UNIVERSITÀ DI PADOVA
Dipartimento di Ingegneria dell'Informazione

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

Algoritmi di analisi dei link per motori di ricerca

Luca Pretto
luca.pretto@unipd.it

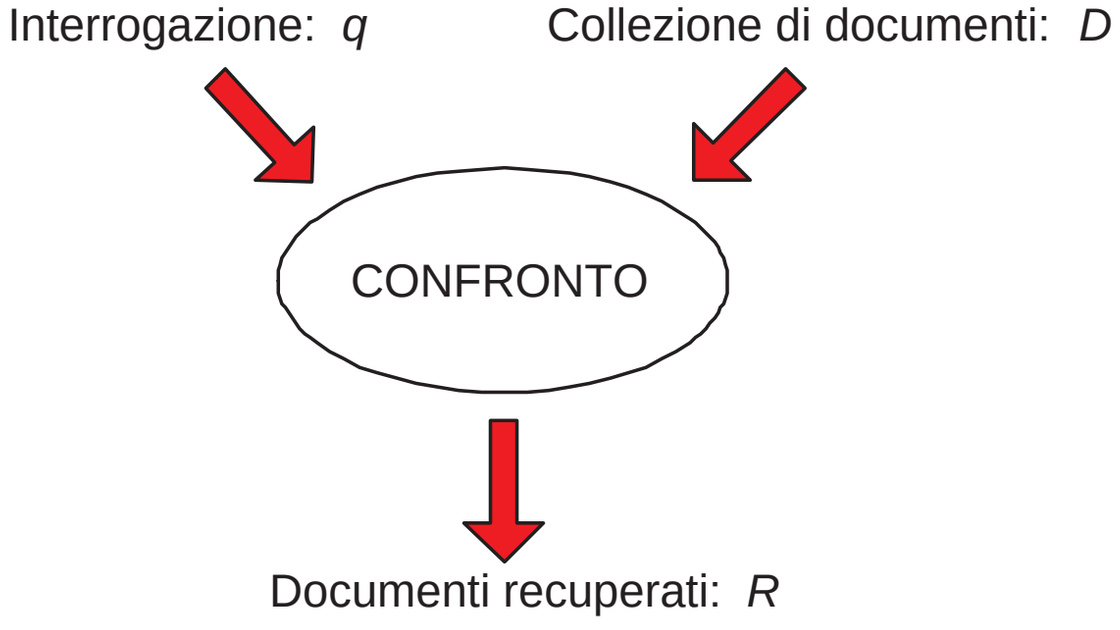
JUG Padova Meeting #25, Padova, 18 febbraio 2006

Reperimento dell'informazione (Information Retrieval)

Parte della *computer science*: studia il recupero dell'informazione (non dei dati) da una collezione di documenti scritti. I documenti recuperati devono soddisfare l'*esigenza informativa dell'utente*, solitamente espressa in linguaggio naturale.
[R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, 1999.]

Information Retrieval System (IRS): sistema software utilizzato per il recupero dell'informazione.

Il processo di recupero dei documenti



È richiesto di recuperare l'insieme dei documenti rilevanti rispetto all'interrogazione formulata dall'utente.

Esempio: recupero con il modello vettoriale

$D = \{d_1, d_2, \ldots, d_J\}$. $w_{i,j}$, $i = 1, 2, \ldots, k$: peso del termine t_i nel documento d_j. $w_{i,q}$, $i = 1, 2, \ldots, k$: peso del termine t_i nell'interrogazione q. k: numero totale di termini nella collezione.

