Matematica nel Web: l'esempio di Google

Prof. S. Serra Capizzano

Dipartimento di Scienza e Alta Tecnologia
Università dell'Insubria

La matematica è intorno a noi

- Senza la matematica molte delle tecnologie che utilizziamo ogni giorno non sarebbero disponibili.
- Alla base della tecnologia avanzata c'è sempre della buona matematica.
- La matematica ci circonda dalle applicazioni quotidiane più comuni (fotografia digitale, video, gps, crittografia, INTERNET, etc.) a quelle più "sofisticate" (TAC, previsioni del tempo, modelli economici, armi "intelligenti" etc.).

Internet e la matematica

La grande mole di dati presente in Internet è sorgente di molti problemi matematici di particolare interesse sia teorico che applicativo

- Information retrieval
- Gestione del flusso di informazioni sulla rete (sicurezza, etc.)
- Calcolo distribuito
- Motori di ricerca (Google)

Google

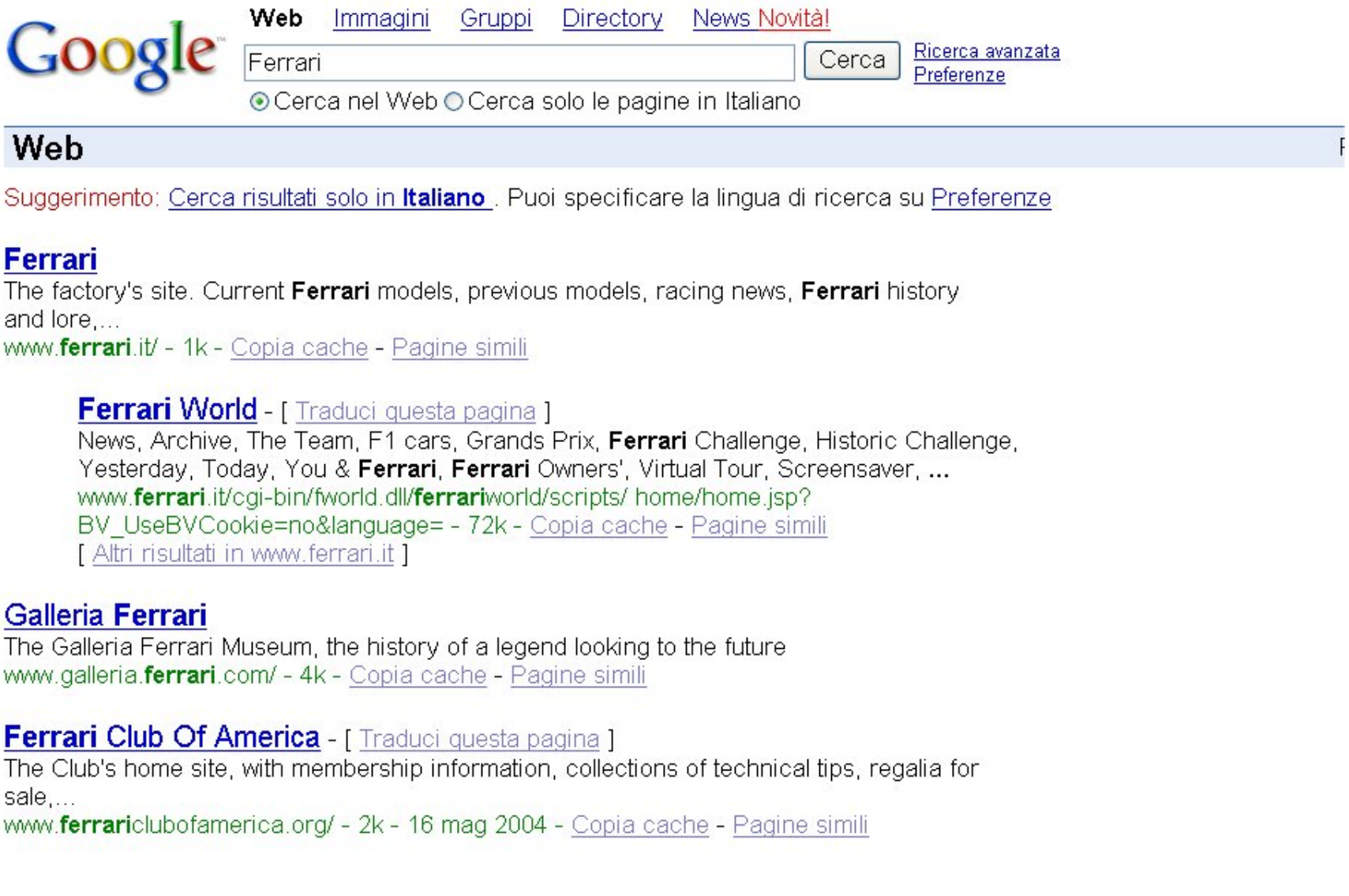
Nascita di Google

Due studenti di dottorato di Stanford, Sergey Brin e Larry Page hanno fatto una fortuna inventando Google.

Ciò che caratterizza Google rispetto ai motori di ricerca precedenti è come vengono ordinati i risultati della ricerca.

Un motore di ricerca non deve solo trovare tutte le pagine corrispondenti ai criteri di ricerca, ma deve anche ordinarle in modo tale da semplificare la scelta all'utente.

Un Esempio di ricerca



Google

Web Immagini Gruppi Directory News Novità!

Ferrari Cerca Ricerca avanzata Preferenze

Cerca nel Web Cerca solo le pagine in Italiano

Web

Suggerimento: Cerca risultati solo in **Italiano** . Puoi specificare la lingua di ricerca su Preferenze

Ferrari
The factory's site. Current **Ferrari** models, previous models, racing news, **Ferrari** history and lore,...
