

Reti complesse Ranking

Matteo Dell'Amico
dellamico@disi.unige.it



Applicazioni di rete 2
A.A. 2006-07

Outline

- 1 Motori di ricerca e ranking
 - Ricerca sul web
 - Ranking

- 2 PageRank
 - L'algoritmo
 - Quanto è grande il web? (reprise)

L'ago nel pagliaio

- Immaginiamo di avere una biblioteca con **25 miliardi di documenti** e **nessun bibliotecario**.
 - Chiunque può aggiungere un libro quando vuole, **senza dirlo a nessuno**.
 - Siamo convinti che l'informazione che ci serve ci sia, e vogliamo **trovarla**.
-
- Messo così, il problema sembra insolubile.
 - Eppure, i motori di ricerca fanno **proprio questo!**

Motori di ricerca

- Che lavoro fa un motore di ricerca?
 - 1 **Esplora** il web (crawling)
 - 2 **Memorizza** le informazioni trovate in un database
 - 3 Usa queste ultime per **rispondere** alle richieste degli utenti.
- Oggi ci concentreremo sull'**ultimo** punto.

Utenti e query

- Gli utenti forniscono poco ai motori di ricerca...
 - Query spesso **vaghe** (2001: 2.5 termini in media, 80% < 3 termini)
 - Nell'85% dei casi, si fermano ai primi **10** risultati
- Il motore di ricerca ha l'improbabile compito di scoprire **quale** pagina web, su 25 miliardi, risponda meglio alla richiesta "hotel sicilia".

Query

- Cosa significa affermare che una pagina web **soddisfa una query**?
- 1 Le parole della query sono **rilevanti** rispetto al contenuto della pagina.
 - 2 La pagina è **autorevole**.
- Concetto **quantitativo**: esiste un grado di **affinità** tra le query e le pagine web.
 - Restituiamo come **primi risultati** quelli **più affini** alla query.

Ranking: prima generazione

1995-1997: AltaVista, Lycos, Excite

- Tecniche tradizionali di Information Retrieval: indicizzazione di testi **piani**.
- Viene ignorata la **struttura ipertestuale** del web.
- Basata sulla tradizione dell'**Information Retrieval**.

Information Retrieval

- Disciplina con più di **40 anni** di storia.
- Strumenti per trovare articoli rilevanti in **collezioni scientifiche**.
- Assunzione implicita: **tutti i documenti sono autorevoli**.
- Questo **non è vero per il web!**

Rilevanza delle pagine

Idea

- Le parole più **rare** della query sono più rilevanti.
 - Le parole che appaiono più **di frequente** nel documento sono più rilevanti.
-
- Tecnica usata frequentemente: **TF*IDF**.
 - **TF** (*Term Frequency*): frequenza del termine all'interno del documento.
 - **IDF** (*Inverted Document Frequency*): (logaritmo dell') inverso della frequenza del termine in tutti i documenti.
 - Il punteggio per ogni documento si ottiene sommando il valore **TF*IDF per ogni elemento della query**.

Seconda generazione

- Possiamo sfruttare il fatto che il **web non è plain-text** per riconoscere le pagine web **autorevoli**?

- Analisi dei **link**
- Analisi dei **click** sui risultati della ricerca
- Analisi degli **anchor text**

```
<a href='http://www.unige.it/'>Università di Genova</a>
```

... a proposito degli anchor text

The screenshot shows a Google search interface. The search bar contains the text "miserable failure". Below the search bar, there are navigation links for "Advanced Search", "Preferences", "Language Tools", and "Search Tips". A "Google Search" button is visible. Below the search bar, there are radio buttons for "the web" (selected) and "pages from Canada". A navigation bar includes "Web", "Images", "Groups", "Directory", and "News". Below the navigation bar, a blue bar indicates the search query: "Searched the web for 'miserable failure'". The search results are listed below, starting with "Biography of President George W. Bush". Each result includes a title, a brief description, and a URL with additional information like "29k" or "36k" and "Cached" or "Similar pages".