Rappresentazioni vettoriali:

$$\boldsymbol{d_j} = \left[\begin{array}{cccc} w_{1,j} & w_{2,j} & \ldots & w_{k,j} \end{array} \right]^T$$

$$\boldsymbol{q} = \left[\begin{array}{cccc} w_{1,q} & w_{2,q} & \ldots & w_{k,q} \end{array} \right]^T$$

Documenti restituiti in ordine di similarità a q:

$$\mathsf{sim}(d_j, q) = \frac{\langle \boldsymbol{d_j}, \boldsymbol{q} \rangle}{\|\boldsymbol{d_j}\| \cdot \|\boldsymbol{q}\|} = \frac{\sum_{i=1}^{k} w_{i,j} \cdot w_{i,q}}{\sqrt{\sum_{i=1}^{k} w_{i,j}^2} \cdot \sqrt{\sum_{i=1}^{k} w_{i,q}^2}}$$

Caratteristiche del World Wide Web

Collezione di documenti eterogenei, con caratteristiche di importanza, autorevolezza, qualità molto diverse.

Non è sufficiente considerare la rilevanza.

Soluzione: uso dei link [Marchiori, 1997]

Algoritmi di analisi dei link

Web come grafo: deduzione della qualità di una pagina Web dalla struttura topologica di un grafo orientato che rappresenta tutto il Web o un suo sottoinsieme.

PageRank $\Longrightarrow$ Google

HITS $\Longrightarrow$ Teoma

Altri algoritmi: SALSA, pHITS, Bayesiano, . . .

L'algoritmo PageRank

Associa a *ogni* pagina Web un numero reale positivo, detto esso stesso PageRank, che dovrebbe fornire l'importanza della pagina nell'*intero* Web.

Ordinamento query-independent e off-line.

Non basta contare il numero di link in ingresso: problema del Web spamming.

Google combina i valori di PageRank con quelli di rilevanza, per ordinare le pagine restituite in risposta a un'interrogazione.

Definizione formale di PageRank

Il PageRank di una generica pagina k, $P_r(k)$, è dato dalla formula:

$$P_r(k) = \frac{d}{N} + (1-d) \sum_{h \to k} \frac{P_r(h)}{o(h)} \qquad k \in S$$

S: insieme delle pagine Web considerate; $N = |S|$.

$o(h)$: numero di link in uscita dalla pagina h.

$0 < d < 1$ è detto *damping factor*.

$h \to k \Leftrightarrow h$ punta a k.

Giustificazione matematica della formula proposta

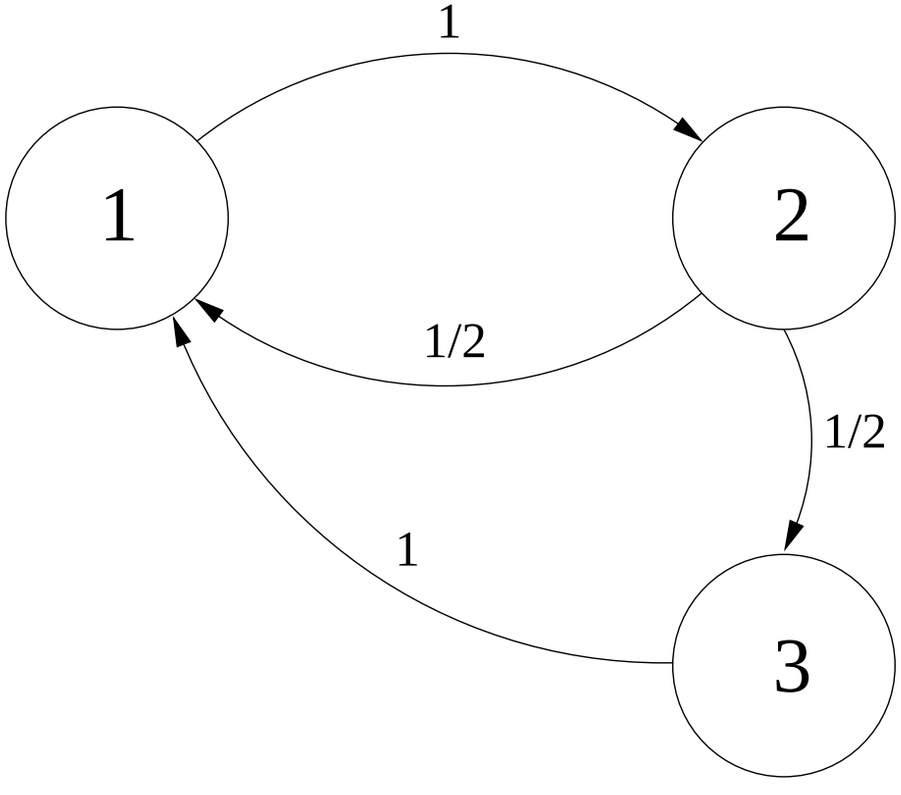
Teoria delle catene di Markov a tempo discreto.