www.**ferrari**.it/ - 1k - Copia cache - Pagine simili

Ferrari World - [Traduci questa pagina]
News, Archive, The Team, F1 cars, Grands Prix, **Ferrari** Challenge, Historic Challenge, Yesterday, Today, You & **Ferrari**, **Ferrari** Owners', Virtual Tour, Screensaver, **...**
www.**ferrari**.it/cgi-bin/fworld.dll/**ferrari**world/scripts/ home/home.jsp?BV_UseBVCookie=no&language= - 72k - Copia cache - Pagine simili
[Altri risultati in www.ferrari.it]

Galleria Ferrari
The Galleria Ferrari Museum, the history of a legend looking to the future
www.galleria.**ferrari**.com/ - 4k - Copia cache - Pagine simili

Ferrari Club Of America - [Traduci questa pagina]
The Club's home site, with membership information, collections of technical tips, regalia for sale,...
www.**ferrari**clubofamerica.org/ - 2k - 16 mag 2004 - Copia cache - Pagine simili

Motori di ricerca

Criteri di ricerca

- Non c'è distinzione fra maiuscole e minuscole
- Si ignora: accentazione, parole "comuni" (e, per, ...)

Ordinamento dei risultati

- Non ci si limita al numero di occorrenze dei termini ricercati
- Assegna una priorità in base alla "vicinanza" dei termini ricercati
- Esamina il contenuto della pagina e delle pagine ad essa correlate
- Considera l'importanza e la qualità di una pagina nel Web: PageRank

L'importanza

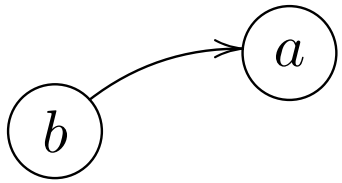
Google cerca di riprodurre sul Web il concetto comune di importanza di una persona in senso "sociale".

Nell'accezione comune una persona è importante se:

1. molte persone parlano di lei,
2. in particolare se chi parla di lei è una persona importante,
3. in particolare se chi parla di lei non è una persona che parla di tanti altri (in tal caso l'importanza viene distribuita fra tutti).

