PageRank™, autovalori e autovettori

2 Maggio 2018

Figure 1: Sergey Brin e Larry Page.

Questi appunti informali mostrano come i concetti di autovalore e autovettore intervengano nella definizione del PageRank™ usato da Google™ per quantificare l'importanza di un sito web. La trattazione è volutamente molto semplificata e verranno illustrate soltanto le idee principali. Per comprendere quanto verrà detto saranno sufficienti i concetti di matrice, prodotto fra matrici, trasformazione lineare e sua rappresentazione matriciale, limite di una successione reale e di una successione di matrici reali. Sarebbe utile avere anche qualche

conoscenza di calcolo delle probabilità ma verranno evitati richiami espliciti alla teoria.

Consideriamo l'insieme di tutti i siti web del mondo: $\{w_1, \ldots, w_n\}$. (Alla data in cui stiamo scrivendo n è superiore a 500 milioni).

Immaginiamo, semplificando, che l'insieme di tutti questi siti sia popolato di visitatori e che, a ogni secondo, una percentuale p_j^i dei visitatori del sito w_j decida di passare al sito w_i (naturalmente questo accade, contemporaneamente, per ogni i e ogni j). Il valore p_j^i sarà un numero reale compreso fra 0 e 1 e potrà anche essere visto come una probabilità.

A questo punto abbiamo una matrice quadrata

$$P = \begin{pmatrix} p_1^1 & p_2^1 & \ldots & p_n^1 \\ p_1^2 & p_2^2 & \ldots & p_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ p_1^n & p_2^n & \ldots & p_n^n \end{pmatrix}.$$

Notate che

- $p_j^i \geq 0$ per ogni i e j;
- p_i^i esprime la frazione di visitatori di w_i che decide di rimanere su quel sito;
- $\sum_{i=1}^n p_j^i = 1$ per ogni j.

Nella realtà è ragionevole supporre $p_j^i > 0$ per ogni i e j e dunque faremo questa ipotesi (esiste sempre una probabilità non nulla per ogni scelta del visitatore).

1 A cosa serve la matrice P?

Supponiamo che a un certo istante t la popolazione dei navigatori del web sia suddivisa, sito per sito, in questo modo: $(s_t^1, \ldots, s_t^n)$. È facile rendersi conto che all'istante successivo $t+1$ la popolazione sarà distribuita così:

$$\left(\sum_{i=1}^n p_i^1 \cdot s_t^i, \ldots, \sum_{i=1}^n p_i^n \cdot s_t^i \right)$$

(verificatelo!).

Quindi la funzione che porta la distribuzione dei visitatori all'istante t nella distribuzione dei visitatori all'istante $t+1$ è una trasformazione lineare T :

$\mathbb{R}^n \to \mathbb{R}^n$ e la sua matrice associata è, rispetto alla base canonica, proprio la matrice quadrata P.

Sappiamo che la matrice associata alla trasformazione lineare

$$T^t = \overbrace{T \circ \ldots \circ T}^{t \text{ volte}}$$

è la matrice P^t. Quindi, indicando con $(s_0^1, \ldots, s_0^n)$ la distribuzione sito per sito dei navigatori del web all'istante iniziale 0, la popolazione all'istante t sarà data da

$$P^t \cdot \begin{pmatrix} s_0^1 \\ s_0^2 \\ \vdots \\ s_0^n \end{pmatrix}.$$

Notate, comunque, che non conosciamo i valori s_0^i. Possiamo invece approssimare i valori p_j^i perché siamo in grado di scoprire, navigando sulla rete, se esiste un link che porti dal sito w_j al sito w_i e qual è il numero m_j di link che portano dal sito w_j ad altri siti (considerando anche un link fittizio che porta dal sito w_j in se stesso). Questo ci permette di valutare come $\frac{1}{m_j}$ la probabilità p_j^i del passaggio da w_j a w_i. In parole povere stiamo supponendo che una percentuale $\frac{1}{m_j}$ dei visitatori presenti su w_j a un certo istante t decida di trasferirsi su w_i all'istante successivo $t+1$. Notate che il valore della probabilità di trasferirsi da w_j a w_i dipende non solo dall'esistenza di almeno un link da w_j a w_i, ma anche dal numero totale di link che partono da w_j.