$P_r(k)$: probabilità limite e stazionaria di un'opportuna catena di Markov a tempo discreto.

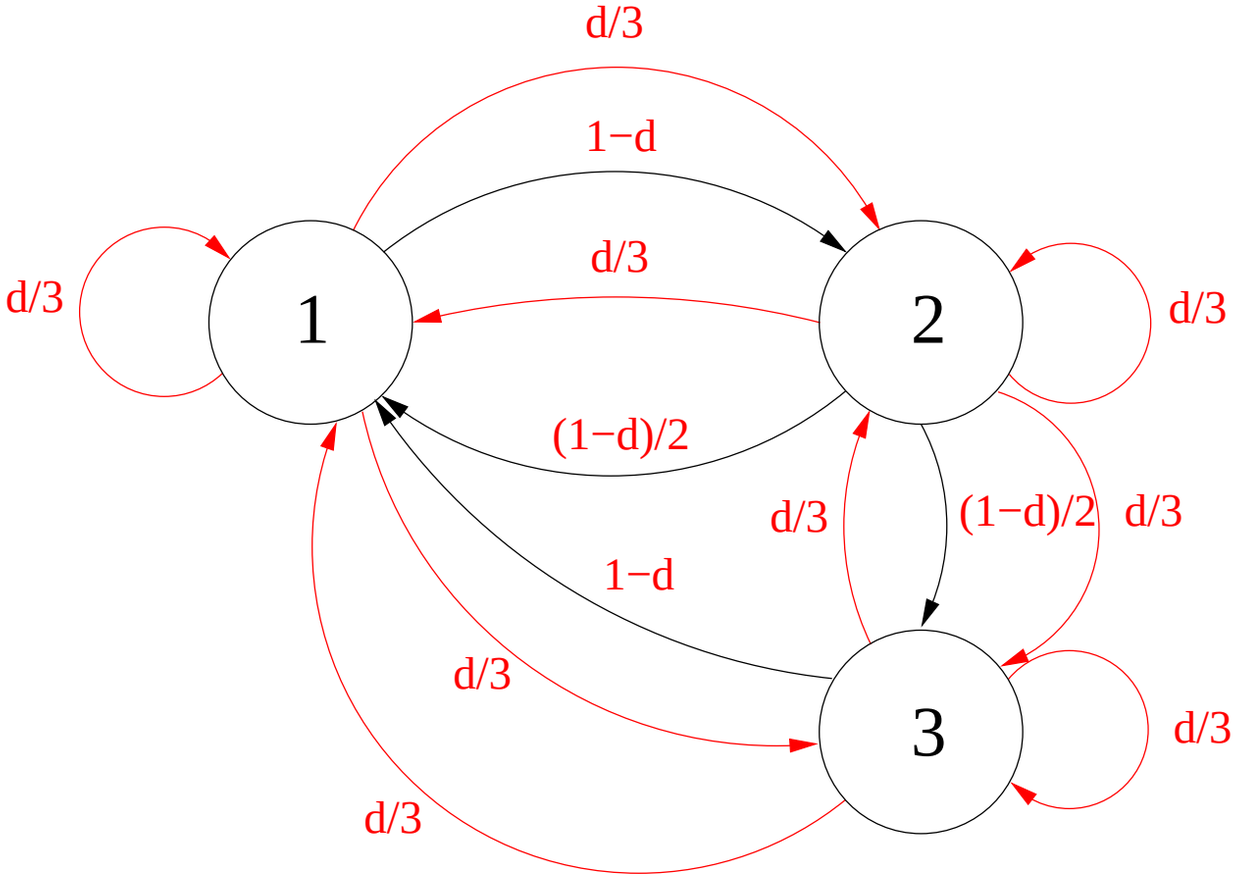
Interpretazione del "Web surfer".

