UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI INGEGNERIA

CORSO DI LAUREA IN INGEGNERIA INFORMATICA

ATTIVITÀ FORMATIVA: ELABORATO

ANALISI DELL'ALGORITMO DI RICERCA WEB PAGERANK

MATTEO RUFFIN

Relatore: Prof. Antonio Ponno

Anno Accademico 2011 - 2012

A tutte le persone che con la loro semplicità, vicinanza, estrema bontà e pazienza mi hanno permesso, giorno dopo giorno, di maturare e diventare ciò che oggi sono; sia nel bene che nel male.

M. R.

Sommario

La necessità di reperire in modo rapido dati ed informazioni aggiornate in risposta ad interrogazioni fornite dall'utenza, comporta un crescente e massiccio uso dei processi di recupero di informazioni.
Questi ultimi, nel corso degli anni sono stati sempre più affinati e migliorati in funzione sia dei cambiamenti subiti dalla rete web che al diversificarsi delle esigenze dell'utenza stessa.
In questo ambito, gode di considerevole rilievo nonché interesse applicativo l'algoritmo di Link Analysis PageRank, la cui trattazione viene riportata nel seguente elaborato strutturato come segue.
Nel primo capitolo viene introdotto il processo di Information Retrieval studiando anche l'utilità derivante dall'impiego di risorse avanzate nell'ambito della ricerca.
Il secondo capitolo è interamente focalizzato alla descrizione degli algoritmi di Link Analysis, quali Indegree, HITS, SALSA, PageRank.
Il terzo capitolo, di carattere matematico, tratta le catene di Markov, che costituiscono la base matematica dell'algoritmo PageRank, studiato approfonditamente nel corso del capitolo successivo.
Nel quarto capitolo, dedicato al PageRank, dopo una breve introduzione, viene affrontato il processo di determinazione della matrice delle probabilità di transizione ed il conseguente calcolo del vettore di PageRank effettuato adottando il metodo delle potenze (*Power Method*). Il capitolo, si conclude con l'esame di un esempio concreto di determinazione del PageRank eseguito su di un grafo individuato dalla struttura topologica di un mini web.

Indice

Capitolo 1

Information Retrieval

L'Information Retrieval IR[1] contraddistingue, data la propria importanza e dinamicità, uno dei processi più esaminati e studiati dai ricercatori la cui finalità volge alla semplificazione dell'individuazione e al conseguente accesso delle informazioni desiderate, reperite tra le enormi moli di dati generalmente disponibili in una generica sessione di ricerca.
In questo contesto, nel processo di determinazione di documenti elettronici[2] il più attinenti possibile alla specifica richiesta[3] elaborata dall'utenza, parecchia attenzione nonchè accuratezza viene affidata all'interpretazione della chiave di ricerca che generalmente, per garantire buoni risultati, deve essere finemente formulata (ad esempio sfruttando le potenzialità fornite dall'impiego di risorse avanzate) e strutturata sintetizzando il soggetto di ricerca attraverso la stesura di uno o più termini la cui conseguente fruizione permetta il reperimento di risultati quanto più rilevanti.

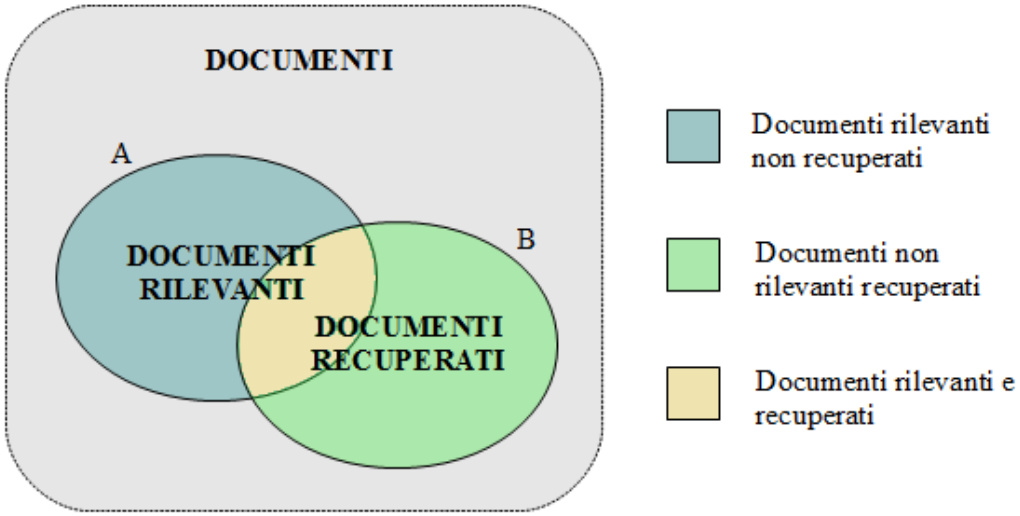
[1] Termine coniato da Calvin Mooers agli inizi degli anni '50 del Novecento.

[2] Nel prosieguo del testo, i termini documenti elettronici e pagine web verranno considerati sinonimi.

[3] Anche in questo caso, nel prosieguo del testo, tale termine verrà sostituito dai suoi più specifici sinonimi keyword, chiave di ricerca o query.

1.1 Il processo di Information Retrieval

Il reperimento di informazioni si concretizza attraverso la determinazione ed organizzazione della maggiore mole possibile di documentazioni pertinenti alla richiesta preventivamente inoltrata dall'utenza, senza incorrere al recupero di un'ingestibile quantità di dati di cui solo un'esigua parte risulta essere di utilità finale.



Fondamentale in questo senso è l'adozione di un'appropriata query compatibile al soggetto della ricerca, la cui particolare formulazione rende possibile all'utenza la determinazione e la conseguente restituzione e stesura di una classifica in cui le documentazioni trovate vengono ordinate in base alla rilevanza, premiando quelle di rilievo (inserendole in cima alla graduatoria generata) e penalizzando quelle di minor pertinenza (inserite in posizioni secondarie della lista).
Di notevole importanza in questo ambito, come esposto in [14], riveste l'introduzione di parametri utilizzati nella valutazione dei diversi processi di

IR volti ad una caratterizzazione di valenza generale:

- ***Richiamo:*** parametro espresso dal rapporto tra la quantità di documentazione rilevante recuperata ed il numero di documenti rilevanti;
- ***Precisione:*** parametro identificato dal rapporto tra la mole di documentazione rilevante recuperata e la quantità di documentazione recuperata;
- ***Rumore:*** parametro rappresentante il recupero e restituzione, in seguito ad un processo di ricerca, di documentazione elettronica considerata non rilevante;
- ***Silenzio:*** parametro rappresentante la mole di documenti rilevanti non recuperati.

Dal loro studio si percepisce quanto delicato ed al contempo accurato debba essere il processo di individuazione di un giusto compromesso tra i parametri **richiamo** e **precisione**, noto che all'aumentare del richiamo la precisione diminuisce mentre al diminuire del richiamo la precisione tende ad aumentare [19].
Inoltre, per rendere il più costruttivo possibile un processo di ricerca, può essere di utilità l'uso di **risorse avanzate**.

1.2 Risorse avanzate utilizzate nell'ambito della ricerca

L'impiego di risorse avanzate eseguito in fase di stesura della richiesta formulata dall'utenza che si avvale di processi di riperimento di informazioni per affrontare determinate sessioni di ricerca, conferisce l'elevazione della qualità intrinseca dell'interrogazione permettendo l'esclusione della documentazione contraddistinta da un'esigua se non nulla rilevanza nei confronti della particolare keyword inoltrata. Come descritto in [3], risulta necessario quindi, definire a priori tutte le caratteristiche che dovranno essere presenti nelle informazioni desiderate risultato della ricerca; in questo modo più specifica e precisa risulterà essere la chiave di ricerca, maggiori saranno le documentazioni pertinenti determinate attraverso essa.

Nella stesura di queste keyword, notevole importanza rivestono risorse avanzate, quali:

- **Operatori booleani**;
- **Operatori di prossimità**;
- **Caratteri Jolly**.

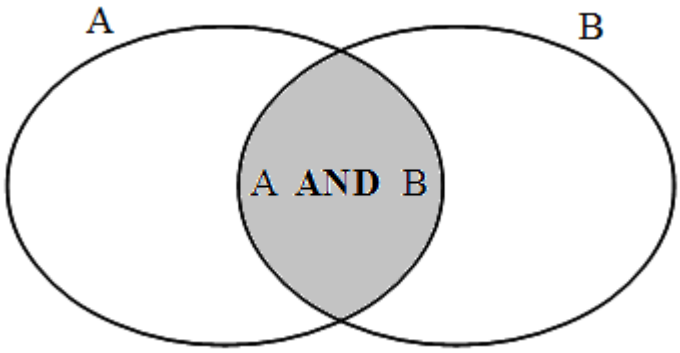
In passato, l'uso e l'implementazione di tali operatori è stato di considerevole utilità soprattutto nel processo di reperimento di informazioni in ambito bibliotecario o più genericamente in archivi elettronici. Con l'avvento delle moderne tecnologie di ricerca, alcuni tra essi sono stati sempre meno utilizzati diventando con il passare del tempo obsoleti, mentre altri, al contrario, sempre più involontariamente usati[4] o direttamente implementati dall'utenza con lo scopo di affinare le interrogazioni formulate in una sessione di ricerca.

1.2.1 Operatori booleani

Gli operatori booleani prendono il loro nome dal matematico inglese George Boole (1815 - 1864). Essi furono introdotti per operare su problemi di carattere logico, ma, nonostante ciò, inizialmente vennero poco utilizzati e sfruttati.
Solo in seguito, con l'avvento dei primi calcolatori e dell'era dell'elettronica digitale, si intuì l'estrema importanza nonché utilità che li contraddistingueva nell'analisi delle proposizioni logiche e, di conseguenza, si cominciò ad usufruirne e sfruttarne la potenza in maniera sempre più pesante.
I principali e più diffusi operatori booleani in ambiente informatico sono **AND**, **OR** e **NOT**.

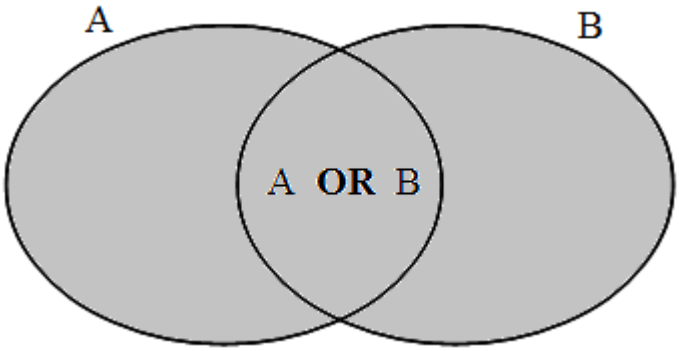
[4] Ci si riferisce all'uso automatico compiuto da alcuni motori di ricerca.

L'operatore AND



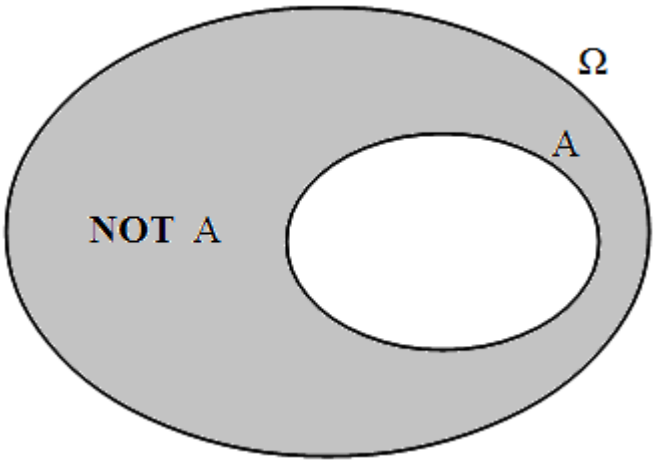
In una sessione di ricerca, l'uso dell'operatore booleano **AND** compiuto interponendolo tra due o più parole chiave, eventualmente poste in ordine di importanza, restituirà come risultato finale l'intera documentazione elettronica contenente tutte le parole indicate nella keyword.
Tale operatore corrisponde all'operazione insiemistica di *intersezione*.