PageRank 1

Numeriamo tutte le pagine del Web da 1 a $N \approx 10^{10}$ e siano a e b due di queste pagine t.c.:



$I(a) =$ importanza della pagina a

1. $I(a)$ cresce se c'è un link (connessione) da b ad a
2. $I(a)$ cresce di più se $I(b)$ è alta
3. $I(a)$ cresce di meno se b ha molti link

PageRank 2

Calcoliamo $I(a)$

- Sia $\sharp b =$ numero di link uscenti da b
- $I(a)$ dipende dall'importanza di tutte le pagine che hanno un link verso a:

$$I(a) = \sum_{b \to a} \frac{I(b)}{\sharp b} \tag{1}$$

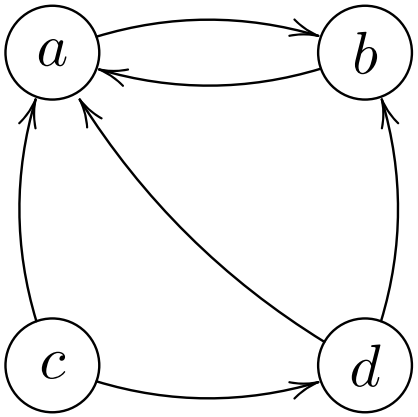
- $a = 1, \ldots, N$ e non riesco a calcolare un singolo $I(a)$ senza avere tutte le $I(b)$ tali che $b \to a$:

Si deve calcolare l'importanza di ogni pagina, contemporaneamente!

Matrice del Web

$$A(a,b) = \begin{cases} \frac{1}{\sharp b} & \text{se } b \to a \\ 0 & \text{altrimenti} \end{cases}$$

Esempio



$$\begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{bmatrix} 0 & 1 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \end{bmatrix} \end{array}$$

Prodotto matrice-vettore

- Dati

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,N} \\ a_{2,1} & a_{2,2} & \dots & a_{2,N} \\ \vdots & \dots & \dots & \vdots \\ a_{N,1} & a_{N,2} & \dots & a_{N,N} \end{bmatrix}, \quad \vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix},$$

- il prodotto matrice vettore $\vec{y} = A\vec{x}$ è definito come

$$y_i = \sum_{j=1}^{N} a_{i,j} x_j, \qquad \text{per } i = 1, \dots, N.$$

Vettore delle importanze

$$\vec{I} = [I(1), I(2), \ldots, I(N)]^T, \qquad \text{dove } I(i) \geq 0,\ i = 1, \ldots, N.$$

Osservazione

Per la definizione di importanza data nell'equazione (1), $\vec{I}$ è l'autovettore di A rispetto all'autovalore 1:

$$\vec{I} = A\vec{I}$$

ovvero

$$I(i) = \sum_{j=1}^{N} A_{i,j} I(j), \qquad \text{dove} \qquad A_{i,j} = \frac{1}{\#j} \text{ se } j \to i.$$

Studio di $\vec{I}$

- Bisogna dimostrare che $\vec{I}$ esiste (ed è unico)
- Bisogna calcolare $\vec{I}$

Problema del nostro modello

- La soluzione esiste ma non è unica
- Un utente continua a navigare senza smettere mai e seguendo solo i link della pagina corrente.

Soluzione: modificare il modello

Introdurre la possibilità per l'utente di spostarsi su una qualsiasi pagina del Web a sua scelta.