Formula generale, con possibilità di personalizzazione.

Interpretazione del "Web surfer"



Interpretazione del "Web surfer"



L'algoritmo HITS

HITS (Hyperlink-Induced Topic Search); J. Kleinberg (1998).

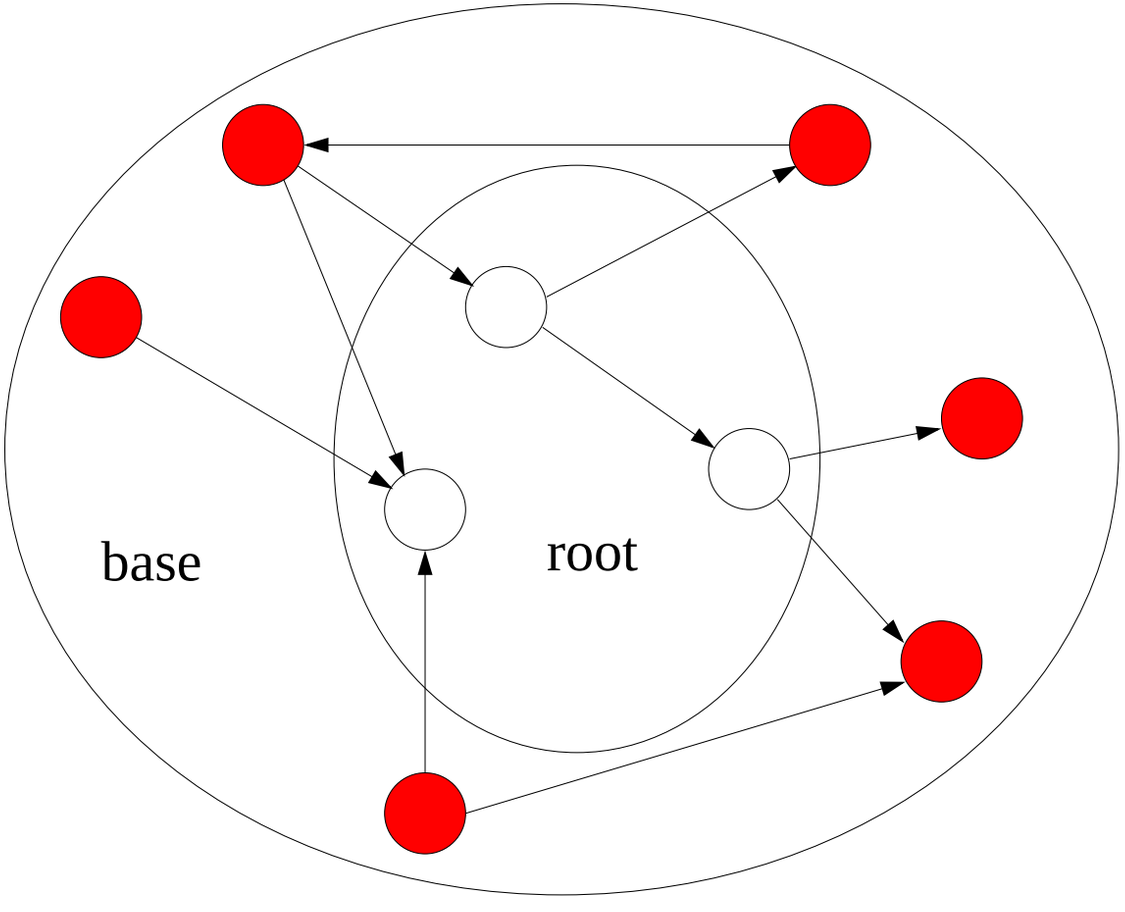
Ranking dipendente dalla query.

Di interesse in altri campi: stemming, latent semantic indexing.