L'operatore OR



In una sessione di ricerca, l'uso dell'operatore booleano **OR** compiuto interponendolo tra due o più parole chiave, restituirà come risultato finale l'intera documentazione elettronica contenente tutti i documenti in cui sono riportate una o più parole indicate nella keyword.
Tale operatore corrisponde all'operazione insiemistica di *unione*.

L'operatore NOT



In una sessione di ricerca, attraverso l'uso dell'operatore booleano **NOT** posto precedendo una parola chiave, viene indicato al motore di ricerca che non sono di interesse, e quindi non vengano restituiti, documenti elettronici contenenti quello specifico termine.
Come si può notare dall'immagine, tale operatore corrisponde all'operazione insiemistica di *complementazione*.

1.2.2 Operatori di Prossimità

Anche gli operatori di prossimità, come gli operatori booleani, rappresentano risorse avanzate sovente utilizzate in fase di reperimento di documentazione elettronica nell'ambito della ricerca nel web.
I più conoscuti ed utilizzati in ambiente informatico sono **ADJ**(n), **NEAR**(n) e **SAME**.

L'operatore ADJ(n)

L'operatore di prossimità **ADJ**(n) permette l'individuazione e conseguente reperimento di tutti i documenti affetti dalla presenza delle parole chiave ricercate **nell'ordine cui esse sono state specificate** al momento della formulazione dell'interrogazione, sia nel caso esse appaiano l'una consecutivamente l'altra, che nel caso più specifico in cui risultino distanziate di una

determinata quantità specificata dal parametro n inserito tra le due parentesi tonde dell'operatore.

L'operatore NEAR(n)

L'operatore di prossimità **NEAR(n)**, simile al precedente operatore **ADJ(n)**, comporta l'individuazione e reperimento di tutti i documenti contraddistinti dalla presenza delle parole chiave ricercate **in qualsiasi ordine esse compaiano**, sia se poste l'una consecutivamente l'altra, che nel caso più specifico in cui esse risultino distanziate di una determinata quantità specificata dal parametro n inserito tra le due parentesi tonde dell'operatore.

L'operatore SAME

L'operatore **SAME** generalizza molto le caratteristiche peculiari possedute dagli operatori di prossimità illustrati in precedenza.
Esso infatti comporta l'individuazione e reperimento di tutte le documentazioni elettroniche nelle quali i termini ricercati (e quindi presenti nella chiave di ricerca) appaiano indistintamente in qualsiasi ordine o distanza l'uno dall'altro.

1.2.3 Caratteri Jolly

Per finire, anche i caratteri Jolly, come gli operatori di prossimità e gli operatori booleani visti nei paragrafi precedenti, identificano risorse avanzate spesso sfruttate in fase di ricerca di documentazione elettronica. Tra i più importanti ed utilizzati in ambiente informatico, si denotano **asterisco ($*$)**, **punto di domanda (?)**, e **virgolette (“ ”)**.

Il carattere Jolly Asterisco ($*$)

Il carattere Jolly Asterisco ($*$) viene di norma utilizzato nelle implementazioni di chiavi di ricerca di tipo avanzato in sostituzione di una (o eventualmente più) lettere che formano termini rappresentanti una generica stringa di caratteri.

Il carattere Jolly Punto di Domanda (?)

Il carattere Jolly punto di domanda (**?**) anch'esso frequentemente sfruttato in implementazioni di keyword di tipologia avanzata, da la possibilità, attraverso il suo impiego, di rimpiazzare più lettere contigue che compongono una parola e quindi può essere impiegato per riprodurre tutte le declinazioni di un eventuale termine.

Il carattere Jolly Virgolette (“ ”)

Infine il carattere Jolly virgolette (**“ ”**) di norma viene implementato in interrogazioni di tipologia avanzata, indicando al motore di ricerca che il loro particolare contenuto deve essere trattato come una frase indivisibile, ossia come una sequenza di caratteri rappresentanti una stringa che deve comparire sottoforma di blocco unitario all'interno del testo delle documentazioni elettroniche eventualmente restituite in seguito la conclusione della specifica sessione di ricerca.

Capitolo 2

Gli Algoritmi di Link Analysis

In passato, come riportato in [5], il processo di reperimento di informazione operato da motori di ricerca considerati di **prima generazione** prevedeva solamente un'analisi dei contenuti testuali delle pagine, confrontandoli con la query formulata nell'interrogazione e verificando la presenza o meno di compatibilità.
Con l'aumentare della quantità di pagine presenti nel web, e con il diversificarsi delle esigenze dell'utenza stessa, si giunse all'introduzione e conseguente uso di motori di ricerca di **seconda generazione**. Essi a differenza dei loro predecessori, oltre che analizzare dati "***on page***" prevedevano lo studio ed esaminazione di una nuova categoria di informazioni, rappresentata da dati estrapolati dalla topologia della struttura del web.
Proprio in tale contesto nacquero e si svilupparono sempre più gli algoritmi di link analysis, quali:

- **Indegree**;
- **HITS** *(Hyperlink Induced Topic Search)*;
- **SALSA** *(Stochastic Approach for Link Structure Analysis)*;
- **PageRank**.

Nei successivi paragrafi verrà riportata una descrizione sintetica di ogni singolo algoritmo, rimandando in seguito la trattazione più approfondita e dettagliata del solo PageRank che identifica l'algoritmio di link analysis utilizzato da Google.

2.1 Indegree

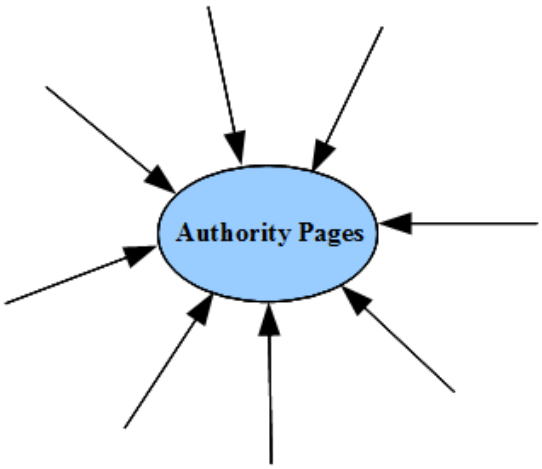
Il capostipite di tutti gli algoritmi di link analysis è stato senza ombra di dubbio **Indegree**. Tale algoritmo, esaminando la struttura topologica ad hyperlink del web, computa un valore di importanza per ciascun documento elettronico prendendo in esame solamente il valore informativo derivante dalla quantità di link afferenti alla pagina stessa e provenienti da altri documenti web. In questo modo, più link entranti presenta una pagina, più elevato è il grado di importanza che la contraddistingue.

2.2 HITS *(Hyperlink Induced Topic Search)*

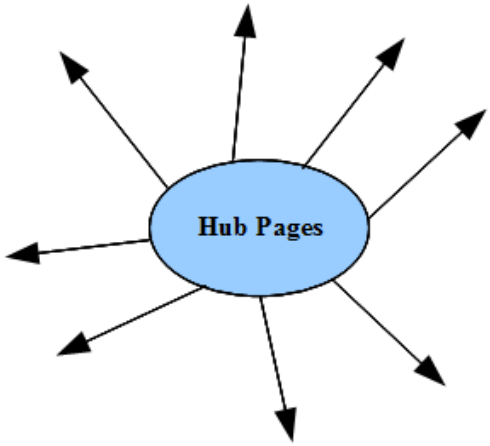
L'algoritmo **HITS**, sviluppato da J.M. Kleinberg, è stato introdotto con lo scopo di implementare un metodo del tutto simile a quello che contraddistingue Indegree evolvendolo attraverso l'esaminazione ed elaborazione approfondita della struttura ad hyperlink che caratterizza il web. Come illustrato in [5], l'innovazione espressa da questo algoritmo *query-dependent*, adatto maggiormente a query di natura *"broad-topic"*[1], riguarda la possibilità di replicare ad una generica interrogazione attraverso la stesura di una classifica implementata sfruttando le potenzialità derivanti dall'impiego di una coppia di nuove entità di pagine con le quali si identificano le documentazioni presenti nel web:

- ***Authority Pages***:
  Documenti caratterizzati dalla presenza di numerosi link che puntano ad essi stessi (link entranti).

[1] Interrogazioni su argomenti di carattere generale che permettevano il reperimento di ingenti quantità di documentazione, a differenza delle query a carattere specifico *(specific queries)* le quali individuavano un ristretto insieme di pagine di interesse.



- ***Hub Pages***:
  Documenti caratterizzati dalla presenza di numerosi link uscenti afferenti ad authority pages.



Concludendo, si può affermare che tale algoritmo concretizza la **relazione di mutuo rinforzo** tra hub ed authority la quale sostiene che [6]:

> *un buon hub rappresenta una pagina afferente a molte buone authority; una buona authority identifica una pagina che viene puntata da molti buoni hub.*

2.3 SALSA *(Stochastic Approach for Link Structure Analysis)*

L'algoritmo **SALSA**, sviluppato da R. Lempel e S. Moran, si basa sui seguenti concetti [16]:

- ***Cammini Random*** *(Random Walks)*: tramite essi in ambito informatico e più precisamente nel contesto del reperimento di informazioni, si modellizza la posizione attualmente occupata da un generico utente all'interno della rete web come indipendente dai documenti elettronici visitati precedentemente.

- ***Visione binaria delle pagine*** *(Autority Pages, Hub Pages)*: essa concretizza i principi affermanti che il valore di authority di un documento elettronico dipende attivamente dalla quantità "punteggio di hub" delle pagine ad esso afferenti, ed in modo analogo ma duale, il valore del punteggio di hub di un documento risulta individuato dalla quantità "punteggio di authority" dei nodi cui esso punta.

2.4 PageRank

L'algoritmo PageRank, sviluppato dai fondatori di Google Sergey Brin e Lawrence Page, trae le proprie origini da un progetto introdotto da Massimo Marchiori, informatico docente all'Università degli studi di Padova.
Tale grandezza, come esposto in [15], costituisce solamente uno dei molteplici fattori esaminati per la valutazione ed individuazione della posizione attribuibile ad una determinata pagina web su di una classifica. Per operare in tale modalità l'algoritmo *query-independent* PageRank, trae beneficio dal concetto di ***Link Popularity***[2] adattandolo e manipolandolo in modo tale da non considerare solamente la quantità dei link afferenti ai documenti elettronici, ma di dipendere attivamente anche dal valore di PageRank posseduto dalle pagine da cui essi partono.

[2] Esso rappresenta un parametro identificante l'affidabilità di un documento elettronico derivante dal conteggio dei link ad esso afferenti.

Capitolo 3

Catene di Markov

La realizzazione di questo capitolo di carattere matematico è di fondamentale importanza dato che, attraverso esso saranno forniti gli elementi necessari alla comprensione dell'algoritmo che caratterizza il funzionamento del motore di ricerca Google.

In particolare, come avremo modo di osservare nel prosieguo del testo, alla base dell'algoritmo di Information Retrieval IR query-independent denominato PageRank si identificano le catene di Markov[1] le quali, attraverso la loro formulazione, permetteranno la determinazione di una serie di valori utilizzabili nella successiva stesura della classifica di pagine web eseguita in ordine di importanza decrescente.

[1] Andrei A. Markov, 1856-1922 scienziato russo fondatore del moderno calcolo delle probabilità.