Google™ [Advanced Search](#) [Preferences](#) [Language Tools](#) [Search Tips](#)
"miserable failure"
Search: the web pages from Canada
[Web](#) [Images](#) [Groups](#) [Directory](#) [News](#)
Searched the web for "miserable failure" Rest

[Biography of President George W. Bush](#)
Home > President > Biography President George W. Bush En Español.
George W. Bush is the 43rd President of the United States. He ...
Description: Biography of the president from the official White House web site.
Category: [Kids and Teens](#) > [School Time](#) > ... > [Bush, George Walker](#)
www.whitehouse.gov/president/gwbbio.html - 29k - [Cached](#) - [Similar pages](#)

[Biography of Jimmy Carter](#)
Home > History & Tours > Past Presidents > Jimmy Carter. Jimmy Carter.
Jimmy Carter aspired to make Government "competent and compassionate ...
Description: Short biography from the official White House site.
Category: [Society](#) > [History](#) > ... > [Presidents](#) > [Carter, James Earl](#)
www.whitehouse.gov/history/presidents/jc39.html - 36k - [Cached](#) - [Similar pages](#)

[Michael Moore.com](#)
February 11, 2004 (67th anniversary of the Great Flint Sit-Down Strike) An Open
Letter from Michael Moore to George "I'ma War President!" Bush. Dear Mr. Bush, ...
Description: Official site of the gadfly of corporations, creator of the film Roger and Me and the television show...

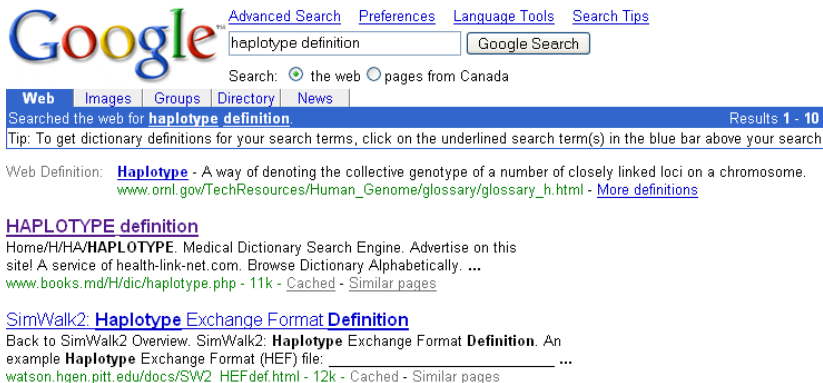
- Google Bombing

```
<a href='...'>miserable failure</a>
```

Terza generazione

- Ancora *in fieri*
- **Analisi semantica**: di che stiamo parlando?
- **Dipendenza dal contesto**: personalizzazione, contesto geografico...

Terza generazione: esempi (1)



The image shows a screenshot of a Google search interface. At the top left is the Google logo. To its right are links for 'Advanced Search', 'Preferences', 'Language Tools', and 'Search Tips'. Below these is a search input box containing the text 'haplotype definition' and a 'Google Search' button. Under the search box, there are radio buttons for 'the web' (which is selected) and 'pages from Canada'. Below the search area is a blue navigation bar with tabs for 'Web', 'Images', 'Groups', 'Directory', and 'News'. The 'Web' tab is active. Below the navigation bar, a blue bar displays the search results: 'Searched the web for **haplotype definition**. Results 1 - 10'. Below this is a tip: 'Tip: To get dictionary definitions for your search terms, click on the underlined search term(s) in the blue bar above your search'. The first search result is a 'Web Definition' for 'Haplotype', described as 'A way of denoting the collective genotype of a number of closely linked loci on a chromosome.' It includes a URL 'www.ornl.gov/TechResources/Human_Genome/glossary/glossary_h.html' and a link to 'More definitions'. Below this is a section titled 'HAPLOTYPE definition' with a link to 'Home/H/HA/HAPLOTYPE. Medical Dictionary Search Engine. Advertise on this site! A service of health-link-net.com. Browse Dictionary Alphabetically. ...' and another link 'www.books.md/H/dic/haplotype.php - 11k - Cached - Similar pages'. The next section is 'SimWalk2: Haplotype Exchange Format Definition' with a link to 'Back to SimWalk2 Overview. SimWalk2: Haplotype Exchange Format Definition. An example Haplotype Exchange Format (HEF) file: _____ ...' and another link 'watson.hgen.pitt.edu/docs/SW2_HEFdef.html - 12k - Cached - Similar pages'.

