Google, ovvero: come diagonalizzare Internet

Marco A. Garuti

6 ottobre 2006

Le pagine web presenti nel database del motore di ricerca Google sono elencate in ordine di importanza. Quando un utente inserisce le parole-chiave per una ricerca, Google seleziona nel suo database le pagine rilevanti[1] e le presenta all'utente in ordine di importanza. Avere a disposizione una "classifica" indipendente dai termini della ricerca, permette a Google di rispondere quasi istantaneamente alla richiesta: la rapidità di risposta è uno dei vantaggi sulla concorrenza che ha consentito a Google di affermarsi come il motore di ricerca attualmente più utilizzato.

La classifica per importanza delle pagine web viene effettuata mediante l'algoritmo PageRank, introdotto dai fondatori di Google Sergey Brin e Larry Page nel 1998[2]. L'algoritmo si basa in buona parte sulla teoria della diagonalizzazione.

I principi sui quali si basa PageRank sono i seguenti:

- Una pagina importante riceve links da pagine importanti.
- Una pagina importante ha pochi links verso altre pagine.

Questi principi vengono formalizzati nella seguente formula: indicando con $r(p)$ il *rango* della pagina web p (cioè la sua importanza relativa) e con $|p|$ il numero di links dalla pagina p verso altre pagine, abbiamo

$$r(p) = \sum_{q \to p} \frac{r(q)}{|q|}. \tag{1}$$

In questa formula, la somma è effettuata su tutte le pagine q che hanno un link verso p. Il contributo di una pagina q è quindi direttamente proporzionale all'importanza (rango) di q ed inversamente proporzionale al numero di links da q verso altre pagine.

Si tratta dunque di una definizione *ricorsiva*: per calcolare il rango di p, dobbiamo conoscere il rango di tutte le pagine q che hanno un link verso p; per calcolare il rango di una di queste pagine q, dobbiamo conoscere il rango delle pagine che hanno un link verso q, e così via all'infinito... Inoltre, se p ha un link verso q e q ha un link verso p, non potremo calcolare $r(p)$ senza conoscere $r(q)$ che non può essere calcolato senza conoscere $r(p)$.

Possiamo riformulare il problema in termini matriciali. Siano $\{p_1, \ldots, p_n\}$ tutte le pagine web di Internet e consideriamo la matrice $A \in M(n \times n, \mathbb{R})$ il cui elemento di posto (i, j) è

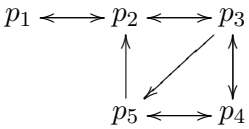
$$a_{i,j} = \begin{cases} \frac{1}{|p_j|} & \text{se esiste un link da } p_j \text{ a } p_i \\ 0 & \text{altrimenti} \end{cases} \tag{2}$$

(la matrice A può essere interpretata in termini probabilistici: $a_{i,j}$ è la probabilità che un internauta che sta visualizzando la pagina p_j clicki sul link alla pagina p_i).

[1] Cioè pagine che contengono le parole-chiave o che hanno un link da o verso pagine che contengono tali parole.

[2] S.Brin, L. Page, R. Motwami, T. Winograd, *The PageRank citation ranking: bringing order to the Web*, Technical Report 1999-0120, Computer Science Department, Stanford University, Stanford, CA, 1999.

Esempio Consideriamo un mini-web di sole 5 pagine:



In questo diagramma, una freccia $p_i \to p_j$ sta ad indicare che la pagina p_i ha un link alla pagina p_j (una doppia freccia $p_i \leftrightarrow p_j$ sta ad indicare che anche p_j ha un link a p_i). La matrice A associata al nostro mini-web è allora

$$A = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{3} & \frac{1}{2} & 0 \end{pmatrix}. \tag{3}$$

La prima colonna ci dice quindi che p_1 ha un solo link (quindi $|p_1| = 1$) verso p_2 (il solo elemento non nullo è sulla seconda riga); la terza colonna ci dice che p_3 ha tre link (quindi $|p_3| = 3$) verso p_2, p_4 e p_5 (gli elementi non nulli sono sulla seconda, quarta e quinta riga).

