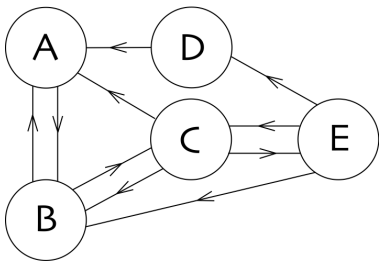
Come funziona Google: catene di Markov e autovalori

Christiane Rousseau

08 febbraio 2012



Fin dall'inizio, Google è diventato "il" motore di ricerca. Ciò deriva dalla supremazia del suo algoritmo di ranking: l'*algoritmo PageRank*. Infatti, con l'enorme quantità di pagine sul World-Wide-Web, molti ricercatori si ritrovano con migliaia o milioni di risultati. Se essi non sono propriamente ordinati, la ricerca può risultare di poco aiuto, dal momento che nessuno riesce ad esplorare milioni di voci.

Come funziona l'algoritmo PageRank?

Lo spiegheremo, ma prima facciamo una ricerca su Google. Il 4 giugno 2010, si ottenevano 16.000.000 di risultati per *Klein project*, nonostante il progetto fosse solamente all'inizio. In quella precisa data, la prima voce era

http://www.mathunion.org/icmi/other-activities/klein-project/introduction/

invece di

http://www.kleinproject.org/

Il primo indirizzo è l'url di una *pagina* che si trova sul sito dell'International Mathematical Union: http://www.mathunion.org. Poiché l'International Mathematical Union è un ente importante, il suo sito ufficiale è il primo disponibile quando si cerca "International Mathematical Union". Inoltre, esso trasmette parte della sua importanza a tutte le sue pagine, una delle quali è

http://www.mathunion.org/icmi/other-activities/klein-project/introduction/

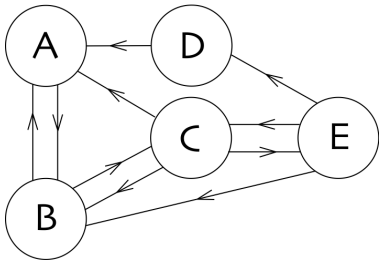
Entro pochi mesi o anni a partire da oggi, ci aspettiamo che la pagina

http://www.kleinproject.org/

appaia per prima in una ricerca per *Klein project*.

Per spiegare l'algoritmo, consideriamo un grafo orientato come modello del web. I vertici sono le *pagine* e i lati orientati sono i *link* tra le pagine. Come abbiamo appena spiegato, ogni pagina corrisponde a un diverso indirizzo url. Quindi, un sito può contenere molte pagine. Il modello non fa differenza tra le singole pagine di un sito e la sua home page. Tuttavia, più probabilmente, l'algoritmo classificherà con grado maggiore la prima pagina di un sito importante.

Un semplice esempio



Consideriamo il semplice web qui a sinistra con cinque pagine chiamate A, B, C, D ed E. Questo web ha pochi link. Se siamo sulla pagina A, allora esiste solo un link alla pagina B, mentre, se siamo sulla pagina C, troviamo tre link e possiamo scegliere se spostarci sulla pagina A, B oppure E. Si osservi che esiste almeno un link a partire da ogni pagina.

Facciamo un gioco che risulta semplicemente un *cammino casuale* sul grafo orientato. Partendo da una pagina, ad ogni passo scegliamo a caso un link e seguiamolo. Per esempio, nel nostro caso, se partiamo sulla pagina B, possiamo andare in A o in C con probabilità $\frac{1}{2}$ per ogni caso mentre, se partiamo da D dobbiamo andare necessariamente in A con probabilità 1. Iteriamo il gioco.

Dove ci troveremo dopo n passi?

Per automatizzare il processo, compattiamo il web nella seguente matrice P, dove ogni colonna rappresenta la pagina di partenza ed ogni riga la pagina di arrivo.

$$P = \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 1 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & 0 \end{pmatrix} \end{array} \begin{array}{c} A \\ B \\ C \\ D \\ E \end{array}$$

Osserviamo che la somma degli elementi di ogni colonna di P è uguale a 1 e che tutti gli elementi sono maggiori o uguali a zero. Le matrici che presentano queste due proprietà sono molto speciali: ognuna di esse è matrice di una *catena di Markov*, anche detta *matrice di transizione markoviana*. Essa ha sempre 1 come autovalore ed esiste un autovettore di autovalore 1 le cui componenti sono tutte minori o uguali a 1 e maggiori o uguali a 0, con somma 1. Tuttavia, prima di richiamare le definizioni di autovalore e autovettore, esploriamo i vantaggi della rappresentazione matriciale del grafo, modello del web.

Consideriamo una variabile casuale X_n che assuma valori nell'insieme delle pagine $\{A, B, C, D, E\}$, che contiene N pagine (qui $N = 5$). X_n rappresenta la pagina in cui ci troviamo dopo n passi del cammino casuale. Quindi, se chiamiamo p_{ij} l'elemento sull'i-esima riga e sulla j-esima colonna della matrice P, allora p_{ij} è la probabilità condizionata di trovarsi sull'i-esima pagina al passo $n+1$, dato che si era sulla j-esima pagina al passo n:

$$p_{ij} = \text{Prob}(X_{n+1} = i \mid X_n = j).$$

Si noti che questa probabilità è indipendente da n! Si dice che una catena di Markov *non ha memoria del passato*. Non è difficile immaginare che le probabilità che si ottengono dopo due passi possano essere contenute nella matrice P^2.