Modifica del modello

Si introducono i vettori

$$\vec{e} = [1, \ldots, 1]^T, \qquad \vec{v} = [v_1, \ldots, v_N]^T,\ v_i > 0,\ \sum_{i=1}^{N} v_i = 1,$$

$\vec{v}$ è il vettore di personalizzazione e la matrice del Web diventa

$$B = pA + (1-p)\vec{v} * \vec{e}^T = pA + (1-p) \begin{bmatrix} v_1 e_1 & \ldots & v_1 e_n \\ \vdots & & \vdots \\ v_n e_1 & \ldots & v_n e_n \end{bmatrix},$$

dove $0 < p < 1$ e

- p è la probabilità di navigare seguendo i link
- $1 - p$ è la probabilità di spostarsi su una pagina a caso nel web

Esistenza e unicità di $\vec{I}$

Si cerca sempre $\vec{I}$ tale che

$$B\vec{I} = \vec{I}.$$

Teorema

Poiché $B > 0$ ed ogni colonna di B ha somma pari a 1,

- esiste $\vec{I}$ tale che $B\vec{I} = \vec{I}$,
- sotto la condizione che $\sum_{i=1}^{N} I(i) = 1$ si ha anche l'unicità di $\vec{I}$,
- vale anche che $\vec{I}(i) \geq 0$ per $i = 1, \ldots, N$.

Calcoliamo $\vec{I}$

Viene generata una successione che converge ad $\vec{I}$:

$$\vec{I}_1, \vec{I}_2, \vec{I}_3, \ldots \longrightarrow \vec{I},$$

qualunque sia l'approssimazione iniziale $\vec{I}_0$.

Un passo dell'algoritmo

Ogni approssimazione $\vec{I}_k$ è ottenuta a partire dalla precedente ($\vec{I}_{k-1}$) mediante essenzialmente un'operazione di prodotto con la matrice B:
Per $k = 1, 2, \ldots$

$$\vec{I}_k = B\vec{I}_{k-1}$$

Costo computazionale

Si ricorda che $\vec{y} = B\vec{x}$ è tale che

$$y_i = \sum_{j=1}^{N} b_{i,j} x_j, \qquad \text{per } i = 1, \ldots, N.$$

Numero di operazioni

Per ogni $i = 1, \ldots, N$ abbiamo

$$N \text{ prodotti e } N-1 \text{ somme } = 2N-1 \text{ operazioni.}$$

In totale

$$2N^2 - N \text{ operazioni.}$$

Tempo di calcolo

- Ad oggi in Internet ci sono circa $N = 8.5 \times 10^9$ pagine attive
- Il calcolatore più veloce al mondo è il Blue Gene dell'IBM ed ha una velocità massima di 360 teraflops (1 tera = 1000 giga) cioè

$$3.6 \times 10^{14} \text{ operazioni al secondo}$$

- Per eseguire un prodotto matrice vettore, cioè calcolare una nuova approssimazione, richiederebbe

$$\begin{aligned} 2N^2 - N \;/\; 3.6 \times 10^{14} &\approx 4 \times 10^5 \text{ sec.} \\ &\approx 6.690 \text{ min.} \\ &\approx 4.5 \text{ giorni} \end{aligned}$$

Algoritmo veloce

$$B\vec{x} = pA\vec{x} + (1-p)\vec{v} * \vec{e}^T\vec{x} = pA\vec{x} + (1-p)\left(\sum_{i=1}^{N} x_i\right)\vec{v}.$$

1. $\vec{e}^T\vec{x} = \sum_{i=1}^{N} x_i$ costa $N-1$ somme.
2. Da ogni pagina del Web partono in media 7 link $\Longrightarrow$ in ogni colonna della matrice A in media ci sono solo 7 elementi diversi da zero. $\Rightarrow$ Considerando nel prodotto $A\vec{x}$ solo gli elementi non nulli questo richiede circa $13N$ operazioni.
3. Rimane poi da moltiplicare i due vettori per uno scalare e sommarli. In totale: $N-1+13N+3N = 17N-1$ operazioni