Tornando alla discussione generale, se indichiamo con $r_i = r(p_i)$ il rango della pagina p_i il *vettore di PageRank* è il vettore colonna

$$\mathbf{r} = \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix}.$$

Il prodotto righe per colonne $A\mathbf{r}$ è il vettore colonna la cui componente i-esima è

$$\sum_{j=1}^{n} a_{i,j} r_j = \sum_{p_j \to p_i} \frac{r_j}{|p_j|} = r_i$$

dove l'ultima uguaglianza è l'equazione (1). In altre parole, l'equazione (1) può essere riscritta come

$$\mathbf{r} = A\mathbf{r}. \tag{4}$$

da cui risulta la

Proposizione 1 *Il vettore di PageRank è un autovettore della matrice A di autovalore 1.*

Quindi, affinché l'algoritmo PageRank funzioni, occorrerà che la matrice A ammetta l'autovalore 1. Se tale autovalore ha moltepicità algebrica 1, l'autovettore $\mathbf{r}$ è determinato a meno di un fattore di proporzionalità. Di solito si normalizza $\mathbf{r}$ mediante la condizione $\sum_{i=1}^{n} r_i = 1$.

Osservazione 1 Sfruttando il fatto che la matrice A è una matrice probabilistica, si può dimostrare che, se λ è un autovalore di A, allora $|\lambda| \leq 1$. Il teorema di Perron-Frobenius[3] dà una condizione necessaria e sufficiente affinché una matrice di questo tipo ammetta l'autovalore 1 con molteplicità 1. La matrice A sicuramente non soddisfa le ipotesi del teorema: una delle condizioni infatti è che la somma degli elementi su una qualunque colonna sia uguale ad 1[4]. Questo si verifica per le colonne corrispondenti a pagine che hanno almeno un link verso un'altra pagina, ma si stima che circa il 25% di tutte le pagine presenti sul web non abbia neanche un link. La soluzione adottata da Google è di *perturbare* la matrice A in modo che soddisfi le ipotesi del teorema di Perron-Frobenius; questo "trucco" ovviamente introduce un elemento arbitrario nel procedimento.

[3]Si veda ad es. O. Axelsson, *Iterative solution methods*, Cambridge University Press, 1994.

[4]Questa condizione si esprime dicendo che A è una matrice stocastica.

In questi appunti, ignoreremo i problemi sollevati nella precedente osservazione, la cui soluzione esula dalle competenze del corso. Supporremo quindi per semplificare che:

1. La matrice A è diagonalizzabile.
2. A ammette l'autovalore 1 con moltepicità algebrica 1.
3. Se $\lambda \neq 1$ è un autovalore di A, allora $|\lambda| < 1$.

Per calcolare il vettore PageRank $\mathbf{r}$ "basta" quindi determinare l'autospazio della matrice A relativo all'autovalore 1. Nell'esempio del mini-web considerato prima, il polinomio caratteristico della matrice data in (3) è

$$t^5 - \frac{13}{12}t^3 - \frac{1}{6}t^2 + \frac{5}{24}t + \frac{1}{24} = (t-1)(t-\frac{1}{2})(t+\frac{1}{2})(t+\frac{3-\sqrt{2}}{6})(t+\frac{3+\sqrt{2}}{6}).$$

La matrice soddisfa dunque le nostre ipotesi e, calcolando, troviamo che un autovettore relativo all'autovalore 1 è per esempio $\mathbf{v}_1 = (2,4,3,2,2)$. Quindi $\mathbf{r}$ è un multiplo di $\mathbf{v}_1$ normalizzato dalla condizione che la somma delle sue coordinate sia 1, cioè

$$\mathbf{r} = (\frac{2}{13}, \frac{4}{13}, \frac{3}{13}, \frac{2}{13}, \frac{2}{13}).$$

Vediamo che la pagina più importante del nostro mini-web è p_2 (che è quella che riceve il maggior numero di links), la seconda classificata è p_3 (che riceve due links, come p_4 e p_5, di cui però uno proviene da p_2, che è più importante) ed ultime a pari merito sono p_4, p_5 e p_1 (che compensa lo svantaggio di ricevere un solo link con il fatto che questo proviene da p_2).

