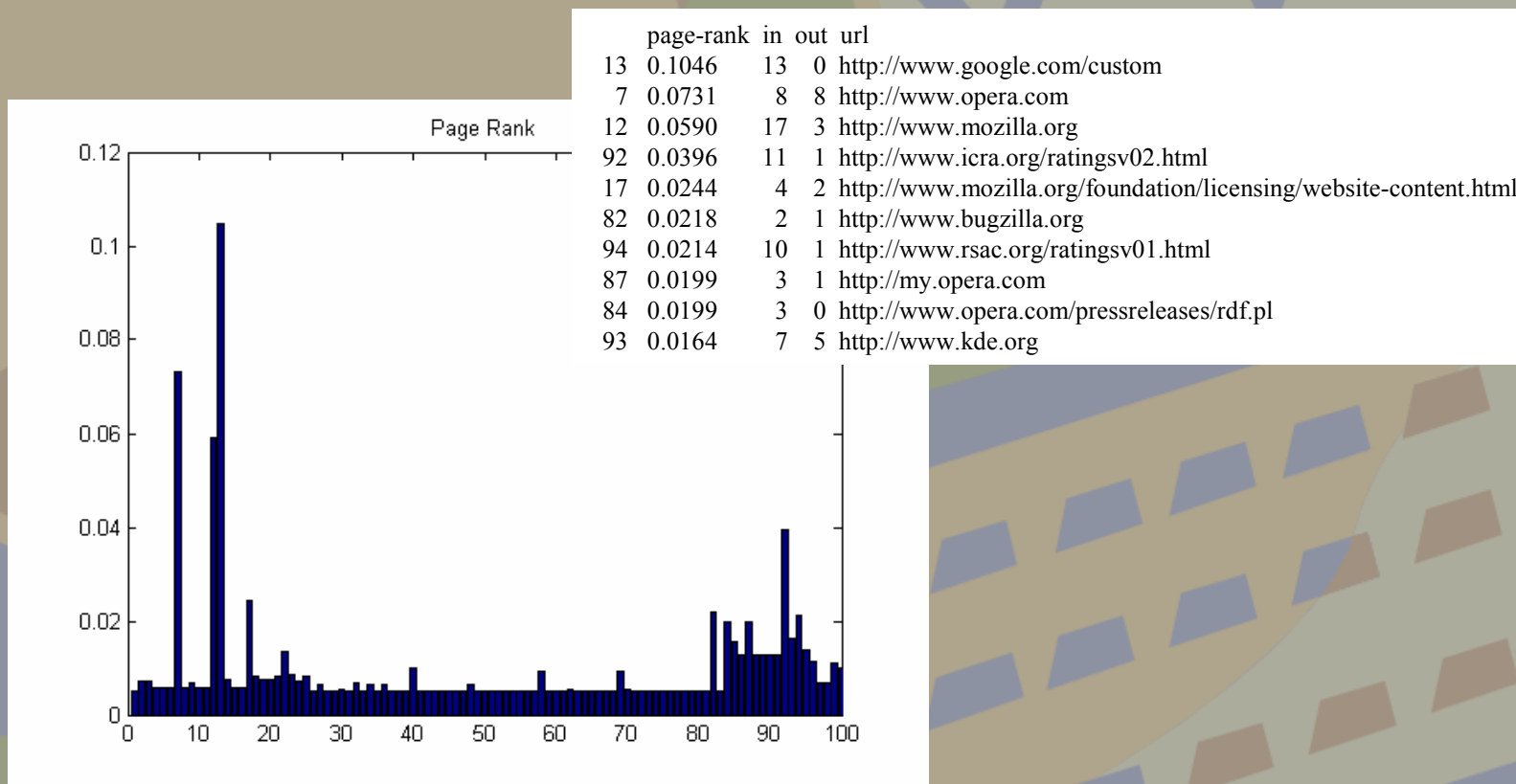
Matriz de adjacências e PageRank de uma amostra com 100 páginas da "web" elaborada a partir de <http://www.fc.up.pt>

- Padrão de esparsidade da matriz W
 - nz representa o número de elementos não nulos na matriz: 243 elementos não nulos numa matriz com 10000 entradas