NOTA: Secondo il precedente metodo di calcolo p_j^i risulta nullo qualora non esistano link da w_j a w_i. Per prendere in considerazione anche la possibilità che un visitatore di w_j decida di passare a w_i scegliendo questo sito a caso (quindi senza uso di link) possiamo scegliere un valore d vicino a 1 e quantificare empiricamente come $\frac{1-d}{n}$ (valore positivo vicino a zero!) la probabilità del trasferimento casuale da w_j a w_i. Dobbiamo allora anche cambiare la valutazione $\frac{1}{m_j}$ per p_j^i in $\frac{d}{m_j} + \frac{1-d}{n}$. In pratica al tempo t c'è una parte dei visitatori di w_j (data da $d \cdot s_t^j$) che decide di spostarsi usando un link e una parte (data da $(1-d) \cdot s_t^j$) che decide di spostarsi su un sito scelto a caso. Comunque, il fatto fondamentale è osservare che i valori p_j^i sono approssimabili consultando i siti presenti sulla rete (tramite del software appositamente predisposto).

Cosa accade alla distribuzione dei visitatori sulla rete quando il tempo t viene

fatto tendere all'infinito? Per scoprirlo dobbiamo fare qualche osservazione matematica sulle potenze P^t della matrice P.

2 Le proprietà della matrice P^t

Per studiare le proprietà della matrice P^t indichiamo con $A = (a^i_j)$ la matrice P^t e con $B = (b^i_j)$ la matrice P^{t+1}. Indichiamo con $\gamma_{min} > 0$ il minimo di tutti termini di P.

Elenchiamo ora alcune proprietà interessanti:

1. *Per qualunque tempo t tutti i termini di P^t sono positivi.* Dimostriamolo per induzione su t. L'affermazione è vera per $t = 1$ perché è vera per P. Se è vera per t è vera anche per $t + 1$, infatti $b^i_j = \sum_{h=1}^n p^i_h \cdot a^h_j \geq p^i_1 \cdot a^1_j > 0$.

2. *La somma dei termini in ciascuna colonna di P^t è uguale a* 1. Dimostriamolo per induzione su t. L'affermazione è vera per $t = 1$ perché è vera per P. Se è vera per t è vera anche per $t + 1$, infatti risulta

$$\sum_{i=1}^n b^i_j = \sum_{i=1}^n \left(\sum_{h=1}^n p^i_h \cdot a^h_j \right) = \sum_{h=1}^n \left(\sum_{i=1}^n p^i_h \cdot a^h_j \right) =$$

$$\sum_{h=1}^n \left(\sum_{i=1}^n p^i_h \right) \cdot a^h_j = \sum_{h=1}^n a^h_j = 1.$$

3. *Per qualunque indice i si ha che*

$$(*) \qquad \max_r a^i_r - \max_r b^i_r \geq (\max_r a^i_r - \min_r a^i_r) \cdot \gamma_{min}.$$

Per dimostrarlo procediamo come segue. Fissiamo un indice i a piacere e scegliamo un indice r_i tale che $a^i_{r_i} = \max_r a^i_r$. Osserviamo che

$$a^i_{r_i} - b^i_s = a^i_{r_i} \cdot \sum_{h=1}^n p^h_s - \sum_{h=1}^n a^i_h \cdot p^h_s =$$

$$\sum_{h=1}^n a^i_{r_i} \cdot p^h_s - \sum_{h=1}^n a^i_h \cdot p^h_s =$$

$$\sum_{h=1}^n (a^i_{r_i} - a^i_h) \cdot p^h_s \geq (a^i_{r_i} - a^i_k) \cdot \gamma_{min}$$

per qualunque s e k. (Attenzione!: qui stiamo sfruttando il fatto che $a^i_{r_i} - a^i_h \geq 0$ per qualunque h.) Ne consegue che $\max_r a^i_r - b^i_s \geq (\max_r a^i_r - a^i_k) \cdot \gamma_{min}$ per qualunque s e k. Se prendiamo s e k in modo tale che $b^i_s = \max_r b^i_r$ e $a^i_k = \min_r a^i_r$ otteniamo che $\max_r a^i_r - \max_r b^i_r \geq (\max_r a^i_r - \min_r a^i_r) \cdot \gamma_{min}$.

4. *Per qualunque indice i si ha che*

$$(**) \qquad \min_r b^i_r - \min_r a^i_r \geq (\max_r a^i_r - \min_r a^i_r) \cdot \gamma_{min}.$$