Google [Advanced Search](#) [Preferences](#) [Language Tools](#) [Search Tips](#)

haplotype definition

Search: the web pages from Canada

Web [Images](#) [Groups](#) [Directory](#) [News](#)

Searched the web for **haplotype definition**. Results 1 - 10

Tip: To get dictionary definitions for your search terms, click on the underlined search term(s) in the blue bar above your search

Web Definition: **Haplotype** - A way of denoting the collective genotype of a number of closely linked loci on a chromosome. www.ornl.gov/TechResources/Human_Genome/glossary/glossary_h.html - [More definitions](#)

HAPLOTYPE definition

[Home/H/HA/HAPLOTYPE. Medical Dictionary Search Engine. Advertise on this site! A service of health-link-net.com. Browse Dictionary Alphabetically. ...](#)

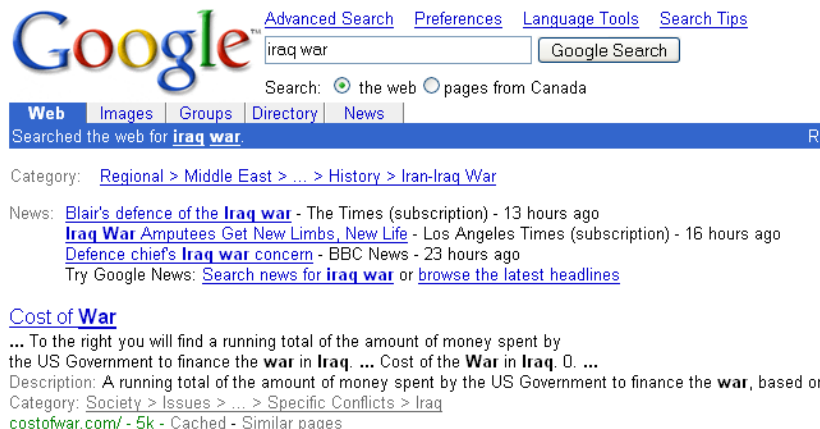
www.books.md/H/dic/haplotype.php - 11k - [Cached](#) - [Similar pages](#)

SimWalk2: Haplotype Exchange Format Definition

[Back to SimWalk2 Overview. SimWalk2: Haplotype Exchange Format Definition. An example Haplotype Exchange Format \(HEF\) file: _____ ...](#)

watson.hgen.pitt.edu/docs/SW2_HEFdef.html - 12k - [Cached](#) - [Similar pages](#)

Terza generazione: esempi (2)



The screenshot shows a Google search interface. At the top left is the Google logo. To its right are links for 'Advanced Search', 'Preferences', 'Language Tools', and 'Search Tips'. Below the logo is a search input field containing the text 'iraq war' and a 'Google Search' button. Underneath the input field, there are radio buttons for 'the web' (which is selected) and 'pages from Canada'. A navigation bar below the search area contains tabs for 'Web', 'Images', 'Groups', 'Directory', and 'News'. A blue banner below the navigation bar displays the search query 'Searched the web for **iraq war**.' with a 'R' icon on the right. Below the banner, the 'Category' is listed as 'Regional > Middle East > ... > History > Iran-Iraq War'. The 'News' section lists three results: 'Blair's defence of the **iraq war** - The Times (subscription) - 13 hours ago', '**Iraq War** Amputees Get New Limbs, New Life - Los Angeles Times (subscription) - 16 hours ago', and 'Defence chief's **iraq war** concern - BBC News - 23 hours ago'. Below the news items, there is a prompt to 'Try Google News: [Search news for iraq war](#) or [browse the latest headlines](#)'. The main result is titled 'Cost of War' and includes a snippet: '... To the right you will find a running total of the amount of money spent by the US Government to finance the **war** in **Iraq**. ... Cost of the **War** in **Iraq**. 0. ...'. Below the snippet is a description: 'Description: A running total of the amount of money spent by the US Government to finance the **war**, based or'. The category for this result is 'Society > Issues > ... > Specific Conflicts > Iraq'. At the bottom of the result, there is a link to 'costofwar.com/ - 5k - Cached - Similar pages'.