Nel caso di Google, $A \in M(n \times n, \mathbb{R})$ dove $n =$ numero di tutte le pagine di Internet è un numero spaventoso! Gli strumenti di calcolo attuali non consentono neanche di calcolare un autovettore della matrice A. Il vettore PageRank viene allora calcolato mediante un approccio *dinamico*. Si considera il vettore iniziale

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

in cui tutte le pagine del web hanno la stessa importanza (rango). Si stila poi una seconda classifica

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

tenendo conto dei link tra le diverse pagine, del rango iniziale $r_i^{(0)}$ di ciascuna pagina e del numero di links $|p_i|$ secondo i principi enunciati:

$$r_i^{(1)} = \sum_{p_j \to p_i} \frac{r_j^{(0)}}{|p_j|} = \sum_{j=1}^{n} a_{i,j} r_j^{(0)}.$$

In termini matriciali $\mathbf{r}^{(1)} = A\mathbf{r}^{(0)}$. Si ripete poi il procedimento, stilando ogni volta una nuova classifica

$$\mathbf{r}^{(k+1)} = \begin{pmatrix} r_1^{(k+1)} \\ \vdots \\ r_n^{(k+1)} \end{pmatrix}$$

che tenga conto della classifica precedente $\mathbf{r}^{(k)}$ e del numero di links di ciascuna pagina secondo i principi enunciati:

$$r_i^{(k+1)} = \sum_{p_j \to p_i} \frac{r_j^{(k)}}{|p_j|} = \sum_{j=1}^{n} a_{i,j} r_j^{(k)},$$

cioè

$$\mathbf{r}^{(k+1)} = A\mathbf{r}^{(k)}. \tag{5}$$

Il rango $r_i^{(k)}$ della pagina p_i nella classifica parziale $\mathbf{r}^{(k)}$ tiene conto quindi solamente delle pagine dalle quali è possibile raggiungere p_i in non più di k clicks. Poiché ci aspettiamo che pagine molto distanti da p_i contribuiscano poco all'importanza di p_i, è ragionevole pensare che

Proposizione 2 *Per k abbastanza grande, la classifica parziale $\mathbf{r}^{(k)}$ è molto vicina alla classifica reale $\mathbf{r}$.*

Possiamo dimostrare rigorosamente questa proposizione. Osserviamo per cominciare che dalle equazioni

$$\mathbf{r}^{(k+1)} = A\mathbf{r}^{(k)}, \quad \mathbf{r}^{(k)} = A\mathbf{r}^{(k-1)}, \quad \dots \quad , \quad \mathbf{r}^{(1)} = A\mathbf{r}^{(0)}$$

ricaviamo per sostituzioni successive

$$\mathbf{r}^{(k+1)} = A\mathbf{r}^{(k)} = A^2\mathbf{r}^{(k-1)} = \cdots = A^{k+1}\mathbf{r}^{(0)}. \tag{6}$$

Abbiamo supposto che A sia diagonalizzabile: sia $\mathbf{v}_1, \dots, \mathbf{v}_n$ una base di autovettori di A di autovalori rispettivamente $\lambda_1, \dots, \lambda_n$ (ripetuti con molteplicità). Siano ora

$$D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

la matrice diagonale ed H la matrice degli autovettori (le colonne di H sono dunque i vettori $\mathbf{v}_1, \dots, \mathbf{v}_n$). Abbiamo quindi che $A = HDH^{-1}$, da cui segue che per ogni numero intero k

$$A^k = \left(HDH^{-1}\right)\left(HDH^{-1}\right)\dots\left(HDH^{-1}\right) = HD^kH^{-1}.$$

Abbiamo dunque che

$$\mathbf{r}^{(k)} = A^k\mathbf{r}^{(0)} = HD^kH^{-1}\mathbf{r}^{(0)}.$$

Ponendo

$$H^{-1}\mathbf{r}^{(0)} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \tag{7}$$

ricaviamo che per ogni numero intero k

$$\begin{aligned} \mathbf{r}^{(k)} &= HD^kH^{-1}\mathbf{r}^{(0)} \\ &= (\mathbf{v}_1 \ \dots \ \mathbf{v}_n) \begin{pmatrix} \lambda_1^k & & \\ & \ddots & \\ & & \lambda_n^k \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \\ &= \left(\lambda_1^k\mathbf{v}_1 \ \dots \ \lambda_n^k\mathbf{v}_n\right) \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \\ &= b_1\lambda_1^k\mathbf{v}_1 + \cdots + b_n\lambda_n^k\mathbf{v}_n. \end{aligned} \tag{8}$$