3.1 Teoria delle Catene di Markov

Definiamo, come riportato in [18, 20], una catena di Markov finita ed omogenea come un processo aleatorio contraddistinto dalla presenza di un insieme finito $S = \{s_1, s_2, \ldots, s_n\}$ di stati distinti (denominati anche vertici o eventi) s_i, le cui coppie ordinate sono caratterizzate da una probabilità di transizione m_{ij} dallo stato s_j allo stato s_i, che risulta essere indipendente dal parametro espressione del tempo t.
Tenendo presente che da s_j si deve passare ad uno degli altri eventi presenti in S compreso s_j stesso, le probabilità di transizione m_{ij} di una catena di Markov dovranno soddisfare la condizione:

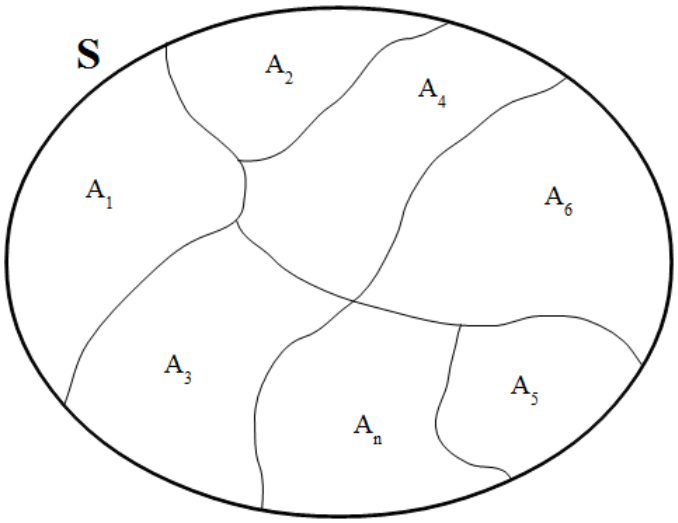
$$\sum_{i=1}^{n} m_{ij} = 1 \ . \tag{3.1}$$

Ora, dato che la probabilità con cui si manifesta uno stato s_i al tempo $t+1$ dipende solamente dalla probabilità con cui si presentano tutti gli stati al tempo t, particolarmente utile è l'impiego di una matrice denominata stocastica[2] (o matrice di transizione) $\mathcal{M} = [m_{ij}]$ l'uso della quale, permette di compattare notevolmente la notazione.
Si introduce inoltre un vettore di probabilità $\mathbf{P}_t = [P_1^{(t)}, P_2^{(t)}, \ldots, P_n^{(t)}]^T$, nel quale il generico elemento $P_j^{(t)}$ individua la probabilità di osservare lo stato s_j al tempo t, in cui vale l'espressione:

$$\sum_{j=1}^{n} P_j^{(t)} = 1 \ . \tag{3.2}$$

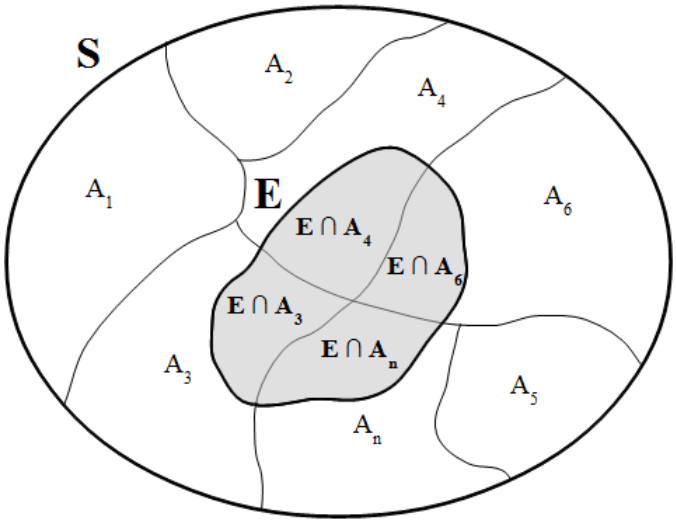
Ora, dato lo spazio di eventi $S = \cup_i A_i$ in cui $A_i \cap A_j = \emptyset$ e dove $\{A_1, \ldots, A_n\}$ rappresenta una "partizione" di S, per un generico evento $E \subset S$

[2] Viene ricordato che in una matrice stocastica la somma dei valori delle colonne vale 1.



si ottiene:

$$P(E) = P(E \cap (\cup_i A_i)) = P(\cup_i E \cap A_i) = \sum_i P(E \cap A_i) = \sum_i P(E|A_i) \cdot P(A_i)\ . \tag{3.3}$$



Siano ora:

- $A_j = s_j$ al tempo t, con $j = 1, \ldots, n$,
- $E = s_i$ al tempo $t + 1$,

si ottiene:

$$P(s_i; t+1) = \sum_j P(s_i; t+1 | s_j; t) \cdot P(s_j; t) \ , \tag{3.4}$$

che riformulato introducendo la notazione $P(s_i; t) = P_i^{(t)}$ e
$P(s_i; t+1 | s_j; t) = m_{ij}(t)$:

$$P_i^{(t+1)} = \sum_{j=1}^{n} m_{ij}(t) \cdot P_j^{(t)} \ , \tag{3.5}$$

che rappresenta un sistema dinamico, discreto, lineare ed omogeneo del tipo:

$$\mathbf{P}_{t+1} = \mathcal{M}(t) \cdot \mathbf{P}_t \ . \tag{3.6}$$

Se, come supponiamo nel seguito, $\mathcal{M} = [m_{ij}]$ è indipendente dal tempo, allora

$$\begin{aligned} \mathbf{P}_1 =& \mathcal{M} \cdot \mathbf{P}_0 \\ \mathbf{P}_2 =& \mathcal{M} \cdot \mathbf{P}_1 = \mathcal{M}^2 \cdot \mathbf{P}_0 \\ \mathbf{P}_3 =& \mathcal{M} \cdot \mathbf{P}_2 = \mathcal{M}^3 \cdot \mathbf{P}_0 \\ & \vdots \\ \mathbf{P}_t =& \mathcal{M}^t \cdot \mathbf{P}_0 \ , \end{aligned} \tag{3.7}$$

ottenendo la soluzione esplicita dell'equazione (3.6) che vale:

$$\mathbf{P}_t = \mathcal{M}^t \cdot \mathbf{P}_0 \ . \tag{3.8}$$

Dato che le catene di Markov sono caratterizzate dalla mancanza di memoria [4], la probabilità di transizione determinata in $t_1 + t_2$ passi (o step) è espressa dalla probabilità condizionata di giungere dallo stato iniziale ad un certo stato k in t_1 step e, consecutivamente, dalla probabilità di passare negli ultimi t_2

passi, dallo stato k allo stato conclusivo.
Analiticamente, poichè

$$\mathcal{M}^{t_1+t_2} = \mathcal{M}^{t_1} \cdot \mathcal{M}^{t_2}, \tag{3.9}$$

si ha

$$m_{ij}^{(t_1+t_2)} = (\mathcal{M}^{t_1} \cdot \mathcal{M}^{t_2})_{ij} = \sum_k m_{ik}^{(t_1)} \cdot m_{kj}^{(t_2)} \tag{3.10}$$

che è nota come **equazione di Chapman-Kolmogorov**.

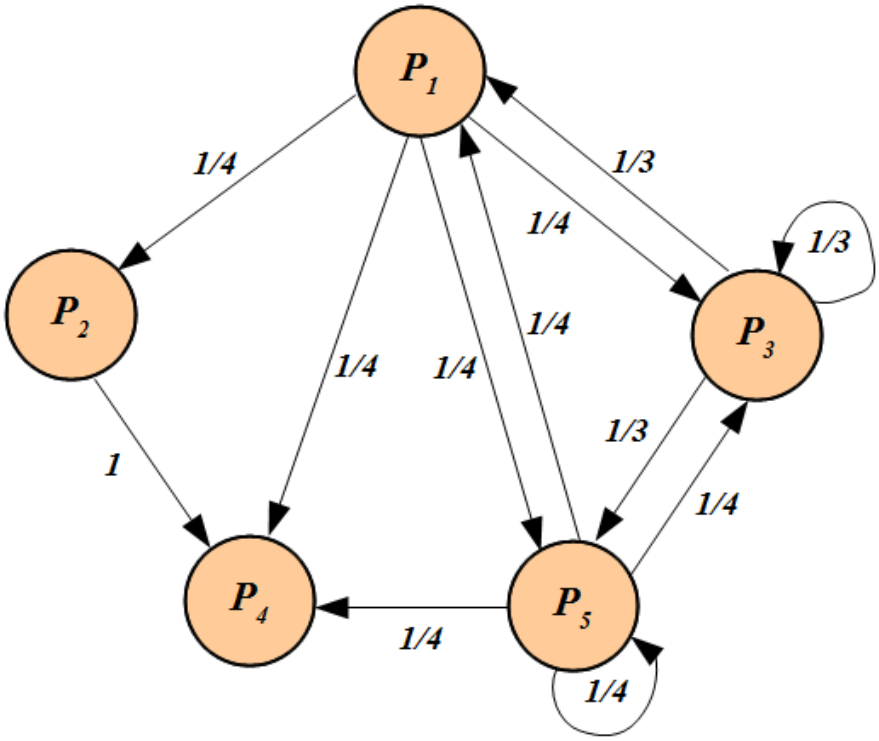
Per semplificare lo studio dei problemi, una catena di Markov può essere proposta anche attraverso una rappresentazione di tipo grafico derivante dalla fruizione di un grafo individuato da nodi collegati tra loro da lati; precisamente essa identifica un grafo di tipo orientato, in quanto la freccia presente in ogni lato individua il verso di percorrenza dello stesso. Un grafo di questo tipo risulta quindi rappresentato dalle seguenti tre componenti:

1. dalla quantità totale dei nodi n;

2. dalla quantità totale dei lati l;

3. dalla coppia ordinata di nodi che ogni lato connette, specificando quale tra essi contraddistingue il nodo di partenza e quale quello di arrivo.

La figura di pagina successiva riporta un esempio di grafo orientato.



Esso risulta formato da 5 nodi $\{P_1, P_2, P_3, P_4, P_5\}$ e da 12 lati orientati ciascuno dei quali presenta un proprio valore di probabilità di transizione, nullo in caso di mancanza di archi di collegamento tra nodi.

3.2 Classificazione degli stati e delle catene di Markov

Di seguito vengono riportate una serie di definizioni introdotte con il fine di agevolare sia la comprensione teorica delle catene di Markov, che le possibili e varie applicazioni pratiche che realmente vengono implementate sfruttando le potenzialità derivanti dalla loro fruizione in campo applicativo e più precisamente nel contesto del recupero di informazioni.
Anzitutto per una più naturale e schematica descrizione, vengono preventivamente introdotte e trattate le definizioni formulabili per la caratterizzazione degli stati delle catene di Markov.

Prendendo in considerazione una catena di Markov e due suoi stati s_i, s_j appartenenti all'insieme degli stati S, come in [2, 4, 18, 20]:

Uno stato s_j viene definito stato accessibile da uno stato s_i se esiste t tale che $m_{ij}^{(t)} > 0$.
Per rappresentare l'accessibilità dello stato s_j da parte di uno stato s_i è stata introdotta e di norma viene impiegata la notazione:

$$s_i \longrightarrow s_j.$$

La condizione di accessibilità può essere definita alternativamente come la capacità, partendo dallo stato s_i, di raggiungere s_j in un numero limitato di transizioni.

Due stati s_i, s_j vengono definiti stati comunicanti se esistono t_1, t_2 tali che $m_{ij}^{(t_1)} > 0$ e $m_{ji}^{(t_2)} > 0$ ovvero se il particolare stato s_j risulta accessibile dallo stato s_i, ed a sua volta lo stato s_i è accessibile dallo stato s_j.
Anche in questo caso è stata introdotta ed è spesso usata la notazione:

$$s_i \longrightarrow s_j \text{ e } s_j \longrightarrow s_i.$$

Un osservazione sicuramente da non tralasciare deriva dal fatto che il concetto di comunicazione definito pocanzi, individua una relazione di equivalenza e quindi sono di validità la proprietà riflessiva, simmetrica e transitiva.

Uno stato viene definito assorbente per una determinata catena di Markov formata da n stati, se per ogni $t \geq 0$ si ottiene $m_{ii}^{(t)} = 1$.
Di conseguenza, uno stato s_i assorbente comporta la presenza nella i-esima colonna della matrice di transizione $\mathcal{M}$, di un unico valore posto ad uno mentre tutti gli altri saranno nulli.