Google [Advanced Search](#) [Preferences](#) [Language Tools](#) [Search Tips](#)

iraq war

Search: the web pages from Canada

Web [Images](#) [Groups](#) [Directory](#) [News](#)

Searched the web for **iraq war**. R

Category: [Regional > Middle East > ... > History > Iran-Iraq War](#)

News: [Blair's defence of the **iraq war**](#) - The Times (subscription) - 13 hours ago
[Iraq War Amputees Get New Limbs, New Life](#) - Los Angeles Times (subscription) - 16 hours ago
[Defence chief's **iraq war** concern](#) - BBC News - 23 hours ago

Try Google News: [Search news for iraq war](#) or [browse the latest headlines](#)

Cost of War

... To the right you will find a running total of the amount of money spent by the US Government to finance the **war** in **Iraq**. ... Cost of the **War** in **Iraq**. 0. ...

Description: A running total of the amount of money spent by the US Government to finance the **war**, based or

Category: [Society > Issues > ... > Specific Conflicts > Iraq](#)

[costofwar.com/ - 5k - Cached - Similar pages](#)

Terza generazione: esempi (3)



Google [Advanced Search](#) [Preferences](#) [Language Tools](#) [Search Tips](#)

the answer to life the universe and e

Search: the web pages from Canada

The following words are very common and were not included in your search.
The "AND" operator is unnecessary -- we include all search terms by default.

Web | [Images](#) | [Groups](#) | [Directory](#) | [News](#)

Searched the web for **[the answer to life the universe and everything](#)**.



the answer to life the universe and everything = 42

[More about calculator.](#)

[The Answer to Life, the Universe, and Everything - Wikipedia](#)

... Google has recently added a calculator function to its search engine, which contains a formula for the question **answer to life the universe and everything**. ...

en2.wikipedia.org/wiki/The_Answer_to_Life,_the_Universe,_and_Everything - 17k - [Cached](#) - [Similar pages](#)

Autorevolezza e link

- **Link Analysis:** si può valutare l'autorevolezza di una pagina web studiando la **struttura dei link** del WWW.
- I link possono essere visti come **raccomandazioni** alla lettura: una pagina **vota** usando i suoi link.
- Le pagine che ricevono più link sono **più importanti**.

Autorevolezza e link (2)

Domanda

- È ragionevole **contare i link entranti** per valutare l'autorevolezza di una pagina?

Risposta

- Non del tutto. I link da pagine diverse potrebbero valere:
 - molto (per esempio, da www.slashdot.org)
 - poco (da www.disi.unige.it/person/DellamicoM)
 - nulla (da un sito "spam")
- Se un link proviene da una pagina importante, è **importante esso stesso**.

PageRank

PageRank

- L'algoritmo che Google usa per **valutare l'autorevolezza** delle pagine.
- Definisce l'importanza di ogni pagina in funzione dell'**importanza delle pagine che la linkano**.
- Se ogni pagina P_j ha l_j link, e viene linkata dall'insieme di pagine B_i , definiamo l'**importanza di P_i** come

$$I(P_i) = \sum_{P_j \in B_i} \frac{I(P_j)}{l_j}$$

PageRank (2)

- La definizione che vista ricorda l'uovo e la gallina. Ha senso definire l'importanza dei nodi in funzione dell'importanza dei nodi?
- Dobbiamo assicurarci che esista un'**unica** attribuzione di valori $I(P_i)$ che soddisfi la "definizione".
- Vogliamo avere un **algoritmo** per trovarla.

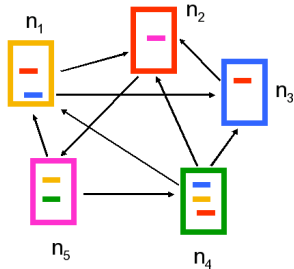
Random Walk

- Pensiamo ad un navigatore “senza meta” che gira sul WWW seguendo link **a caso**.
- Possiamo rappresentare il suo cammino come una **catena di Markov**.
 - **Stato**: pagina visualizzata dal navigatore.
 - **Matrice di transizione**: la probabilità che durante la passeggiata si segua un dato arco.
 - Dalla stessa pagina, assegnamo la **stessa probabilità** ad ogni arco uscente.
- Se la catena così definita possiede una **distribuzione stazionaria**, allora questa **verifica la definizione** data in precedenza.

