Diagonalizzazione e Google

In questa paginetta si ripete, senza molta originalità, la descrizione del processo con cui Google 'indicizza' le pagine web e sceglie quella di maggiore "importanza". Questo è legato ad un problema di diagonalizzazione e al celebre teorema di Perron–Frobenius (una cui dimostrazione si può trovare in un altro di questi fogli). Vediamo di dare brevemente una formulazione del problema.

Siano $P_1, \dots, P_N$ le pagine web in un dato momento (N è un numero molto elevato, ma finito) o, come più spesso accade, tutte le pagine web presenti in un dato momento e contenenti alcune parole chiave. Data una pagina P_j, indichiamo con ℓ_j il numero di links presenti nella pagina e con k_{ij} il numero di links che portano dalla pagina P_j alla pagina P_i. Si ha quindi $0 \leq \ell_j = \sum_{i=1}^{N} k_{ij}$. Possiamo così costruire la matrice $H = (h_{ij})_{1 \leq i,j \leq N}$, detta la *matrice di hyperlink* ove

$$h_{ij} = \begin{cases} \frac{k_{ij}}{\ell_j} & \text{se } \ell_j \neq 0 \\ 0 & \text{altrimenti} \end{cases}$$

e osserviamo che una pagina priva di links ($\ell_j = 0$) dà origine ad una colonna nulla.

Possiamo definire la *rilevanza* della pagina P_i come un numero reale $x_i \geq 0$, legato alla rilevanza delle altre pagine dalla relazione

$$x_i = \sum_{j=1}^{N} h_{ij} x_j, \qquad \text{per } i = 1, \dots, N.$$

Ovvero i links che portano dalla pagina P_j alla pagina P_i trasmettono a quest'ultima la frazione h_{ij} di importanza della pagina P_j. Ciò significa esattamente che il vettore $x = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}$ è un autovettore relativo all'autovalore 1 per la matrice H. Quindi, classificare per importanza tutte le pagine web (o solo quelle che contengono certi termini) corrisponde al problema di trovare un autovettore relativo all'autovalore 1 con entrate tutte positive. Le entrate più grandi corrisponderanno alle pagine più rilevanti.

Nulla ci dice che 1 sia un autovalore o che vi sia un unico autovettore (a meno di fattori di proporzionalità), né che quest'ultimo abbia entrate positive. Un celebre Teorema di Perron e Frobenius garantisce che una matrice reale con tutte le entrate positive ha un solo autovalore positivo con valore assoluto massimo e che il corrispondente spazio di autovettori ha dimensione 1 ed è generato da un vettore con tutte le entrate positive (per una dimostrazione vedi il foglio dedicato). Vediamo quindi come modificare il problema in modo da soddisfare le ipotesi del Teorema.

Un piccolo miglioramento si ha se si interpretano i numeri h_{ij} come la *probabilità che un utente che si trova nella pagina P_j passi alla pagina P_i, seguendo uno dei links*. Da una pagina priva di links si ha uguale probabilità di spostarsi verso qualsiasi pagina, e quindi possiamo considerare la matrice $A = (a_{ij})_{1 \leq i,j \leq N}$ con

$$a_{ij} = \begin{cases} \frac{1}{N} & \text{se } \ell_j = 0 \\ 0 & \text{altrimenti} \end{cases}.$$

La nuova matrice che prendiamo in considerazione è quindi $L = H + A$ che è una matrice stocastica (ovvero ha entrate in $[0, 1]$ e la somma delle entrate di ogni colonna è uguale a 1) e quindi vi è l'autovalore 1 e tutti gli altri autovalori hanno valore assoluto minore o uguale ad 1(†).

(†) Chi non conosca questo fatto, può vederlo così. Se una matrice B ha le entrate $b_{ij} \geq 0$ e la somma delle entrate di ogni colonna è uguale ad 1, la riga $(1, \dots, 1)$ è un autovettore relativo all'autovalore 1 per la trasposta ${}^tB$, ma B e ${}^tB$ hanno lo stesso polinomio caratteristico, quindi 1 è autovalore per B. Dato un vettore $z = \begin{pmatrix} z_1 \\ \vdots \\ z_N \end{pmatrix}$ e, posto $\|z\|_1 = \sum_{i=1}^{N} |z_i|$, si ha $\|z\|_1 = 0$ solo se $z = 0$ e

$$\|Bz\|_1 = \sum_{i=1}^{N} \left| \sum_{j=1}^{N} b_{ij} z_j \right| \leq \sum_{i=1}^{N} \sum_{j=1}^{N} b_{ij} |z_j| = \sum_{j=1}^{N} |z_j| \sum_{i=1}^{N} b_{ij} = \|z\|_1.$$

Quindi, se z è autovettore per B, si ha $Bz = \lambda z$, e quindi $|\lambda| \|z\|_1 = \|Bz\|_1 \leq \|z\|$, con $\|z\| \neq 0$, ovvero $|\lambda| < 1$.

La matrice L descrive quindi il comportamento di un utente che si muove nel web seguendo i links (quando ci sono). Però non è detto che questo sia il comportamento generico, può succedere che un utente si comporti in modo casuale nel passare da una pagine all'altra. In generale il comportamento generico sarà una combinazione dei due comportamenti, ovvero una parte guidata dai links e il resto in modo casuale. Il comportamento casuale à descritto dalla matrice $C = (c_{ij})_{1 \leq i,j \leq N}$ con

$$c_{ij} = \frac{1}{N} \qquad \text{per ogni } i, j,$$

ed anche questa è una matrice stocastica. Il comportamento generale sarà quindi determinato dalla matrice

$$G = \alpha L + (1 - \alpha)C, \qquad \text{con } \alpha \in (0, 1).$$

G è la matrice di Google. I tecnici di Google hanno stimato che un valore attendibile sia $\alpha = 0.85$, ovvero che un utente generico nel muoversi nel web si comporti nell'85% dei casi seguendo i links e nel rimanente 15% in modo casuale. La matrice G che si ottiene ha tutte le entrate reali e positive ed è ancora una matrice stocastica. Quindi (per il Teorema di Perron e Frobenius) esiste un unico autovettore per l'autovalore 1, con le entrate $x_1, \ldots, x_N$ tutte positive, e tale che $x_1 + \cdots + x_N = 1$. Le grandezze delle entrate di questo vettore danno l'importanza delle corrispondenti pagine web.

Quindi, costruita la matrice G, si tratta di trovare un autovettore relativo all'autovalore 1, ovvero di risolvere un sistema lineare (o, almeno, trovarne una soluzione approssimata, visto che, in genere, il rango $N - 1$ è molto alto).