Uno stato s_i viene definito stato transitorio se esiste uno stato s_j con $s_j \neq s_i$, il quale risulta accessibile da s_i mentre non vale il contrario, ovvero s_i non si rivela accessibile da s_j.
Tale definizione può anche essere riproposta attraverso la seguente notazione:

$$\exists\ s_j \in S,\ s_j \neq s_i, \text{ tale che } s_i \longrightarrow s_j \text{ e } s_j \not\longrightarrow s_i$$

Uno stato s_i è definito stato ricorrente se non risulta essere transitorio ovvero, preferendo in questo frangente l'uso diretto della notazione sintetica, se:

$$\forall s_j \in S \text{ risulta } s_i \longrightarrow s_j \Rightarrow s_j \longrightarrow s_i$$

Osservazione:
Dalle ultime definizioni riportate può essere dedotta una coppia di proprietà, la prima delle quali sostiene che *uno stato assorbente s_i ($m_{ii}^{(t)} = 1$) identifica un particolare stato ricorrente* mentre la seconda afferma che *considerando la definizione di ricorrenza, qualsiasi stato presente in una catena di Markov può essere o ricorrente oppure di tipo transitorio.*

Di seguito, verranno introdotte e riportate le definizioni utili alla classificazione delle catene di Markov. Prima di cominciare tale trattazione, è utile soffermarsi definendo la matrice stocastica $\mathcal{M}$

$$\mathcal{M} = [m_{ij}] = \begin{bmatrix} m_{11} & \dots & m_{1n} \\ m_{21} & \dots & m_{2n} \\ \vdots & \ddots & \vdots \\ m_{n1} & \dots & m_{nn} \end{bmatrix} ,$$

strettamente positiva se per ogni $i, j = 1, ..., n$ si ha che $m_{ij} > 0$.
La notazione utilizzata per indicare una matrice di questa tipologia è:

$$\boxed{\mathcal{M} > 0}$$

Una catena di Marcov si definisce regolare se esiste un valore di t tale che la matrice $\mathcal{M}^t$ non presenta alcun elemento nullo.
Quindi, si può affermare che una catena di Markov regolare possiede una matrice di transizione la cui potenza è strettamente positiva.

Una catena di Markov si definisce riducibile se esistono due o più classi di stati comunicanti in cui una non risulta accessibile dall'altra.
Da un punto di vista pratico, una catena è riducibile quando sono presenti alcuni stati nei quali ci si può intrappolare e dai quali non si può più uscire.

Una catena di Markov si definisce irriducibile se esiste un'unica classe di stati che risultano essere tutti comunicanti tra loro.

3.3 Teorema della convergenza markoviana

Molto utile, ai fini della computazione del vettore di PageRank, è la potenzialità insita nel teorema della convergenza markoviana.
Esso, come riportato in [18], afferma che:
In qualsiasi catena di Markov finita con matrice stocastica positiva, ogni distribuzione di probabilità $\mathbf{P}_t$ *converge per* $t \to +\infty$ *verso l'unica distribuzione di probabilità invariante* $\overline{\mathbf{P}}=[\overline{p}_1, \overline{p}_2, \ldots, \overline{p}_n]$ *che soddisfa* $\overline{\mathbf{P}} = \mathcal{M} \cdot \overline{\mathbf{P}}$.

Dimostrazione: Inizialmente, come in [18], si dimostra che la soluzione del seguente sistema rappresenta l'unica distribuzione esistente di probabilità invariante.

$$\begin{cases} m_{11}p_1 + m_{12}p_2 + \cdots + m_{1m}p_m = p_1 \\ m_{21}p_1 + m_{22}p_2 + \cdots + m_{2m}p_m = p_2 \\ \vdots \\ m_{m1}p_1 + m_{m2}p_2 + \cdots + m_{mm}p_m = p_m \end{cases}$$

Esso può essere espresso anche in modo alternativo (per ogni equazione che lo compone, si fanno transitare da secondo a primo membro e successivamente si raggruppano le p_i):

$$\begin{cases} (m_{11} - 1)p_1 + m_{12}p_2 + \cdots + m_{1m}p_m = 0 \\ m_{21}p_1 + (m_{22} - 1)p_2 + \cdots + m_{2m}p_m = 0 \\ \vdots \\ m_{m1}p_1 + m_{m2}p_2 + \cdots + (m_{mm} - 1)p_m = 0 \end{cases}$$

Ora, riportando i coefficienti presenti nelle equazioni su di una matrice $\mathcal{S}$:

$$\mathcal{S} = \mathcal{M} - \mathcal{I} = \begin{bmatrix} m_{11} - 1 & m_{12} & \cdots & m_{1m} \\ m_{21} & m_{22} - 1 & \cdots & m_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ m_{m1} & m_{m2} & \ldots & m_{mm} - 1 \end{bmatrix},$$

si può intuire come la somma degli elementi presenti in ciascuna colonna, dato che $\mathcal{M}$ identifica una matrice stocastica, risulti essere nulla e quindi, come anche il determinante di $\mathcal{S}$ sia nullo. Allora il sistema inizialmente introdotto ha soluzione non nulla $\mathbf{P}=[p_1, p_2, \ldots, p_m]$ in cui tutte le componenti p_i sono

positive.
Moltiplicando queste ultime per un'opportuna costante anch'essa positiva, si può ottenere una soluzione che rappresenta una distribuzione di probabilità invariante:

$$\overline{\mathbf{P}} = [\overline{p}_1, \overline{p}_2, \ldots, \overline{p}_n] \ , \tag{3.11}$$

in cui $\sum_i \overline{p}_i = 1$.

Giunti a questo punto, verrà ora dimostrato che una qualsiasi distribuzione di probabilità $\mathbf{P}_t$ converge verso $\overline{\mathbf{P}}$.
Sfruttando le potenzialità derivanti dall'uso della notazione matriciale, si ha:

$$\mathbf{P}_{t+1} = \mathcal{M} \cdot \mathbf{P}_t \ \ e \ \ \overline{\mathbf{P}} = \mathcal{M} \cdot \overline{\mathbf{P}} \ , \tag{3.12}$$

sottraendo membro a membro le quali, si perviene alla formulazione:

$$(\mathbf{P}_{t+1} - \overline{\mathbf{P}}) = \mathcal{M} \cdot (\mathbf{P}_t - \overline{\mathbf{P}}) \ , \tag{3.13}$$

la cui scrittura può essere notevolmente semplificata attraverso l'ausilio dei vettori:

$$\begin{aligned} \mathbf{Z} &= [z_1, z_2, \cdots, z_m] = \mathbf{P}_t - \overline{\mathbf{P}}, \\ \mathbf{Z}' &= [z'_1, z'_2, \cdots, z'_m] = \mathbf{P}_{t+1} - \overline{\mathbf{P}}. \end{aligned} \tag{3.14}$$

Quindi dalla (3.13) attraveso le (3.14), si perviene a:

$$\mathbf{Z}' = \mathcal{M} \cdot \mathbf{Z}, \tag{3.15}$$

che riproposta in un'altra forma, diventa:

$$z'_i = \sum_k m_{ik} \cdot z_k \ . \tag{3.16}$$

Ora, dato che sia $\mathbf{P}_t$ che $\overline{\mathbf{P}}$ costituiscono distribuzioni di probabilità, si ottiene:

$$\begin{aligned} \sum_i z_i &= 0, \\ \sum_i z'_i &= 0. \end{aligned} \tag{3.17}$$

Di conseguenza, ***considerando nullo*** il vettore **Z'**, dalla seconda delle (3.14) si deriva che $\mathbf{P}_{t+1}=\overline{\mathbf{P}}$ e conseguentemente anche $\mathbf{P}_k=\overline{\mathbf{P}}$ per ogni $k > t+1$ verificando, in questo modo, la convergenza di $\mathbf{P}_t$ verso $\overline{\mathbf{P}}$.

Al contrario di quanto sopra, si supponga che il vettore **Z'** e conseguentemente il vettore **Z *non siano nulli***. In questo caso, dalle formulazioni (3.17), si intuisce la possibile presenza di indici k per cui $z_k > 0$, ed indici per cui $z_k < 0$; analogamente, ciò si verificherà anche per le componenti z'_i di **Z'**.

Introducendo i sottoinsiemi I, ed I' l'ausilio dei quali permette di raggruppare gli indici k tali per cui $z_k > 0$ ed indici i per i quali $z'_i > 0$:

$$\begin{aligned} I &= \{k : z_k > 0\}; \\ I' &= \{i : z'_i > 0\}, \end{aligned} \tag{3.18}$$

direttamente dalla (3.16), sommando membro a membro al variare di i nell'insieme I', si determina la formulazione:

$$\sum_{i\in I'} z'_i = \sum_{i\in I'}\sum_k m_{ik}\cdot z_k = \sum_k (\sum_{i\in I'} m_{ik}) z_k \leq \sum_{k\in I} (\sum_{i\in I'} m_{ik}) z_k. \tag{3.19}$$

nella quale l'uso del $\leq$ deriva dal fatto che la somma relativa ai soli valori contraddistinti dall'indice $k \in I$ risulterà maggiore di quella individuata sommando tutti i termini, dato che in quest'ultima potranno essere presenti valori negativi.
Ora, considerando che:

$$\sum_{i\in I'} m_{ik} \leq \underbrace{1-\alpha}_{=\beta}, \tag{3.20}$$

dove α rappresenta il valore minimo tra tutti gli m_{ik} e β una quantità compresa tra 0 ed 1, si determina direttamente dalla (3.19), l'espressione:

$$\sum_{i \in I'} z'_i \leq \beta \cdot \sum_{k \in I} z_k \ . \tag{3.21}$$

A questo punto dalla (3.21), indicando con il vettore:

- $\mathbf{M}_t$ la somma delle *componenti positive* di $\mathbf{P}_t - \overline{\mathbf{P}}$;
- $\mathbf{M}_{t+1}$ la somma delle *componenti positive* di $\mathbf{P}_{t+1} - \overline{\mathbf{P}}$,

si ottiene:

$$\mathbf{M}_{t+1} \leq \beta \cdot \mathbf{M}_t \ , \tag{3.22}$$

che riproposta, diventa:

$$\mathbf{M}_t \leq \beta^t \cdot \mathbf{M}_0 \ . \tag{3.23}$$

In questo modo:

$$\lim_{t \to \infty} \mathbf{M}_t = 0 \ . \tag{3.24}$$

Per le *componenti negative*, si procede in modo del tutto analogo.

In conclusione si trova che:

$$\boxed{\lim_{t \to \infty} \mathbf{P}_t = \overline{\mathbf{P}}}$$

Capitolo 4

Introduzione al PageRank

Il processo di reperimento di informazioni, come esposto in [5], data l'enorme quantità di documentazioni elettroniche collegate da fitte reti di link e dal comportamento dell'utenza sempre più "***low effort***" (caratterizzato dalla stesura di interrogazioni formate da query sempre più corte ed imprecise), rappresenta per i ricercatori e sviluppatori di software una sfida sempre più ostica ma allo stesso tempo intrigante, volta sia al miglioramento e perfezionamento delle risorse esistenti, sia allo studio, progettazione, elaborazione ed implementazione di nuovi e sempre più evoluti nonchè efficienti algoritmi di ricerca web.
In questo ambito, di particolare interesse è l'algoritmo di link analysis PageRank implementato adottando il Metodo delle Potenze *(Power Method)* il cui studio è affrontato nel seguente capitolo.

4.1 Il Metodo delle Potenze *(Power Method)*

Uno dei più importanti e conosciuti motori di ricerca di seconda generazione in grado di esaminare sia dati *on page* che *off page*, è senza ombra di dubbio Google.

La fama di cui esso gode a livello internazionale è frutto di particolari accorgimenti nonchè scelte strutturali ed implementative apportategli dai realizzatori le quali, partendo da un'interrogazione sottoposta dall'utenza del web, ne permettono la stesura di una classifica (eseguita per ordine di importanza) di documenti elettronici inerenti alla query stessa, in modo estremamente rapido; quasi "istantaneo", dato che è stato stimato che il tempo medio di risposta si aggira in un intorno di un quarto di secondo [10].