Matrice di transizione

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

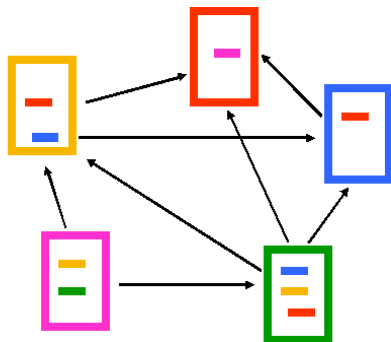
$$P = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$



- Matrice di transizione: nient'altro che la **matrice di adiacenza** resa stocastica (somma 1 sulle righe).

Vicoli ciechi (1)

$$P = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$

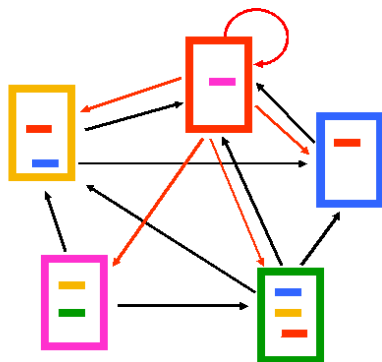


Problema

- Che fare per i “vicoli ciechi”?

Vicoli ciechi (2)

$$P' = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$



Soluzione

- Decidiamo che si ritorna ad **una pagina a caso**.

Esiste una distribuzione stazionaria?

- Abbiamo visto che la probabilità limite converge alla distribuzione stazionaria se gli stati sono
 - **ricorrenti-positivi**: la matrice sottostante è **connessa**.
 - **aperiodici**: la matrice sottostante non è **n-partita**.

Adattamento

- Con una certa probabilità α , effettuiamo un salto verso un nodo **a caso**.
- La matrice sottostante diventa **completamente connessa**, quindi verifica le proprietà richieste.

La matrice di Google

$$P'' = \alpha \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \end{bmatrix} + (1-\alpha) \begin{bmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{bmatrix}$$

- La matrice P'' ci **garantisce la convergenza** alla distribuzione stazionaria.
- Il valore α , nella pratica, assume valori vicini a 0.85.

Metodo delle potenze

- Adottiamo il **metodo delle potenze** per trovare la soluzione a $\pi = \pi P''$:
 - 1 Partiamo con un vettore l^0
 - 2 Impostiamo $i := 0$
 - 3 Finché il calcolo non converge:
 - 1 Calcoliamo $l^{i+1} := l^i P''$
 - 2 Incrementiamo $i := i + 1$
- D'ora in poi assumeremo che la **somma delle componenti** di l^0 sia 1.
- Questo ci garantisce che la somma valga **1 per qualsiasi l^k** (a meno di errori di approssimazione).

Convergenza (1)

- Richiamo dalla lezione sull'algebra lineare: abbiamo un vettore iniziale I^0 ed una matrice P'' .
- Chiamiamo $\lambda_1 \dots \lambda_n$ gli **autovalori** di P'' e $v_1 \dots v_n$ i rispettivi **autovettori**.
- Grazie alle proprietà delle matrici stocastiche ergodiche, sappiamo che $\mathbf{1} = \lambda_1 > |\lambda_2| \geq \dots \geq |\lambda_n|$.

Convergenza (2)

- Supponiamo per semplicità che I^0 possa essere espresso come **combinazione lineare** degli autovettori:

$$I^0 = c_1 v_1 + \dots + c_n v_n$$

- Per definizione di autovalori ed autovettori,

$$I^1 = I^0 P'' = c_1 v_1 + c_2 \lambda_2 v_2 + \dots + c_n \lambda_n v_n$$

$$I^k = I^{k-1} P'' = c_1 v_1 + c_2 \lambda_2^k v_2 + \dots + c_n \lambda_n^k v_n$$

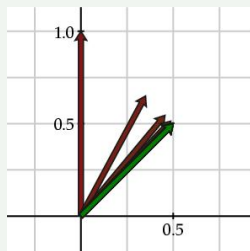
- Dato che i λ_i^k tendono a 0 se $i \geq 2$, I_k tende a $I = c_1 v_1$, cioè uno scalare moltiplicato per v_1 .
- La velocità di convergenza dipende da $\lambda_2/\lambda_1 = \lambda_2$.