Due parti:

1. costruzione del digrafo;
2. individuazione dei nodi authority e hub nel digrafo.

Costruzione del digrafo



Ranking dei nodi

Nodo i: peso di authority a_i e peso di hub h_i. Valore iniziale 1. Formule iterative di aggiornamento:

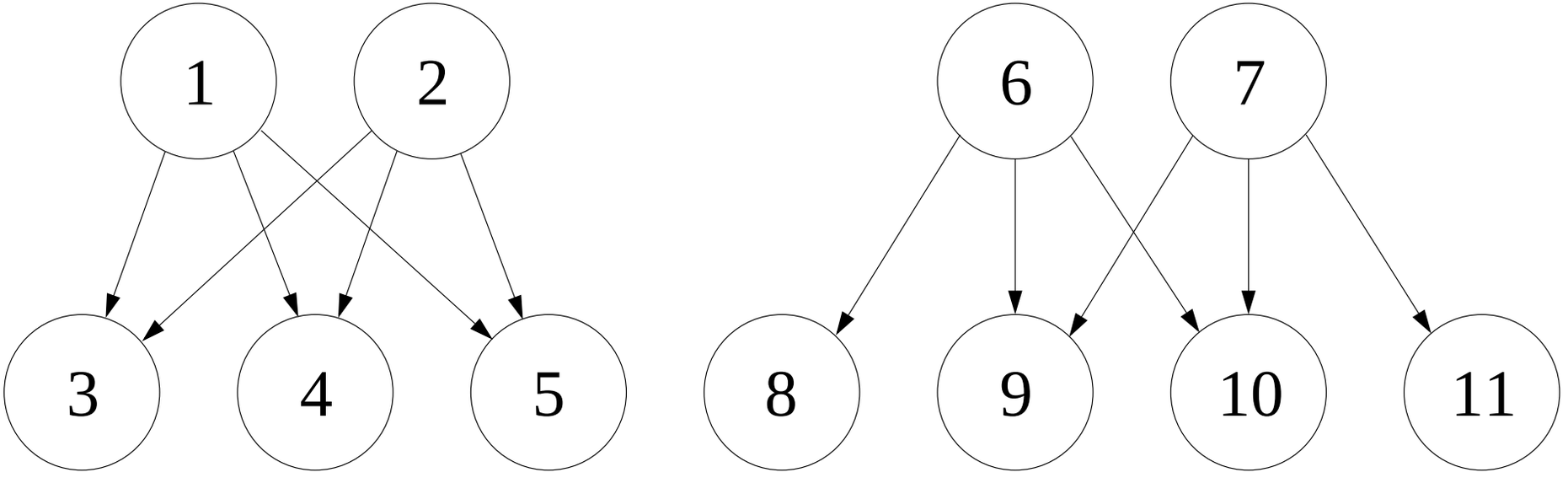
$$a_i^{(k)} = \sum_{j:j\to i} h_j^{(k-1)} \qquad h_i^{(k)} = \sum_{j:i\to j} a_j^{(k)}.$$

$\mathbf{A}$: matrice di adiacenza del digrafo.

$$\mathbf{a}^{(k)} = \mathbf{A}^T\mathbf{h}^{(k-1)} \qquad \mathbf{h}^{(k)} = \mathbf{A}\mathbf{a}^{(k)}.$$

Inoltre: normalizzazione.

Lotta tra comunità



La comunità di sinistra vince e si prende tutto!

Ordinamento delle pagine

Dopo $K > 0$ passi, trascurando la normalizzazione:

$$\mathbf{a}^{(K)} = (\mathbf{A}^T\mathbf{A})^{K-1}\mathbf{A}^T\mathbf{u} \qquad \mathbf{h}^{(K)} = (\mathbf{A}\mathbf{A}^T)^K\mathbf{u}$$

dove $\mathbf{u} = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^T$.

$\Rightarrow$ Ordinamento per autorevolezza $\neq$ rilevanza $+$ importanza.

Un esempio: confronto tra Google e Teoma con interrogazione "newspapers".