In particolare, come in [9], il fulcro del processo implementativo di tale classifica, è indubbiamente rappresentato dall'algoritmo di IR denominato PageRank il quale, traendo beneficio dal concetto di link popularity, considera *una determinata pagina web importante se essa oltre che ricevere link provenienti da pagine importanti, afferisce scarsamente verso altre pagine.* Queste nozioni possono essere rappresentate attraverso l'introduzione e conseguente uso **ricorsivo**[1] della seguente formulazione:

$$r(P) = \sum_{Q \to P} \frac{r(Q)}{|Q|} \ , \tag{4.1}$$

nella quale il fattore $r(P)$ denota il rango (o importanza relativa) posseduta da una determinata pagina web P, mentre $|Q|$ rappresenta la quantità di link uscenti dalla pagina Q ed afferenti ad altre pagine. La somma nella formulazione (4.1) viene eseguita sulle pagine Q che puntano a P.

[1] Dato che tale processo viene eseguito iterativamente per tutte le pagine presenti nel web.

Ora, sfruttando la potenza ed agevolezza computazionale che contraddistinguono il calcolo matriciale, consideriamo:

- L'insieme contenente le n pagine $\{P_1, P_2, ..., P_n\}$ presenti nel web;
- La matrice delle probabilità di transizione $\mathcal{A} = [a_{ij}]$ così composta[2]:

$$a_{ij} = \begin{cases} \frac{1}{|P_j|} & \textit{se la pagina } P_j \textit{ afferisce al documento elettronico } P_i; \\ 0 & \textit{altrimenti}. \end{cases}$$

  nella quale la quantità $|P_j|$ identifica il numero di link uscenti dalla pagina P_j.

In particolare, dato che la matrice $\mathcal{A}$ implementata dal motore di ricerca Google considera miliardi e miliardi di pagine web, difficilmente la potenza di calcolo garantita dai moderni calcolatori è in grado di affrontare il processo di determinazione del vettore di PageRank tramite la **computazione dell'autospazio della matrice $\mathcal{A}$ relativa all'autovalore unitario**[3], e quindi per la determinazione del vettore $\mathbf{r}$ viene adottata una differente e più agevole tecnica, denominata **metodo delle potenze *(power method)***. Partendo dal seguente vettore colonna $\mathbf{r}_0$

$$\mathbf{r}_0 = \begin{bmatrix} r_1^{(0)} \\ \vdots \\ r_n^{(0)} \end{bmatrix} = \begin{bmatrix} \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{bmatrix} ,$$

nel quale tutte le singole pagine web vengono preventivamente inizializzate attribuendo a ciascuna la stessa importanza relativa (individuata dal rapporto $\frac{1}{n}$ dove n rappresenta la quantità totale di pagine) e conoscendo la quantità di link $|P_i|$, attraverso la seguente formulazione:

$$r_i^{(1)} = \sum_{P_j \to P_i} \frac{r_j^{(0)}}{|P_j|} = \sum_{j=1}^{n} a_{ij} \cdot r_j^{(0)} , \tag{4.2}$$

[2] Le procedure di determinazione di $\mathcal{A}$ verrano esaminate nel corso del paragrafo (4.2).

[3] Compiuto impostando e risolvendo il sistema $(\lambda \cdot \mathbf{I} - \mathcal{A}) \cdot \mathbf{r} = 0$ (in cui $\lambda = 1$ identifica l'autovalore posto ad "1" mentre $\mathbf{I}$ rappresenta la matrice diagonale di dimensione n), si determina il valore dell'autovettore $\mathbf{r}$ (vettore di PageRank) relativo all'autovalore λ.

che riproposta in forma matriciale, diventa

$$\mathbf{r}_1 = \mathcal{A} \cdot \mathbf{r}_0 \ , \tag{4.3}$$

viene computata una seconda classifica $\mathbf{r}_1$:

$$\mathbf{r}_1 = \begin{bmatrix} r_1^{(1)} \\ \vdots \\ r_n^{(1)} \end{bmatrix} \ .$$

Il procedimento appena esaminato, viene poi ripetuto completamente nelle successive iterazioni, calcolando di volta in volta una nuova e più aggiornata classifica, la quale risulta diretta espressione dei valori del vettore $\mathbf{r}$ computati al passo precedente.
Si giunge in questo modo alla seguente espressione (riportata in termini matriciali) di carattere del tutto generale, rappresentante il metodo delle potenze:

$$\boxed{\mathbf{r}_{t+1} = \mathcal{A} \cdot \mathbf{r}_t}$$

4.2 La Matrice delle Probabilità di Transizione

Durante il processo di calcolo del vettore di PageRank, data la potenza ed estrema agevolezza computazionale godute dalla notazione matriciale, parecchia attenzione viene posta alla determinazione e ai conseguenti aggiustamenti apportati alla matrice delle probabilità di transizione.
In particolare, nel preliminare processo di individuazione della matrice $\mathcal{C} = [c_{ij}]$ chiamata anche "matrice di adiacenza", i componenti c_{ij} in essa presenti, vengono computati nel seguente modo:

$$c_{ij} = \begin{cases} 1 & \textit{se la pagina } P_j \textit{ punta alla pagina } P_i; \\ 0 & \textit{altrimenti.} \end{cases}$$

In questo frangente, un possibile problema riscontrabile è rappresentato dall'individuazione nella struttura topologica del grafo web, di pagine isolate (frequentemente chiamate anche *dead-ends* o *dangling nodes*).

La presenza di questi dead-ends[4], sopratutto durante la fase preliminare di computazione della matrice di transizione $\mathcal{A}$, derivata dalla normalizzazione per colonne della matrice $\mathcal{C}$, comporta l'individuazione di colonne corrispondenti a tali pagine isolate composte interamente ed esclusivamente di valori nulli, le quali "declassificano" la matrice $\mathcal{A}$ non permettendole di essere considerata una matrice stocastica.
Per ovviare a questo inconveniente, come in [13], risulta di considerevole utilità l'introduzione e conseguente applicazione, in sostituzione dei valori nulli presenti nelle sopracitate colonne, della quantità $\frac{\mathbf{e}^T}{n}$ (in cui $\mathbf{e}$ identifica un vettore riga formato da valori unitari).
La matrice **stocastica** $\mathcal{A}'$ ottenuta attraverso questo processo di normalizzazione, non identificherà la matrice definitivamente utilizzabile per la computazione dell'algoritmo PageRank dato che, tale $\mathcal{A}'$ può rappresentare una matrice riducibile[5].
Per il calcolo del vettore $\mathbf{r}$ di PageRank, un passo fondamentale risulta dunque individuato dalla trasformazione della matrice stocastica $\mathcal{A}'$ in una

[4] Viene stimato che circa il 25% delle pagine web non possieda alcun link.

[5] In analogia a quanto già notato in precedenza nella teoria markoviana, una catena si definisce riducibile se in essa sono presenti stati "trappola".

matrice irriducibile $\mathcal{A}''$ determinata sommando ad $\mathcal{A}'$ la matrice di perturbazione $\mathcal{E} = \frac{(\mathbf{e} \cdot \mathbf{e}^T)}{n}$:

$$\mathcal{A}'' = d \cdot \mathcal{A}' + (1-d) \cdot \mathcal{E} = d \cdot \mathcal{A}' + (1-d) \cdot \frac{(\mathbf{e} \cdot \mathbf{e}^T)}{n} \ . \tag{4.4}$$

Tale $\mathcal{A}''$ rappresenta una matrice irriducibile il cui parametro d (denominato *damping factor* o fattore di smorzamento) identifica un valore compreso tra 0 ed 1, la cui descrizione e completa trattazione sono rimandati al paragrafo (4.3).

Solo in seguito, nonostante l'esistenza di numerose altre tecniche adottabili per il calcolo della matrice irriducibile, venne elaborata una nuova e più moderna matrice di perturbazione $\mathcal{E} = (\mathbf{e} \cdot \mathbf{v}^T)$ il cui uso consentiva la formulazione:

$$\mathcal{A}'' = d \cdot \mathcal{A}' + (1-d) \cdot \mathcal{E} = d \cdot \mathcal{A}' + (1-d) \cdot (\mathbf{e} \cdot \mathbf{v}^T) \ , \tag{4.5}$$

nella quale il vettore di probabilità strettamente positivo $\mathbf{v}^T$ introdotto, permetteva a Google di variare il valore di PageRank di determinate pagine web aumentandolo o diminuendolo a proprio piacimento in base ad esigenze di mercato.

La matrice stocastica ed irriducibile $\mathcal{A}''$, computata attraverso i passi precedentemente descritti, riveste estrema importanza dato che rappresenta la matrice adottata per il calcolo del vettore di PageRank $\mathbf{r}$.

4.3 Il Fattore di Smorzamento d *(Damping factor)*

Il parametro d riveste un ruolo di primaria importanza nella robusta, affidabile ed indipendente[6] computazione dell'algoritmo PageRank di Google. Esso, come in [16, 21], da un punto di vista puramente algebrico-matematico, rappresenta la probabilità derivante dalla decisione compiuta dall'utente di esplorare nuove pagine web collegate tramite outlink al documento elettronico correntemente visitato; tale parametro, generalmente compreso tra 0 ed 1, nella documentazione originale risulta posto a 0.85.
Contrariamente, la grandezza $1 - d$ (che nel caso specifico vale 0.15) esprime la possibilità che il generico web surfer decida di intraprendere un percorso alternativo, compiuto verso altre pagine, non attenendosi in questo modo a quello "indicato" dai link uscenti presenti nella pagina web correntemente visitata.

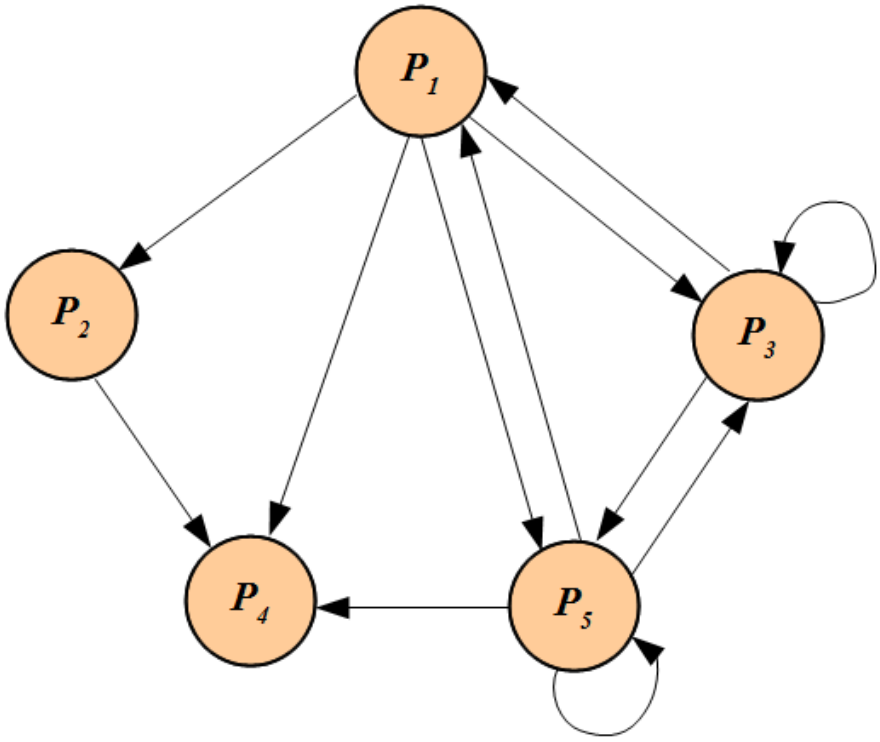
4.3.1 Esempio di calcolo della matrice di transizione

Dopo avere esaminato dal punto di vista teorico tutti i passaggi ed operazioni che interessano la matrice di transizione partendo dal proprio stato iniziale di matrice di adiacenza fino al raggiungimento dello stato finale (nel quale essa risulta utilizzabile nella successiva computazione del vettore $\mathbf{r}$ di PageRank), risulta ora di estrema utilità soffermarci e considerare un piccolo esempio in cui verranno riepilogati tutti i calcoli eseguiti su di essa.