Per dimostrarlo procediamo come segue. Fissiamo un indice i a piacere e scegliamo un indice r_i tale che $a^i_{r_i} = \min_r a^i_r$. Osserviamo che

$$b^i_s - a^i_{r_i} = \sum_{h=1}^{n} a^i_h \cdot p^h_s - a^i_{r_i} \cdot \sum_{h=1}^{n} p^h_s =$$
$$\sum_{h=1}^{n} a^i_h \cdot p^h_s - \sum_{h=1}^{n} a^i_{r_i} \cdot p^h_s =$$
$$\sum_{h=1}^{n} (a^i_h - a^i_{r_i}) \cdot p^h_s \geq (a^i_k - a^i_{r_i}) \cdot \gamma_{min}$$

per qualunque s e k. (Attenzione!: qui stiamo sfruttando il fatto che $a^i_h - a^i_{r_i} \geq 0$ per qualunque h.) Ne consegue che $b^i_s - \min_r a^i_r \geq (a^i_k - \min_r a^i_r) \cdot \gamma_{min}$ per qualunque s e k. Se prendiamo s e k in modo tale che $b^i_s = \min_r b^i_r$ e $a^i_k = \max_r a^i_r$ otteniamo che

$$\min_r b^i_r - \min_r a^i_r \geq (\max_r a^i_r - \min_r a^i_r) \cdot \gamma_{min}.$$

Possiamo ora dimostrare il seguente teorema:

Teorema 2.1. *Esiste $L = \lim_{t\to\infty} P^t$. Le colonne di L sono tutte uguali fra loro, i termini di L sono tutti positivi e la somma dei termini in ciascuna colonna di L è uguale a* 1.

Dimostrazione. Le proprietà 3 e 4 viste prima implicano che per ogni i

$$\min_r a^i_r \leq \min_r b^i_r \leq \max_r b^i_r \leq \max_r a^i_r.$$

Questo significa che il minimo di ogni riga di P^t non diminuisce al crescere del tempo t e che il massimo di ogni riga di P^t non aumenta al crescere del tempo t. Quindi esistono $\ell'_i = \lim_{t\to\infty} \min_r a^i_r$ (coincidente con $\lim_{t\to\infty} \min_r b^i_r$)

e $\ell''_i = \lim_{t\to\infty} \max_r a^i_r$, e sono entrambi positivi (visto che tutti i termini della matrice P sono positivi). Inoltre $\lim_{t\to\infty} \max_r a^i_r \geq \lim_{t\to\infty} \min_r a^i_r$.

Passando al limite per t che tende all'infinito nella disuguaglianza $(**)$, si ha

$$0 = \ell'_i - \ell'_i \geq (\lim_{t\to\infty} \max_r a^i_r - \lim_{t\to\infty} \min_r a^i_r) \cdot \gamma_{min} \geq 0$$

e dunque $(\lim_{t\to\infty} \max_r a^i_r - \lim_{t\to\infty} \min_r a^i_r) \cdot \gamma_{min} = 0$.

Dato che $\gamma_{min} > 0$, se ne deduce che $\lim_{t\to\infty} \max_r a^i_r = \lim_{t\to\infty} \min_r a^i_r$.

Ciò significa che tutti i termini della i-esima riga della matrice P^t tendono a un unico valore $c^i = \ell'_i = \ell''_i > 0$ quando t tende all'infinito. Dunque esiste $L = \lim_{t\to\infty} P^t$, le colonne di L sono tutte uguali fra loro e i loro termini sono tutti positivi. Per la proprietà 2 la somma dei termini in ciascuna colonna di L è uguale a 1 (visto che questa proprietà vale per la matrice P^t e si conserva passando al limite per t che tende all'infinito). □

Il teorema precedente dimostra che, partendo da una distribuzione iniziale $(s^1_0, \ldots, s^n_0)$ dei visitatori sui siti della rete, esiste un'unica distribuzione limite a cui si tende al passare del tempo. Questa distribuzione limite è data da

$$\lim_{t\to\infty} P^t \cdot \begin{pmatrix} s^1_0 \\ \vdots \\ s^n_0 \end{pmatrix} = L \cdot \begin{pmatrix} s^1_0 \\ \vdots \\ s^n_0 \end{pmatrix} = \begin{pmatrix} c^1 & \ldots & c^1 \\ \vdots & \ddots & \vdots \\ c^n & \ldots & c^n \end{pmatrix} \cdot \begin{pmatrix} s^1_0 \\ \vdots \\ s^n_0 \end{pmatrix} = \begin{pmatrix} c^1 \cdot (s^1_0 + \ldots + s^n_0) \\ \vdots \\ c^n \cdot (s^1_0 + \ldots + s^n_0) \end{pmatrix}$$

dove $c^1 + \ldots + c^n = 1$ e tutti i c^i sono positivi. Si noti che i valori $c^1, \ldots, c^n$ NON dipendono dalla distribuzione iniziale $(s^1_0, \ldots, s^n_0)$.