Tempo di calcolo

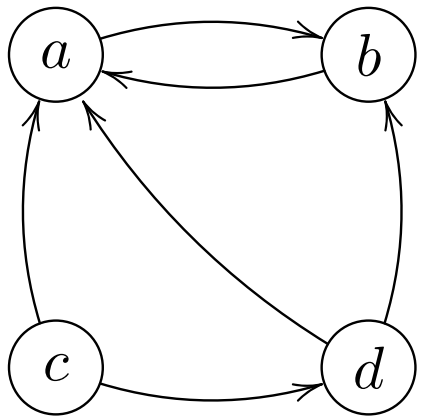
$$17N \;/\; (3.6 \times 10^{14}) \approx 0.0004 \text{ sec.}$$

$$50 \text{ link } \Rightarrow\; 103N \;/\; (3.6 \times 10^{14}) \approx 0.0024 \text{ sec.}$$

Cosa accade variando p

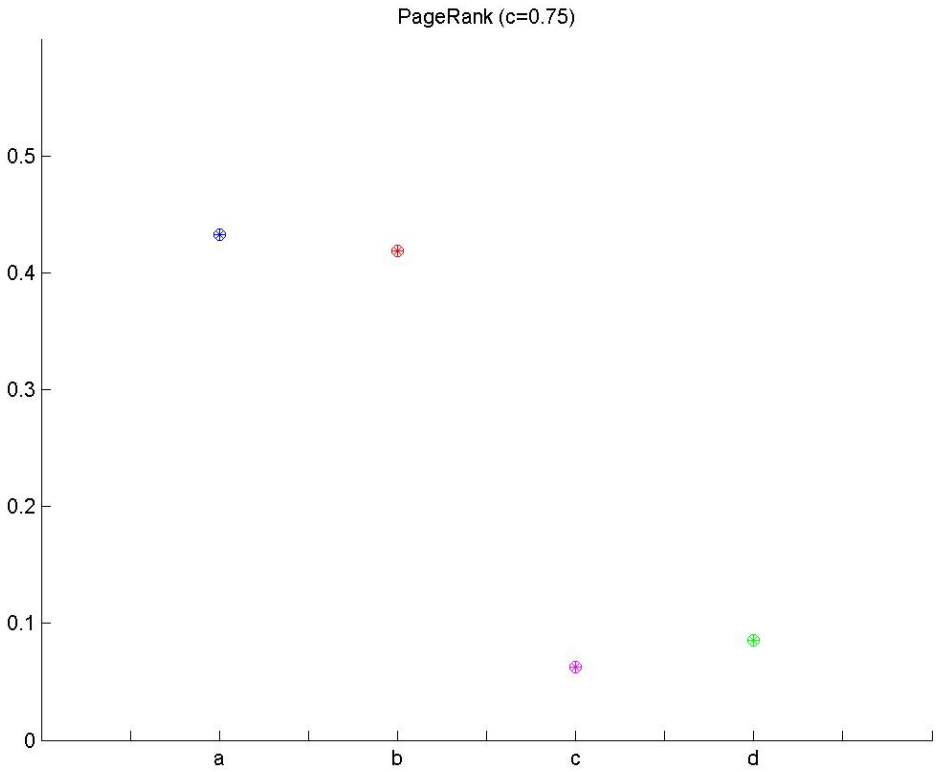
- Per p che tende a 1 il metodo iterativo diventa spaventosamente lento
- Il valore utilizzato nella pratica è $p = 0.85$.