Di seguito è riportato il grafo di un ristretto insieme di pagine web formato da 5 nodi $\{P_1, P_2, P_3, P_4, P_5\}$ e 12 link; esso per conformità di trattazione è il medesimo introdotto ed utilizzato come esempio durante la descrizione di

[6] Indipendenza riferita ai contenuti insiti nella chiave di ricerca.

un grafo orientato eseguita nella fase finale del paragrafo (3.1):



Inizialmente, dalla topologia del mini web in esame, viene individuata la matrice di adiacenza $\mathcal{C}$ formata in modo esclusivo da valori nulli o unitari; essa varrà:

$$\mathcal{C} = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix} .$$

Una volta determinata la matrice di adiacenza, dividendo tutti gli elementi della $j-$esima colonna per la quantità di 1 in essa (se presenti), viene normalizzata per colonne la matrice $\mathcal{C}$, ottenendo in questo modo la nuova matrice

$\mathcal{A}$:

$$\mathcal{A} = \begin{bmatrix} 0 & 0 & \frac{1}{3} & 0 & \frac{1}{4} \\ \frac{1}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{3} & 0 & \frac{1}{4} \\ \frac{1}{4} & 1 & 0 & 0 & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{1}{3} & 0 & \frac{1}{4} \end{bmatrix} .$$

Esaminando la matrice di transizione $\mathcal{A}$ si nota la presenza di una colonna composta interamente da valori nulli (precisamente la quarta colonna della matrice). Questa caratteristica di $\mathcal{A}$ indica che essa non identifica una matrice stocastica; per fare in modo che lo diventi, si interviene algebricamente sostituendo a tutti gli elementi della quarta colonna posti a 0, degli uni e successivamente normalizzando la stessa.
In questo modo, sostituendo i valori nulli presenti nella quarta colonna della matrice $\mathcal{A}$ con la quantità $\frac{1}{n}$ (dove n esprime il numero di nodi presenti nel grafo in esame), otteniamo la matrice $\mathcal{A}'$:

$$\mathcal{A}' = \begin{bmatrix} 0 & 0 & \frac{1}{3} & \frac{1}{5} & \frac{1}{4} \\ \frac{1}{4} & 0 & 0 & \frac{1}{5} & 0 \\ \frac{1}{4} & 0 & \frac{1}{3} & \frac{1}{5} & \frac{1}{4} \\ \frac{1}{4} & 1 & 0 & \frac{1}{5} & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{1}{3} & \frac{1}{5} & \frac{1}{4} \end{bmatrix} .$$

La nuova matrice $\mathcal{A}'$, identifica a tutti gli effetti una matrice stocastica, ma non ancora irriducibile.
Per renderla irriducibile, scegliendo di considerare il fattore di smorzamento come grandezza variabile, viene sommata ad $\mathcal{A}'$ la matrice di perturbazione $\mathcal{E} = \frac{(\mathbf{e} \cdot \mathbf{e}^T)}{n}$, ottenendo:

$$\mathcal{A}'' = d \cdot \mathcal{A}' + (1-d) \cdot \frac{(\mathbf{e} \cdot \mathbf{e}^T)}{n} =$$

$$= \begin{bmatrix} \frac{1-d}{n} & \frac{1-d}{n} & \frac{3+(n-3)\cdot d}{3\cdot n} & \frac{5+(n-5)\cdot d}{5\cdot n} & \frac{4+(n-4)\cdot d}{4\cdot n} \\ \frac{4+(n-4)\cdot d}{4\cdot n} & \frac{1-d}{n} & \frac{1-d}{n} & \frac{5+(n-5)\cdot d}{5\cdot n} & \frac{1-d}{n} \\ \frac{4+(n-4)\cdot d}{4\cdot n} & \frac{1-d}{n} & \frac{3+(n-3)\cdot d}{3\cdot n} & \frac{5+(n-5)\cdot d}{5\cdot n} & \frac{4+(n-4)\cdot d}{4\cdot n} \\ \frac{4+(n-4)\cdot d}{4\cdot n} & \frac{1+(n-1)\cdot d}{n} & \frac{1-d}{n} & \frac{5+(n-5)\cdot d}{5\cdot n} & \frac{4+(n-4)\cdot d}{4\cdot n} \\ \frac{4+(n-4)\cdot d}{4\cdot n} & \frac{1-d}{n} & \frac{3+(n-3)\cdot d}{3\cdot n} & \frac{5+(n-5)\cdot d}{5\cdot n} & \frac{4+(n-4)\cdot d}{4\cdot n} \end{bmatrix} .$$

Ora, sostituendo la variabile damping factor d con l'usuale valore cui esso risulta fissato ($d = 0.85$) ed attribuendo alla grandezza n la quantità totale di nodi presenti nel grafo ($n = 5$), otteniamo:

$$\mathcal{A}'' = \begin{bmatrix} \frac{3}{100} & \frac{3}{100} & \frac{47}{150} & \frac{1}{5} & \frac{97}{400} \\ \frac{97}{400} & \frac{3}{100} & \frac{3}{100} & \frac{1}{5} & \frac{3}{100} \\ \frac{97}{400} & \frac{3}{100} & \frac{47}{150} & \frac{1}{5} & \frac{97}{400} \\ \frac{97}{400} & \frac{22}{25} & \frac{3}{100} & \frac{1}{5} & \frac{97}{400} \\ \frac{97}{400} & \frac{3}{100} & \frac{47}{150} & \frac{1}{5} & \frac{97}{400} \end{bmatrix} .$$

La matrice strettamente positiva $\mathcal{A}''$ risultante, frutto delle computazioni descritte in precedenza, è sia stocastica che irriducibile e rappresentera la matrice considerata nella computazione del vettore $\mathbf{r}$ di PageRank.

4.4 Il Calcolo Iterativo del PageRank

Come già citato più volte nei precedenti paragrafi, ciò che contraddistingue una qualsiasi sessione volta al reperimento di informazioni contenute nella documentazione elettronica, è l'individuazione e conseguente valutazione del posizionamento delle pagine web determinate come risultato della ricerca svolta su di una parola chiave inoltrata dall'utente.
Analiticamente, come descritto in [9, 13], in seguito alla determinazione della matrice stocastica ed irriducibile:

$$\mathcal{A}'' = d \cdot \mathcal{A}' + (1-d) \cdot \mathcal{E} = d \cdot \mathcal{A}' + (1-d) \cdot \ (\mathbf{e} \cdot \mathbf{v}^T) \ , \tag{4.6}$$

nella quale la matrice di perturbazione scelta per l'implementazione vale $\mathcal{E} = (\mathbf{e} \cdot \mathbf{v}^T)$, e partendo dal seguente vettore colonna:

$$\mathbf{r}_0 = \frac{\mathbf{e}^T}{n} = \begin{bmatrix} r_1^{(0)} \\ \vdots \\ r_n^{(0)} \end{bmatrix} = \begin{bmatrix} \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{bmatrix} \ ,$$

attraverso cui tutti i documenti elettronici vengono inizializzati con la stessa importanza relativa, eseguendo iterativamente[7] la formulazione:

$$\begin{aligned} \mathbf{r}_{t+1} =& \mathcal{A}'' \cdot \mathbf{r}_t = \\ =& [d \cdot \mathcal{A}' + (1-d) \cdot \mathcal{E}] \cdot \mathbf{r}_t = \\ =& d \cdot \mathbf{r}_t \cdot \mathcal{A}' + (1-d) \cdot \mathcal{E} \cdot \mathbf{r}_t = \\ =& d \cdot \mathbf{r}_t \cdot \mathcal{A}' + (1-d) \cdot (\underbrace{\mathbf{e} \cdot \mathbf{r}_t}_{=1} \cdot \mathbf{v}^T) = \\ =& d \cdot \mathbf{r}_t \cdot \mathcal{A}' + (1-d) \cdot \mathbf{v}^T \ , \end{aligned} \tag{4.7}$$

viene implementato il metodo delle potenze il cui fine consiste nel computare il vettore $\mathbf{r}$ di PageRank.

[7] Compiuta fino il raggiungimento del grado di convergenza desiderato.

4.5 Aggiornamento dell'algoritmo di PageRank

Il processo di aggiornamento dell'algoritmo utilizzato per il calcolo del vettore di PageRank rappresenta un'importante azione volta al miglioramento della ricerca e seguente estrapolazione di informazioni utili all'utente finale, operata su grandi moli di documenti elettronici presenti nel web.

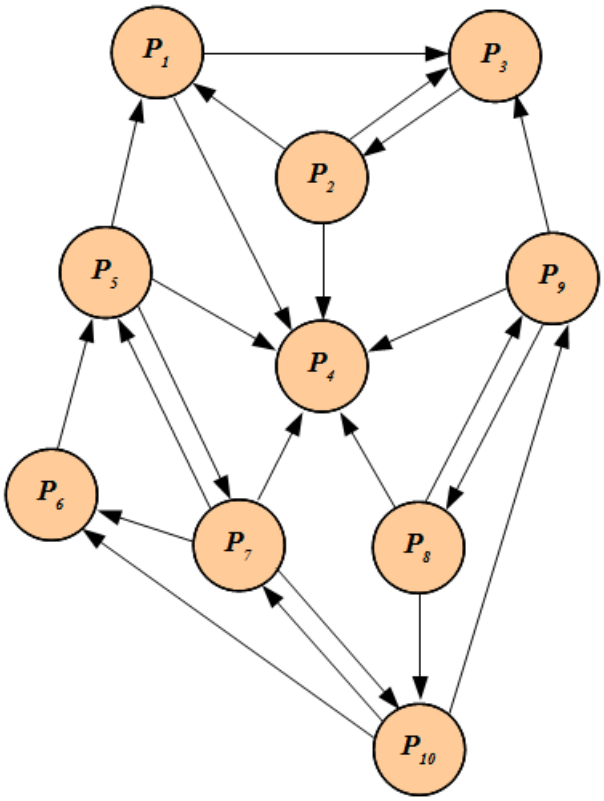
Si calcola che Google annualmente apporti al proprio algoritmo all'incirca 500 cambiamenti, tutti finalizzati al reperimento in maniera sempre più precisa e rapida di dati utili all'utenza finale. Di particolare risalto, in questo ambito, godono metodi nuovi ed alternativi, rispetto quello delle potenze o quello derivante da modificazioni in esso apportate, volti alla determinazione ottimizzata del vettore $\mathbf{r}$.
Oggetto di studio da parte dei ricercatori, data l'importanza e delicatezza che rivestono tali processi, è stata sia l'ottimizzazione del ricalcolo del PageRank che rappresenta un'operazione particolarmente onerosa anche se esguita da calcolatori di ultimissima generazione, che la possibile validità di una computazione ossia quanto essa possa essere sfruttata prima che mutazioni del web obblighino ad un aggiornamento del vettore $\mathbf{r}$ tramite il proprio ricalcolo [13].

In conclusione, negli ultimi mesi sono stati adottati nel processo di determinazione del PageRank nuovi accorgimenti, l'adozione dei quali provvede al posizionamento nei vertici della classifica risultante dalla computazione dell'algoritmo di pagine di qualità [17] migliorando il reperimento di documenti elettronici ufficiali [22] e di pagine le cui informazioni riportate siano il più aggiornate possibili [10], valorizzando in questo modo documenti elettronici di levatura superiore il cui valore aggiunto andrà ad incidere sulla qualità finale della ricerca.

4.6 Esempo Conclusivo

Di seguito andremo ad esaminare un concreto e completo esempio di calcolo del vettore **r** di PageRank e di determinazione della classifica finale computata per una coppia di query formulate dall'utenza. Per semplicità di trattazione, è stata prevista la presenza di un numero ristretto di nodi (documenti elettronici) e lati (link), sopratutto se si considera la quantità totale realmente presente nel web che risulta nell'ordine dei miliardi.