Indicando con S il numero totale di navigatori sul web si ha ovviamente che $S = s^1_0 + \ldots + s^n_0$ e dunque

$$\lim_{t\to\infty} P^t \cdot \begin{pmatrix} s^1_0 \\ \vdots \\ s^n_0 \end{pmatrix} = S \cdot \begin{pmatrix} c^1 \\ \vdots \\ c^n \end{pmatrix}.$$

Dividendo i due termini dell'ultima uguaglianza per S si ottiene che

$$\lim_{t\to\infty} P^t \cdot \begin{pmatrix} s^1_0/S \\ \vdots \\ s^n_0/S \end{pmatrix} = \begin{pmatrix} c^1 \\ \vdots \\ c^n \end{pmatrix}.$$

Ciascun valore c^i rappresenta (sostanzialmente) il PageRank™ del sito w_i: indica la percentuale della popolazione della rete che si trova, dopo un tempo

infinito, nel sito considerato. È chiaro che un sito dove si trovano molti visitatori è, in linea di principio, un sito ritenuto importante dalla comunità della rete, ed è proprio questo fattore a essere misurato dal PageRank™.

NOTA IMPORTANTE: osservate che il Teorema 2.1 non vale più se togliamo l'ipotesi che tutti i p_j^i siano strettamente positivi. Infatti se prendiamo

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

vediamo che **non** esiste $\lim_{t\to\infty} P^t$.

3 Alcune osservazioni finali

Nel seguito indicheremo il vettore $\begin{pmatrix} c^1 \\ \vdots \\ c^n \end{pmatrix}$ col simbolo (c). Sappiamo già che $c^1 + \ldots + c^n = 1$ (perché la somma dei termini in ciascuna colonna di L fa 1).

Risulta dunque che

$$L \cdot (c) = \begin{pmatrix} c^1 \cdot (c^1 + \ldots + c^n) \\ \vdots \\ c^n \cdot (c^1 + \ldots + c^n) \end{pmatrix} = (c).$$

Quando si ha che $M \cdot (x) = \lambda \cdot (x)$ per una matrice $M \in \mathcal{M}_n(\mathbb{R})$ e un vettore non nullo $(x) \in \mathbb{R}^n$ si dice che (x) è un autovettore di M associato all'autovalore λ. Quindi, nel nostro caso si ha che (c) è un autovettore per L associato all'autovalore 1 e che la somma delle sue componenti fa 1.

Osserviamo anche che $P \cdot L = P \cdot \lim_{t\to\infty} P^t = \lim_{t\to\infty} P^{t+1} = \lim_{t\to\infty} P^t = L$ e dunque $P \cdot L = L$. Da questo e da quanto detto sopra segue che

$$P \cdot (c) = P \cdot L \cdot (c) = L \cdot (c) = (c).$$

Quindi si ha che (c) è un autovettore anche per la matrice P (associato all'autovalore 1).

D'altra parte, se $(\bar{c}) = (\bar{c}^1, \ldots, \bar{c}^n)$ è un autovettore di L associato all'autovalore 1 e risulta anche $\bar{c}^1 + \ldots + \bar{c}^n = 1$, allora si ha che

$$(\bar{c}) = L \cdot (\bar{c}) = \begin{pmatrix} c^1 \cdot (\bar{c}^1 + \ldots + \bar{c}^n) \\ \vdots \\ c^n \cdot (\bar{c}^1 + \ldots + \bar{c}^n) \end{pmatrix} = (c).$$

Dunque l'unico autovettore $(\bar{c})$ di L associato all'autovalore 1 per il quale $\bar{c}^1 + \ldots + \bar{c}^n = 1$ è il vettore (c).

Se invece $(\bar{c})$ è un autovettore per la matrice P associato all'autovalore 1 si ha che

$$L \cdot (\bar{c}) = (\lim_{t\to\infty} P^t) \cdot (\bar{c}) = \lim_{t\to\infty} \left(P^t \cdot (\bar{c})\right) = (\bar{c}).$$

Quindi si ha che $(\bar{c})$ è anche un autovettore per L associato all'autovalore 1.

Perciò se $(\bar{c}) = (\bar{c}^1, \ldots, \bar{c}^n)$ è un autovettore di P associato all'autovalore 1 e risulta anche $\bar{c}^1 + \ldots + \bar{c}^n = 1$, allora si ha che $(\bar{c}) = (c)$.

Dunque l'unico autovettore $(\bar{c})$ di P associato all'autovalore 1 per il quale $\bar{c}^1 + \ldots + \bar{c}^n = 1$ è il vettore (c).

Tutto questo ci permette di dare una definizione alternativa di PageRank™: l'n-upla dei PageRank™ per i nostri n siti è data dall'unico autovettore di L (o, equivalentemente, di P) che sia associato all'autovalore 1 e i cui termini abbiano somma uguale a 1.