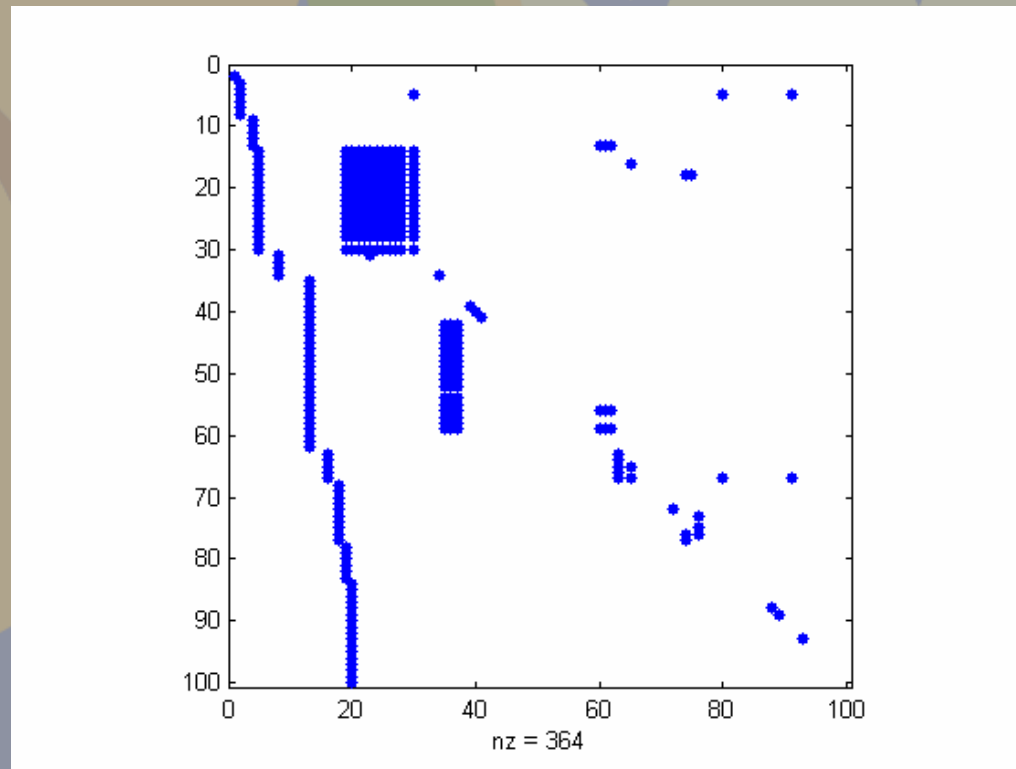
Matriz de adjacências e PageRank de uma amostra com 100 páginas da “web” elaborada a partir de <http://www.fc.up.pt>

- Gráfico dos PageRank e discriminação das 10 páginas com maior PageRank



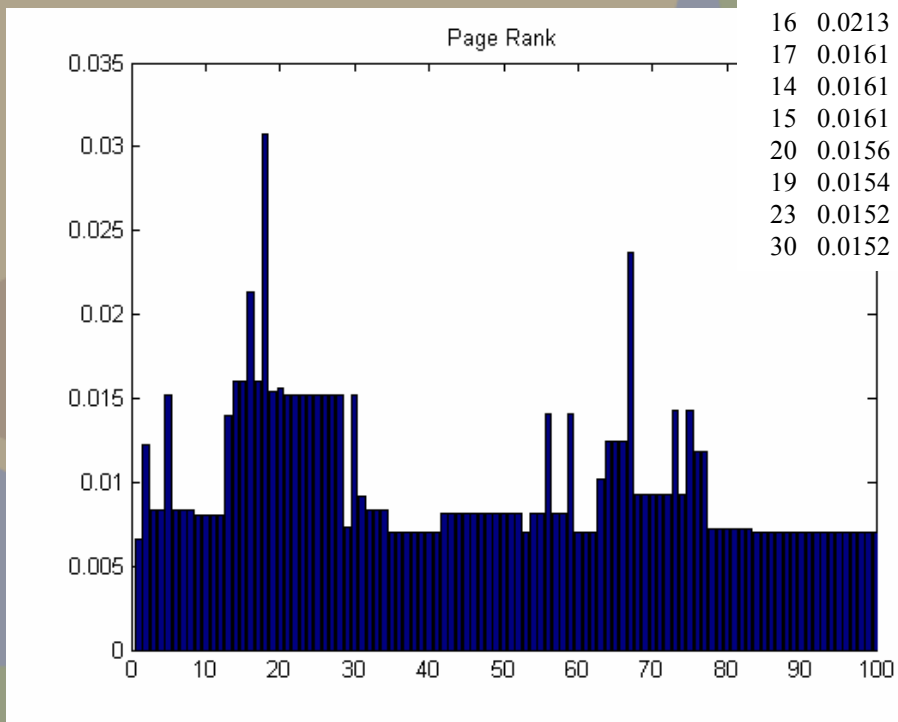
Matriz de adjacências e PageRank de uma amostra com 100 páginas da “web” elaborada a partir de <http://www.up.pt>