Nella fattispecie, il grafo orientato corrispondente alla struttura topologica del mini web preso in esame in questo esempio, è contraddistinto dalla presenza di 10 nodi $\{P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9, P_{10}\}$ interconnessi tra loro da 23 lati:



Partendo da esso, il processo di computazione della matrice stocastica ed irriducibile, prevede in prima istanza l'individuazione della matrice di

adiacenza $\mathcal{C}$ formata solo ed esclusivamente da 0 o da 1, il cui generico elemento c_{ij}, viene posto ad uno se e solo se risulta essere presente un arco che collega il nodo j al nodo i, in caso contrario verrà settato a zero.
Nel caso in esame, otteniamo:

$$\mathcal{C} = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix},$$

che come logico aspettarsi conterrà 23 elementi posti ad 1 (tanti quanti sono i collegamenti presenti nel mini web). A questo punto, dividendo tutti gli 1 individuati nella $j-$esima colonna per la quantità di 1 presenti nella stessa, normalizziamo la matrice $\mathcal{C}$ per colonne, ottenendo in questo modo la nuova rappresentazione matriciale $\mathcal{A}$ così formata:

$$\mathcal{A} = \begin{bmatrix} 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ \frac{1}{2} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{4} & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{3} & 0 & 0 \end{bmatrix}.$$

Da una seguente esaminazione della matrice, si denota la presenza di una colonna composta esclusivamente da valori nulli (più precisamente la quarta colonna); tale evento, comporta la classificazione di $\mathcal{A}$ come matrice di tipologia non stocastica. Per fare in modo che lo diventi, si interviene algebricamente operando la sostituzione di tutti gli elementi nulli presenti in tale colonna con degli 1, ed implementando successivamente un nuovo processo

di normalizzazione su di essa.
Tramite questi interventi, otteniamo da $\mathcal{A}$ la nuova matrice stocastica $\mathcal{A}'$:

$$\mathcal{A}' = \begin{bmatrix} 0 & \frac{1}{3} & 0 & \frac{1}{10} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \frac{1}{10} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{3} & 0 & \frac{1}{10} & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ \frac{1}{2} & \frac{1}{3} & 0 & \frac{1}{10} & \frac{1}{3} & 0 & \frac{1}{4} & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{10} & 0 & 1 & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{10} & 0 & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{10} & \frac{1}{3} & 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{10} & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{10} & 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{10} & 0 & 0 & \frac{1}{4} & \frac{1}{3} & 0 & 0 \end{bmatrix} .$$

Essa identifica a tutti gli effetti una matrice stocastica, ma non ancora irriducibile.
Per ottenere l'irriducibilità, decidendo anche in questa circostanza di fissare il parametro damping factor d con il valore 0.85 documentato, ed adottando la matrice di perturbazione $\mathcal{E} = \frac{(\mathbf{e} \cdot \mathbf{e}^T)}{n}$ in cui n come consuetudine identifica il numero totale di pagine presenti nel mini web in esame, si determina la seguente formulazione:

$$\mathcal{A}'' = d \cdot \mathcal{A}' + (1-d) \cdot \mathcal{E} = d \cdot \mathcal{A}' + (1-d) \cdot \frac{(\mathbf{e} \cdot \mathbf{e}^T)}{n} =$$

$$= \begin{bmatrix} \frac{3}{200} & \frac{179}{600} & \frac{3}{200} & \frac{1}{10} & \frac{179}{600} & \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{3}{200} \\ \frac{3}{200} & \frac{3}{200} & \frac{173}{200} & \frac{1}{10} & \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{3}{200} \\ \frac{11}{25} & \frac{179}{600} & \frac{3}{200} & \frac{1}{10} & \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{179}{600} & \frac{3}{200} \\ \frac{11}{25} & \frac{179}{600} & \frac{3}{200} & \frac{1}{10} & \frac{179}{600} & \frac{3}{200} & \frac{91}{400} & \frac{179}{600} & \frac{179}{600} & \frac{3}{200} \\ \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{1}{10} & \frac{3}{200} & \frac{173}{200} & \frac{91}{400} & \frac{3}{200} & \frac{3}{200} & \frac{3}{200} \\ \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{1}{10} & \frac{3}{200} & \frac{3}{200} & \frac{91}{400} & \frac{3}{200} & \frac{3}{200} & \frac{179}{600} \\ \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{1}{10} & \frac{179}{600} & \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{179}{600} \\ \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{1}{10} & \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{179}{600} & \frac{3}{200} \\ \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{1}{10} & \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{179}{600} & \frac{3}{200} & \frac{179}{600} \\ \frac{3}{200} & \frac{3}{200} & \frac{3}{200} & \frac{1}{10} & \frac{3}{200} & \frac{3}{200} & \frac{91}{400} & \frac{179}{600} & \frac{3}{200} & \frac{3}{200} \end{bmatrix} .$$

La matrice stocastica ed irriducibile $\mathcal{A}''$ così ottenuta verrà utilizzata nel calcolo del vettore PageRank $\mathbf{r}$ affrontato implementando il metodo delle

potenze; nel caso specifico, usufruendo della formulazione $\mathbf{r}_{t+1} = \mathcal{A}'' \cdot \mathbf{r}_t$ (esaminata nel corso del paragrafo (4.1)) per 15 iterazioni successive (riportate nella seguente tabella i cui valori che la compongono sono stati determinati tramite il software Numbers presente nel pacchetto Apple iWork '09), partendo dallo stato iniziale $\mathbf{r}_0$ si giunge allo stato finale prefisso ($\mathbf{r}_{15}$ nel nostro caso), ottenendo il vettore $\mathbf{r}$ di PageRank.

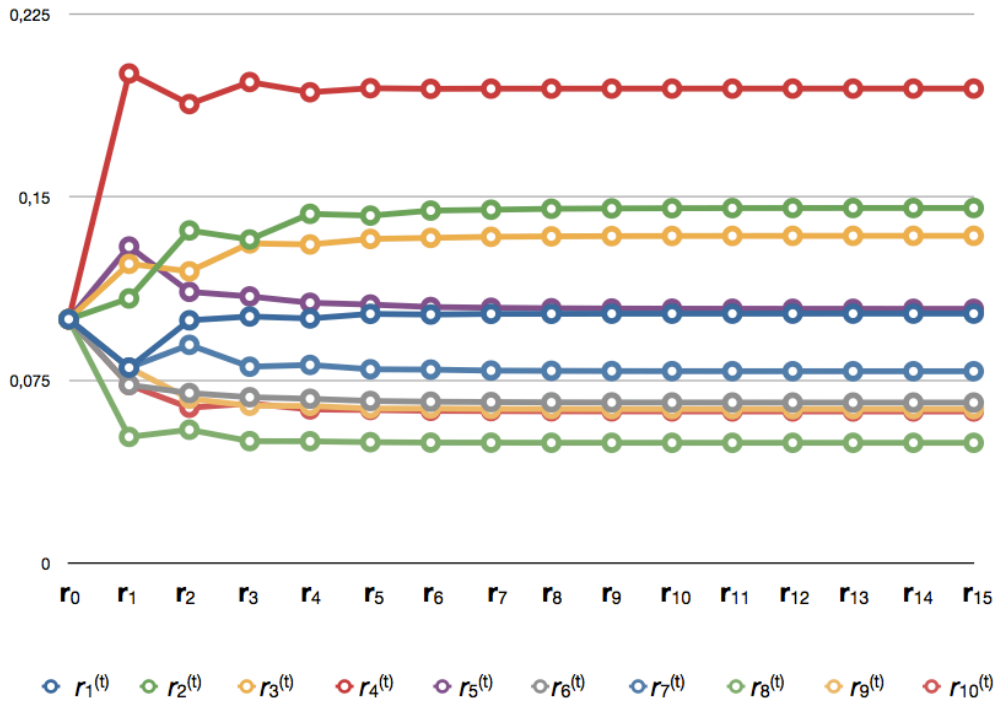
$\mathbf{r}_0$	$\mathbf{r}_1 = \mathcal{A}'' \cdot \mathbf{r}_0$	$\mathbf{r}_2 = \mathcal{A}'' \cdot \mathbf{r}_1$	$\mathbf{r}_3 = \mathcal{A}'' \cdot \mathbf{r}_2$
$\frac{1}{10}$	0.080166667	0.09955375	0.101116357
$\frac{1}{10}$	0.1085	0.13631625	0.132624676
$\frac{1}{10}$	0.122666667	0.119575972	0.131027129
$\frac{1}{10}$	0.200583333	0.18806	0.197074559
$\frac{1}{10}$	0.12975	0.111205833	0.109331046
$\frac{1}{10}$	0.073083333	0.069791944	0.068076375
$\frac{1}{10}$	0.080166667	0.089519028	0.080561901
$\frac{1}{10}$	0.051833333	0.054763472	0.050093848
$\frac{1}{10}$	0.080166667	0.067442639	0.064569899
$\frac{1}{10}$	0.073083333	0.063771111	0.065524210

$\mathbf{r}_4 = \mathcal{A}'' \cdot \mathbf{r}_3$	$\mathbf{r}_5 = \mathcal{A}'' \cdot \mathbf{r}_4$	$\mathbf{r}_6 = \mathcal{A}'' \cdot \mathbf{r}_5$	$\mathbf{r}_7 = \mathcal{A}'' \cdot \mathbf{r}_6$
0.100305459	0.102189110	0.101915235	0.102193013
0.143124397	0.142403375	0.144463557	0.144811555
0.130597585	0.132854932	0.133290460	0.133728543
0.192887376	0.194551346	0.194290164	0.194371299
0.106735660	0.105990874	0.104989440	0.104685256
0.067435934	0.066538463	0.066239267	0.066091862
0.081293660	0.079505330	0.079375132	0.078971654
0.050046142	0.049673200	0.049512466	0.049483227
0.064509787	0.063443300	0.063418457	0.063253178
0.063063998	0.062850070	0.062505820	0.062410411

$\mathbf{r}_8 = \mathcal{A}'' \cdot \mathbf{r}_7$	$\mathbf{r}_9 = \mathcal{A}'' \cdot \mathbf{r}_8$	$\mathbf{r}_{10} = \mathcal{A}'' \cdot \mathbf{r}_9$	$\mathbf{r}_{11} = \mathcal{A}'' \cdot \mathbf{r}_{10}$
0.102212324	0.102261647	0.102273464	0.102283830
0.145190822	0.145340740	0.145433283	0.145479805
0.133905266	0.134012582	0.134067061	0.134095478
0.194367812	0.194383383	0.194385913	0.194387147
0.104481120	0.104368236	0.104311521	0.104281089
0.065985987	0.065938411	0.065912444	0.065898803
0.078865333	0.078782512	0.078742161	0.078721025
0.049443294	0.049434945	0.049425985	0.049423159
0.063224758	0.063188461	0.063177728	0.063170123
0.062323284	0.062289081	0.062270439	0.062259541

$\mathbf{r}_{12} = \mathcal{A}'' \cdot \mathbf{r}_{11}$	$\mathbf{r}_{13} = \mathcal{A}'' \cdot \mathbf{r}_{12}$	$\mathbf{r}_{14} = \mathcal{A}'' \cdot \mathbf{r}_{13}$	$\mathbf{r}_{15} = \mathcal{A}'' \cdot \mathbf{r}_{14}$
0.102288494	0.102290977	0.102292328	0.102293015
0.145504064	0.145517408	0.145524171	0.145527876
0.134111015	0.134118936	0.134123260	0.134125480
0.194388770	0.194389116	0.194389470	0.194389594
0.104265108	0.104256426	0.104251982	0.104249587
0.065891328	0.065887530	0.065885490	0.065884409
0.078709419	0.078703559	0.078700305	0.078698656
0.049421109	0.049420175	0.049419662	0.049419392
0.063166339	0.063164426	0.063163367	0.063162832
0.062254354	0.062251444	0.062249964	0.062249157

L'ultima colonna della tabella, riporta i valori del vettore $\mathbf{r}$ risultanti alla quindicesima iterazione. Come si può notare osservando e confrontando i valori computati ad ogni passo di esecuzione, il metodo delle potenze garantisce una convergenza assai rapida; ciò è ancora più apprezzabile se si riportano gli andamenti ottenuti in forma grafica.