In base alle nostre ipotesi, uno solo degli autovalori λ_i è uguale ad 1 e tutti gli altri sono in valore assoluto minori di 1. Riordinandoli, possiamo pensare che $\lambda_1 = 1$ e $|\lambda_i| < 1$ per $i = 2, \ldots, n$. Ricordiamo poi che, per la Proposizione 1, il vettore di PageRank $\mathbf{r}$ è un multiplo di $\mathbf{v}_1$: scegliamo $\mathbf{r} = b_1\mathbf{v}_1$. Non è difficile dimostrare che, se A è stocastica, il coefficiente $b_1 = 1$.

> Per la formula (7), b_1 è il prodotto della prima riga di H^{-1} per il vettore iniziale $\mathbf{r}^{(0)} = (\frac{1}{n}, \ldots, \frac{1}{n})$. Prendendo la trasposta del prodotto $A = HDH^{-1}$ troviamo che $A^t = (H^{-1})^t D H^t$. Quindi $(H^{-1})^t$ è una matrice di autovettori per A^t ed in particolare la prima colonna $\mathbf{w}_1$ di $(H^{-1})^t$, cioè la prima riga di H^{-1}, è un autovettore di A^t di autovalore 1. Ricordiamo che dire che A è stocastica vuol dire che la somma degli elementi sulle sue colonne è uguale ad uno, cioè $(1, \ldots, 1)A = (1, \ldots, 1)$; trasponendo questa espressione abbiamo quindi che $(1, \ldots, 1)$ è un autovettore di autovalore 1 per A^t. Dunque $\mathbf{w}_1 = \alpha(1, \ldots, 1)$, dove la costante α è determinata dalla condizione che il prodotto della prima riga $\mathbf{w}_1$ di H^{-1} per la prima colonna $\mathbf{r}$ di H deve essere uguale ad 1. Ma abbiamo normalizzato $\mathbf{r}$ imponendo che la somma delle sue coordinate sia uguale ad 1; quindi $\alpha = 1$. Ne deduciamo che
>
> $$b_1 = (1, \ldots, 1) \begin{pmatrix} \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{pmatrix} = 1$$

Dalla formula (8) ricaviamo allora che

$$\mathbf{r}^{(k)} = \mathbf{r} + b_2\lambda_2^k\mathbf{v}_2 + \cdots + b_n\lambda_n^k\mathbf{v}_n.$$

Per $i = 2, \ldots, n$, siccome $|\lambda_i| < 1$, per k molto grande λ_i^k è trascurabile. Quindi $\mathbf{r}^{(k)} \approx \mathbf{r}$ per k abbastanza grande.

Osservazione 2 L'ipotesi che A sia diagonalizzabie è superflua: in quanto precede, si può sostituire la matrice diagonale D con una forma di Jordan di A e prendere come H la relativa matrice di autovettori generalizzati.

Osservazione 3 Nella pratica, la classifica delle pagine web utilizzata da Google è data da $\mathbf{r}^{(k)}$ per un opportuno valore di k, anziché da $\mathbf{r}$. Il calcolo effettivo di $\mathbf{r}^{(k)}$ presenta comunque notevoli problemi computazionali. Una scelta oculata della modificazione della matrice A (vedi osservazione 1) consente a Google di determinare una classifica utile dopo un numero di iterazioni (circa $k \approx 100$) abbastanza piccolo per calcolare A^k per moltiplicazione diretta (come già segnalato, nessun calcolatore attuale è lontanamente in grado di determinare le matrici D ed H). I calcolatori di Google impiegano comunque alcune settimane per eseguire questo calcolo. Per tenere sempre aggiornata la classifica, Google ricalcola il vettore Pagerank a ciclo continuo.