Convergenza (3)

- Applichiamo il metodo delle potenze alla matrice

$$\begin{bmatrix} 0.65 & 0.35 \\ 0.35 & 0.65 \end{bmatrix}$$

- Gli autovalori sono $\lambda_1 = 1$ e $\lambda_2 = 0.3$. I vettori I^k convergono piuttosto **rapidamente** al vettore stazionario in verde.

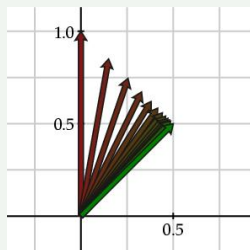


Convergenza (4)

- Prendiamo una seconda matrice:

$$\begin{bmatrix} 0.85 & 0.15 \\ 0.15 & 0.85 \end{bmatrix}$$

- Gli autovalori sono $\lambda_1 = 1$ e $\lambda_2 = 0.7$. Dato che λ_2 è più grande, i vettori I^k convergono **più lentamente** al vettore stazionario.



Convergenza (5)

- È possibile dimostrare che $\lambda_2 \leq \alpha$.
 - Asintoticamente, la precisione dopo k iterazioni è λ_2^k .
 - Sappiamo quindi che la velocità di convergenza dell'algoritmo è **garantita** indipendentemente dalla topologia del grafo.
-
- Google riesce ad ottenere valori accurati dopo **50-100** iterazioni dell'algoritmo.
 - Tempo di calcolo: circa **un giorno**.
 - Precisione garantita: $\alpha^{50} \simeq 0.85^{50} \simeq \mathbf{0.0003}$.

Implementazione (1)

- La matrice P'' è **enorme**: nessun computer riuscirebbe ad immagazzinarla.
 - Fortunatamente, possiamo **semplificare** il calcolo.
-
- Ci basta propagare PageRank “solo” sui link.
 - Teniamo traccia della **probabilità di effettuare un salto casuale** ad ogni passo, aumentandola ogni volta che incontriamo vicoli ciechi.

Implementazione (2)

Passo iterativo

- Inizializziamo il nuovo vettore: $l^{i+1} := \vec{0}$
- Impostiamo la probabilità p_s di salto casuale: $p_s := 1 - \alpha$
- Per ogni nodo j :
 - Se j è un vicolo cieco, $p_s := p_s + \alpha l^i[j]$
 - Altrimenti, per ognuno degli m link uscenti da j :
 - Sia k il nodo che riceve il link.

$$l^{i+1}[k] := l^{i+1}[k] + \alpha \frac{l^i[j]}{m}$$

- Effettuiamo i salti. Per ogni nodo j :

$$l^{i+1}[j] := l^{i+1}[j] + \frac{p_s}{n}$$

Campionamento quasi uniforme (1)

Approccio

- Effettuare un **cammino casuale** sul WWW, seguendo link a caso.

Problema

- Tutte le pagine hanno la **stessa probabilità** di essere raggiunte?

Campionamento quasi uniforme

Problema

- Tutte le pagine hanno la **stessa probabilità** di essere raggiunte?

Risposta

- **Ora lo sappiamo!** PageRank calcola proprio la probabilità che un cammino raggiunga la pagina data.
- Con la parte di Web che abbiamo esplorato, stimiamo PageRank.
- Prendiamo un nodo con una probabilità **inversamente proporzionale** al suo PageRank stimato.

Riferimenti



David Austin.

How Google Finds Your Needle in the Web's Haystack.

<http://www.ams.org/featurecolumn/archive/pagerank.html>



S. Brin e L. Page.

The Anatomy of a Large Scale Search Engine.

Proc. 7th International WWW Conference, 1998.



Taher Haveliwala, Sepandar Kamvar.

The Second Eigenvalue of the Google Matrix.

Stanford University Tech. Rep., 2003.



M. Henzinger, A. Heydon, M. Mitzenmacher e M. Najork.

On Near-Uniform URL Sampling.

Proc. 9th International WWW Conference, 2000.