Questa rappresentazione, è molto più intuitiva della forma tabellare precedentemete riportata. In particolare, si riescono a comprendere in modo più agevole le variazioni che subiscono gli elementi che contraddistinguono il vettore di PageRank computato per l'esempio in esame; si nota come dal valore comune inizialmente fissato $\mathbf{r}_0$ sino al raggiungimento della terza iterazione eseguita implementando il metodo delle potenze, si susseguano variazioni di PageRank che solo da $\mathbf{r}_4$ in poi, tende ad assestarsi definitivamente convergendo.
Inoltre da un esame incrociato eseguito sfruttando sia la precisione dei valori presenti nella tabella, che l'intuitività espressa dal diagramma soprariportato, si può facilmente intuire che la pagina P_4 rappresenta il nodo del mini web esaminato che con $r_4^{(t)}$ (t=15 dato che ci riferiamo all'ultima iterazione calcolata attraverso l'uso del power method) uguale al valore 0.194389594, identifica il documento elettronico con valore di PageRank più elevato; di seguito poi, in ordine decrescente, si notano: P_2 con $r_2^{(15)} = 0.145527876$, P_3 con $r_3^{(15)} = 0.134125480$, P_5 con $r_5^{(15)} = 0.104249587$, P_1 con $r_1^{(15)} =$

0.102293015, P_7 con $r_7^{(15)} = 0.078698656$, P_6 con $r_6^{(15)} = 0.065884409$, P_9 con $r_9^{(15)} = 0.063162832$, P_{10} con $r_{10}^{(15)} = 0.062249157$ ed infine la pagina P_8 che costituisce la pagina che con $r_8^{(15)} = 0.049419392$ rappresenta il documento elettronico con valore di PageRank minimo tra quelle presenti nell'esempio.

Considerando ora, $\mathbf{r}_{15}$ come vettore di PageRank $\mathbf{r}$ risultante dalla particolare conformazione del grafo presa in esame:

$$\mathbf{r} = \begin{bmatrix} 0.102293015 \\ 0.145527876 \\ 0.134125480 \\ 0.194389594 \\ 0.104249587 \\ 0.065884409 \\ 0.078698656 \\ 0.049419392 \\ 0.063162832 \\ 0.062249157 \end{bmatrix},$$

simuliamo [13], il comportamento di un determinato utente il quale effettua un'interrogazione eseguita tramite la formulazione di una concreta chiave di ricerca composta inizialmente da due termini "`studenti ingegneria`", per ognuno dei quali viene successivamente estrapolato da un indice (è stato stimato che la dimensione dell'indice realmente utilizzato da Google è di circa 100 milioni di gigabyte [10]) contenente per ogni parola le referenze web, ossia la lista dei documenti elettronici in cui si denota l'effettiva presenza del termine stesso.
Ipotizzando che, per l'esempio in esame, in tale file siano disponibili i seguenti dati:

$$\begin{array}{ll} \vdots & \\ \texttt{corsi} & : P_1, P_3, P_5, P_6 \\ \texttt{frequentanti} & : P_1 \\ \texttt{ingegneria} & : P_2, P_4, P_5 \\ \texttt{matematici} & : P_1 \\ \texttt{studenti} & : P_3, P_4, P_5, P_6 \\ \vdots & \end{array}$$

otteniamo, per la keyword formulata precedentemente, un insieme congiunto $\{P_2, P_3, P_4, P_5, P_6\}$ che comparato al vettore $\mathbf{r}$ precedentemente determinato, permetterà la stesura di una classifica dei documenti elettronici eseguita in ordine di importanza decrescente, computata tramite l'ordinamento dei corrispondenti valori di PageRank loro associato, ottenendo in questo modo:

$$r_4 = 0.194389594$$
$$r_2 = 0.145527876$$
$$r_3 = 0.134125480$$
$$r_5 = 0.104249587$$
$$r_6 = 0.065884409$$

Come si può notare, la pagina web P_4 identifica il documento elettronico che gode di maggiore importanza tra quelli rilevanti; essa precede P_2, P_3, P_5 ed infine P_6 che individua quella di minore rilievo.

Di considerevole utilità può essere una successiva sperimentazione compiuta considerando una seconda keyword, formulata da un generico web surfer. Essa risulta formata in totale da tre termini "`frequentanti corsi matematici`", per ognuno dei quali, viene estrapolato dal file introdotto in precedenza, la lista di pagine web contraddistinte dalla presenza del singolo termine ottenendo, in questo modo, l'insieme congiunto $\{P_1, P_3, P_5, P_6\}$ che comparato al vettore di PageRank calcolato inizialmente, permetterà la stesura della seguente classifica implementata in ordine di importanza decrescente:

$$r_3 = 0.134125480$$
$$r_5 = 0.104249587$$
$$r_1 = 0.102293015$$
$$r_6 = 0.065884409$$

In questo caso, la pagina web che gode di maggiore importanza nell'insieme composto da quelle rilevanti risulta essere P_3, che precede P_5, P_1 ed infine P_6 che rappresenta nuovamente il documento elettronico di minore rilievo.

Concludendo, l'esempio appena esposto, ha permesso di evidenziare ed esaminare ancora una volta come le caratteristiche che contraddistinguono il calcolo del PageRank, quali l'indipendenza che esso nutre nei confronti della generica query o la rapidità di determinazione e successiva stesura della classifica implementata in ordine di importanza decrescente, abbiano consentito a Google il raggiungimento dell'enorme successo e conseguente affermazione di cui esso attualmente gode a livello mondiale.

Capitolo 5

Conclusioni

L'obiettivo finale di questo lavoro risulta individuato da una più completa e fine comprensione dei meccanismi che permettono la determinazione di documentazione elettronica rilevante e conseguente stesura di una classifica implementata in base alla qualità insita nelle pagine web determinate nel processo di calcolo del vettore **r** di PageRank.
Nell'analisi, è stata attribuita considerevole visibilità a tutti gli aspetti riguardanti il processo di information retrieval, sia teorici che pratici, inquadrando nel modo più ampio e preciso possibile tale argomentazione.
In questo senso risulta essere molto descrittiva tutta la parte iniziale inerente al processo di riperimento di informazioni e di come esso possa essere ottimizzato sopratutto in ambito informatico adottando strumenti quali risorse avanzate come operatori booleani, di prossimità o sfruttando le potenzialità fornite dall'uso dei caratteri jolly.
Di notevole rilievo gode anche la successiva descrizione dei principali algoritmi di link analysis, implementata dando particolare risalto ed importanza al soggetto dell'elaborato rappresentato dall'algoritmo PageRank, la cui descrizione dettagliata viene riportata in seguito la stesura di un capitolo di ambito strettamente matematico riguardante il calcolo markoviano.
Nella fase finale, tutte le conoscenze acquisite nello studio preliminare, vengono messe in pratica attraverso l'esaminazione di un completo esempio derivante dalla struttura topologica di un mini web. In esso è descritta completamente sia la procedura affrontata per il calcolo del vettore **r** di PageRank, che i passi necessari volti alla successiva stesura di una classifica computata in base alla chiave di ricerca formulata dal generico utente

del web; le azioni compiute per la determinazione della classifica, sono state affrontate per due keyword differenti comprovando l'indipendenza che caratterizza la computazione del PageRank nei confronti della query inoltrata dal generico web surfer e di come tale qualità rappresenti uno dei vantaggi che consentono a Google di essere considerato uno dei più importanti ed affermati motori di ricerca attualmente disponibili.

Bibliografia

[1] Alicante A. (2008) - *Storia Infirmation Retrieval.* Storia dell'Informatica e del Calcolo Automatico, SICSI VIII CILCO.

[2] Amodio P. - *Catene di Markov.* Ricerca operativa - Met. e mod. per le decisioni (Informatica - Matematica), Dipartimento di Matematica, Università di Bari.

[3] Bedussi F. - *Guida all'uso degli operatori logici.* Disponibile a: `http://www.motoridiricerca.it/operatori.htm` [Ultimo accesso 8 Luglio 2012]

[4] Cambini R. (2003) - *Catene di Markov a stati finiti approccio teorico e computazionale.* Appunti per l'Insegnamento di Teoria delle Decisioni (Modelli Probabilistici), Anno Accademico 2002/2003, Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa.

[5] Danapota - *APPUNTI DI INFORMATION RETRIEVAL: PageRank, HITS, SALSA, Inverted Index, TF-IDF, Web Models.*

[6] De Santis A. - *Motori di ricerca e Web Spamming.* Sicurezza su reti II (A.A. 2007-2008), Università degli studi di Salerno.

[7] Dell'Orto F. - *Ricerche avanzate su Google: Gli Operatori Booleani.* Disponibile a: `http://posizionamento.triplaw.it/news/56/ricerche-avanzate-su-google-gli-operatori-booleani` [Ultimo accesso 10 giugno 2012]

[8] Fagnola F. & Sasso E. (2008) - *CATENE DI MARKOV.*

[9] Garuti A. M. (2006) - *Google, ovvero: come diagonalizzare Internet*; note per il corso di Geometria ed Algebra Lineare.

[10] Google - *Panoramica tecnologica.* Disponibile a: `http://www.google.it/intl/it/about/corporate/company/tech.html` [Ultimo accesso 17 luglio 2012]

[11] Guarnieri M. (2012) - *Elementi di elettrotecnica circuitale. Seconda edizione.* Edizioni Progetto - Padova.

[12] Guirreri S. S. (2007) - *Le catene di Markov come metodologia utilizzata dai motori di ricerca per classificare le pagine web su internet.* Facoltà di Economia, Dipartimento di Scienze Statistiche e Matematiche "S. Vianelli", Dottorato di Ricerca in Statistica e Finanza Quantitativa - XXI Ciclo, Università degli studi di Palermo.

[13] Langville N. A. and Meyer D. C. (2005) - *A Survey of Eigenvector Methods for Web Information Retrieval.* SIAM Rev. 47, pp. 135 - 161.

[14] Maioli C. (2007) - *Introduzione all'Information Retrieval: rappresentazione, memorizzazione, organizzazione e accesso a informazioni.* Università di Bologna.

[15] motoricerca - *Guida al posizionamento dei siti web nei motori di ricerca.* Disponibile a: `http://www.motoricerca.info/articoli/pagerank.phtml` [Ultimo accesso 25 marzo 2012]

[16] Petterle L. (2010-2011) - *Implementazione di Algoritmi di Link Analysis e Campionamento del web per la loro Valutazione.* Tesi di laurea, Dipartimento di Ingegneria dell'Informazione, Università di Padova.

[17] PMI Servizi - *Google, il bollettino novità del mese di gennaio.* Disponibile a: `http://news.pmiservizi.it/news/internet-news/google-novita-gennaio.html` [Ultimo accesso 5 maggio 2012]

[18] Prodi G. (1992) - *METODI MATEMATICI E STATISTICI per le scenze applicate.* McGraw-Hill.

[19] Ridi R. (2012) - *NOZIONI DI INFORMATION RETRIEVAL.*

[20] Salinelli E. & Tomarelli F. (2002) - *Modelli Dinamici Discreti.* Springer.

[21] SEO Guida - *Google PageRank*. Disponibile a: `http://www.seoguida.com/approfondimenti/page-rank.php` [Ultimo accesso 12 maggio 2012]

[22] Valente L. - *Algoritmo di Google, alcuni cambiamenti recenti*. Disponibile a: `http://consulenzaseo.tumblr.com/post/12806648301/algoritmo-di-google-alcuni-cambiamenti-recenti` [Ultimo accesso 5 maggio 2012]

[23] Wikipedia - *Logo Google*. Disponibile a: `http://it.wikipedia.org/wiki/Logo_Google` [Ultimo accesso 21 luglio 2012]

[24] you-can.it - *Come funziona il motore di ricerca*. Disponibile a: `http://www.you-can.it/Motori_di_ricerca/come_funziona_il_motore_di_ricerca.asp` [Ultimo accesso 12 maggio 2012]