Esempio: $\vec{v} = \frac{1}{N}[1, \ldots, 1]^T$



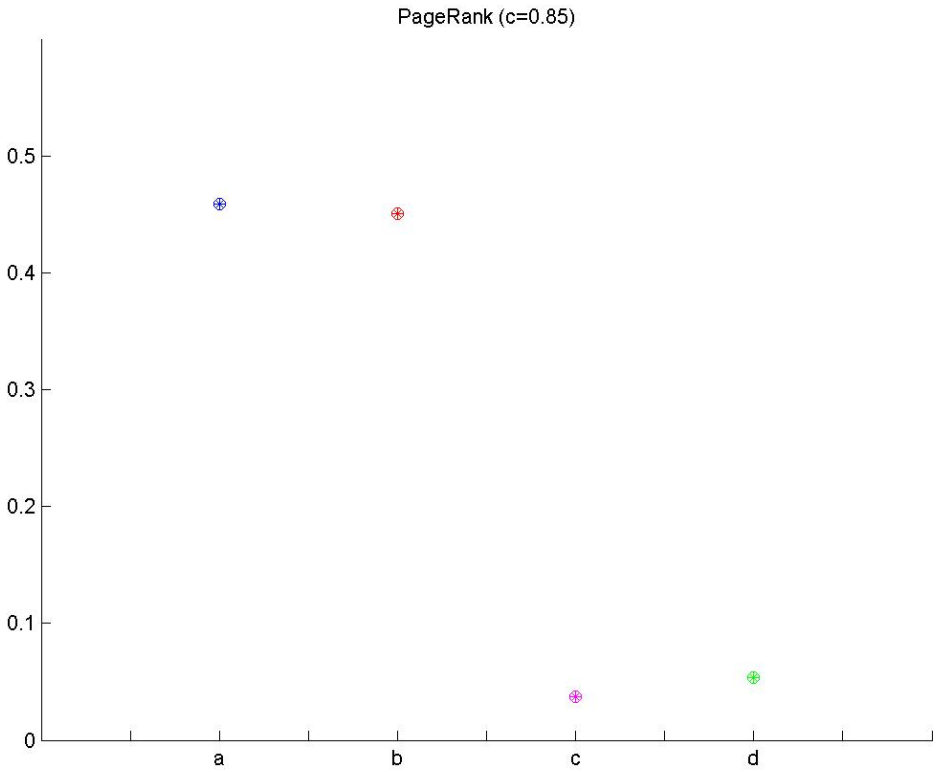
Esempio

p=0.75



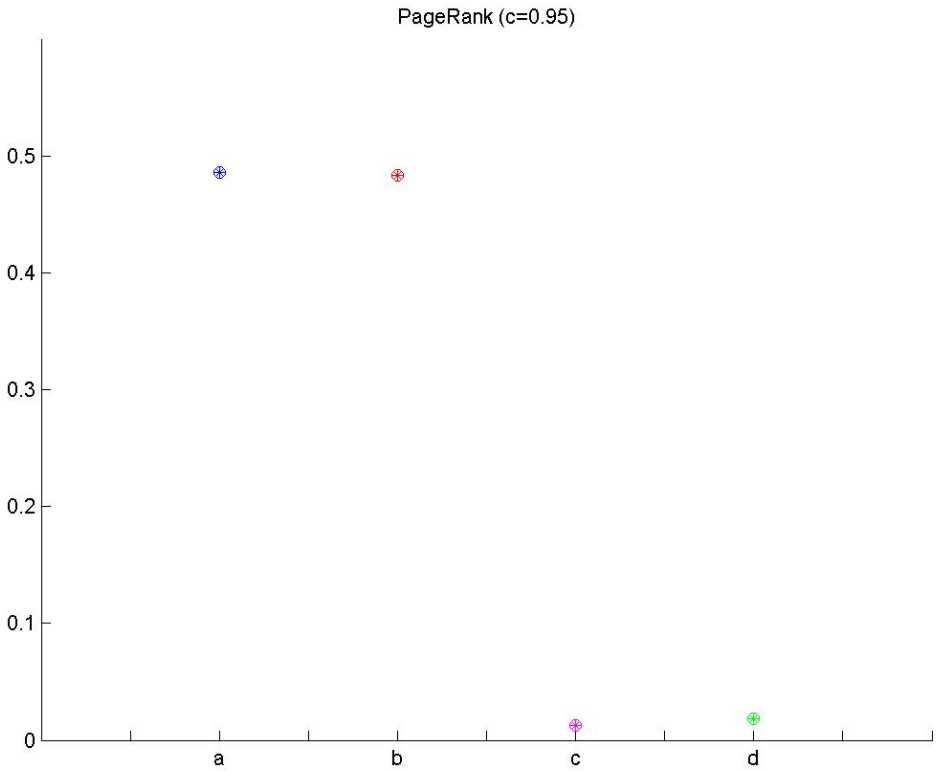
Esempio

p=0.85



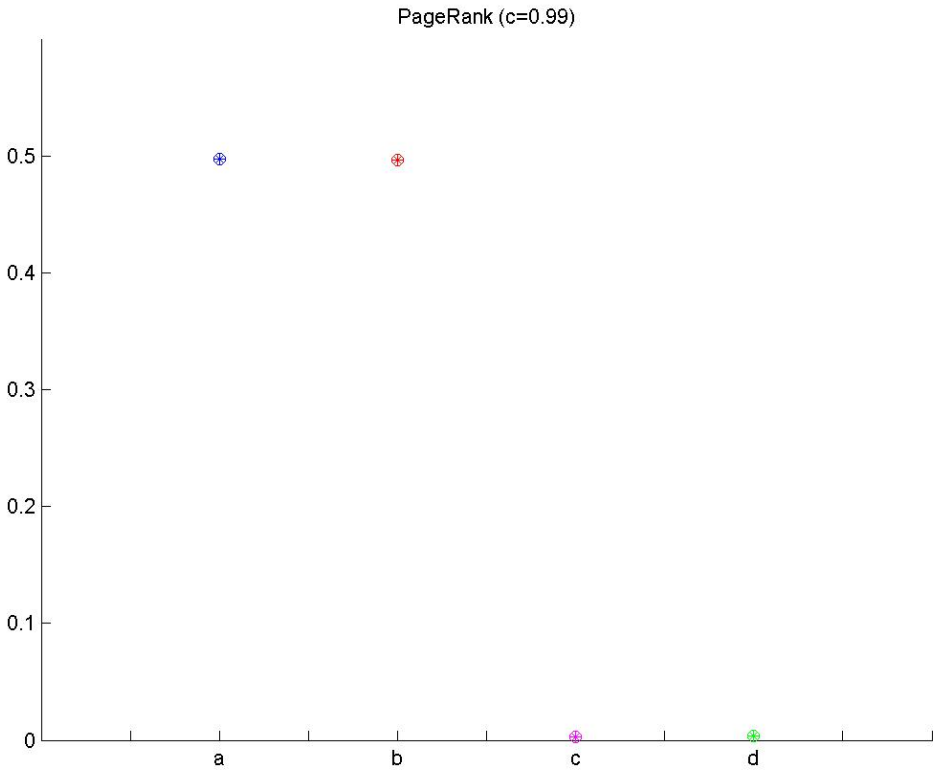
Esempio

p=0.95



Esempio

p=0.99



Esempio

Risultati per alcuni valori di p.

p	Num. iter.	$I(a)$	$I(b)$	$I(c)$	$I(d)$
0.75	68	0.4325	0.4191	0.0625	0.0859
0.85	119	0.4588	0.4502	0.0375	0.0534
0.95	377	0.4861	0.4830	0.0125	0.0184
0.99	1925	0.4972	0.4966	0.0025	0.0037

Possibili miglioramenti

- Metodi iterativi per approssimare $\vec{I}$ che convergono più velocemente
- Scelta di p.
- Studiare modifiche al modello per superare le patologie del caso ideale in cui $p \to 1$.

Riferimenti

- L. Page, S. Brin, R. Motwani, T. Winograd, *The pagerank citation ranking: Bringing order to the Web*, http://dbpubs.stanford.edu:8090/pub/1999-66.
- A. N. Langville, C. D. Meyer, *A survey of eigenvector methods for Web information retrieval*, SIAM Rev. 47 (2005), pp. 135-161.
- A. Arasu, J. Novak, A. Tomkins, J. Tomlin, *Page rank computation and the structure of the Web.* (2002) http://www2002.org/CDROM/poster/173.pdf
- R. Horn, S. Serra Capizzano, *A general setting for the parametric Google matrix*, Internet Math. 3-4 (2006), pp. 385-411.