- Padrão de esparsidade da matriz W
 - nz representa o número de elementos não nulos na matriz: 364 elementos não nulos numa matriz com 10000 entradas



Matriz de adjacências e PageRank de uma amostra com 100 páginas da “web” elaborada a partir de <http://www.up.pt>

- Gráfico dos PageRank e discriminação das 10 páginas com maior PageRank



page-rank	in	out	url
18	0.0308	14	10 http://www.b-on.pt
67	0.0237	5	0 http://www.elsevier.com
16	0.0213	13	5 http://www.scopus.com
17	0.0161	12	0 http://isi15.isiknowledge.com/portal.cgi
14	0.0161	12	0 http://www.up.pt/ ;
15	0.0161	12	0 http://minerva.up.pt
20	0.0156	11	32 http://biblioteca.up.pt/paginas/periodicos.shtml
19	0.0154	11	21 http://biblioteca.up.pt/paginas/bases.shtml
23	0.0152	11	16 http://biblioteca.up.pt/paginas/catalogo.shtml
30	0.0152	11	16 http://biblioteca.up.pt/paginas/contactos.shtml

Algumas referências

- *Gerais:*
 - Uma brev(íssima) explicação da Google: *Google Technology* (<http://www.google.com/technology/>)
 - O paper original: *The Anatomy of a Large-Scale Hypertextual Web Search Engine* (<http://www-db.stanford.edu/~backrub/google.html>), Lawrence Page e Sergey Brin
 - Uma explicação breve: *Google's PageRank Explained and how to make the most of it* (<http://www.webworkshop.net/pagerank.html>), Phil Craven
 - Uma explicação nem breve nem longa: *The Google Pagerank Algorithm and How It Works* (<http://www.iprcom.com/papers/pagerank/>), Ian Rogers
 - Uma explicação longa: *A Survey of Google's PageRank* (<http://pr.efactory.de/>)
- *Mais técnicas:*
 - Sobre a criação de matrizes tipo PageRank em pequenas “web”: *The World's Largest Matrix Computation* (<http://www.matesco.unican.es/aplicaciones/Cleve's%20Corner%20The%20World%E2%80%99s%20Largest%20Matrix%20Computation.htm>), Cleve Moler
 - Sobre o cálculo em máquinas paralelas: *Decomposition of the Google PageRank and Optimal Linking Strategy* (<http://wwwhome.math.utwente.nl/~litvakn/1712.pdf>), Konstatin Avrachenov e Nelly Litvak
 - Sobre métodos para resolver numericamente problemas do tipo do PageRank: *Arnoldi-type algorithms for computing stationary distribution vectors, with application to PageRank* (<http://www.cs.ubc.ca/~greif/Papers/gg2004.pdf>), Gene H. Golub and Chen Greif
 - Sobre o 2º valor próprio: *The Second Eigenvalue of the Google Matrix* (<http://dbpubs.stanford.edu:8090/pub/showDoc.Fulltext?lang=en&doc=2003-20&format=pdf&compression=&name=2003-20.pdf>), Taher H. Haveliwala and Sepandar D. Kamvar
 - Sobre uma outra interpretação para o PageRank: *Google PageRank as mean playing time for pinball on the web*, D. Higham, AML, vol 18, 1359-1362, 2005