Dimostriamolo (il lettore può saltare la dimostrazione se preferisce). Il teorema delle probabilità totali afferma che

$$\text{Prob}(X_{n+2} = i \mid X_n = j) = \sum_{k=1}^{N} \text{Prob}(X_{n+2} = i \text{ et } X_{n+1} = k \mid X_n = j).$$

Per definizione di probabilità condizionata

$$\text{Prob}(X_{n+2} = i \mid X_n = j) = \sum_{k=1}^{N} \frac{\text{Prob}(X_{n+2} = i \text{ et } X_{n+1} = k \text{ et } X_n = j)}{\text{Prob}(X_n = j)}.$$

Usiamo un trucco familiare: moltiplicare e dividere per una stessa quantità

$$\text{Prob}(X_{n+2} = i \mid X_n = j) =$$
$$\sum_{k=1}^{N} \frac{\text{Prob}(X_{n+2} = i \text{ et } X_{n+1} = k \text{ et } X_n = j)}{\text{Prob}(X_{n+1} = k \text{ et } X_n = j)} \frac{\text{Prob}(X_{n+1} = k \text{ et } X_n = j)}{\text{Prob}(X_n = j)}.$$

Il primo quoziente è uguale a

$$\text{Prob}(X_{n+2} = i \mid X_{n+1} = k \text{ et } X_n = j) = \text{Prob}(X_{n+2} = i \mid X_{n+1} = k),$$

in quanto in una catena di Markov *non vi è memoria di ciò che è avvenuto prima del passo immediatamente precedente*. Perciò,

$$\begin{aligned}
&\text{Prob}(X_{n+2} = i \mid X_n = j) = \\
&= \sum_{k=1}^{N} \text{Prob}(X_{n+2} = i \mid X_{n+1} = k)\text{Prob}(X_{n+1} = k \mid X_n = j) \\
&= \sum_{k=1}^{N} p_{ik}p_{kj} \\
&= (P^2)_{ij}.
\end{aligned}$$

Nel nostro esempio,

$$P^2 = \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & 0 & \frac{11}{18} \\ 0 & \frac{2}{3} & \frac{4}{9} & 1 & \frac{1}{9} \\ \frac{1}{2} & 0 & \frac{5}{18} & 0 & \frac{1}{6} \\ 0 & 0 & \frac{1}{9} & 0 & 0 \\ 0 & \frac{1}{6} & 0 & 0 & \frac{1}{9} \end{pmatrix} \end{array} \begin{array}{c} A \\ B \\ C \\ D \\ E \end{array}$$

Iterando quest'idea, è chiaro che l'elemento $(P^m)_{ij}$ della matrice P^m descrive la probabilità $\text{Prob}(X_{n+m} = i \mid X_n = j)$. Per esempio,

$$P^{32} = \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0.293 & 0.293 & 0.293 & 0.293 & 0.293 \\ 0.390 & 0.390 & 0.390 & 0.390 & 0.390 \\ 0.220 & 0.220 & 0.220 & 0.220 & 0.220 \\ 0.024 & 0.024 & 0.024 & 0.024 & 0.024 \\ 0.073 & 0.073 & 0.073 & 0.073 & 0.073 \end{pmatrix} \end{array} \begin{array}{c} A \\ B \\ C \\ D \\ E \end{array}$$

Tutte le colonne di P^{32} sono identiche se scegliamo una precisione di 3 cifre decimali e lo stesso vale per le colonne di P^n quando $n > 32$. Se scegliamo una precisione più alta, troviamo comunque una stabilizzazione, ma per n più grande di 32. Perciò, dopo n passi, con n sufficientemente grande, la probabilità di essere su una determinata pagina non dipende da quella in cui siamo partiti!

Inoltre, consideriamo il vettore

$$\pi^t = (0.293, 0.390, 0.220, 0.024, 0.073)$$

(π è un vettore colonna e il suo trasposto π^t è un vettore riga). É semplice verificare che $P\pi = \pi$. Se consideriamo la i-esima coordinata del vettore π^t come la probabilità di essere sulla pagina i-esima in un dato istante n e, quindi, π^t come la distribuzione di probabilità delle pagine al tempo n, allora esso è anche la distribuzione di probabilità al tempo $n + 1$. Per questa ragione, il vettore π è detto *distribuzione stazionaria*. Tale distribuzione stazionaria permette di ordinare le pagine. Nel nostro esempio, ordiamo le pagine come B, A, C, E, D e dichiariamo B la pagina più importante.

Il caso generale

Il caso generale può essere trattato esattamente come il nostro esempio. Rappresentiamo il web come un grafo orientato in cui gli N vertici rappresentano le N pagine del web e il lati orientati rappresentano i link tra le pagine. Compattiamo il grafo in una matrice $N \times N$, P, in cui la j-esima colonna rappresenta la j-esima pagina di partenza e l'i-esima riga, la i-esima pagina di arrivo. Nel nostro esempio, noi abbiamo trovato un vettore π soddisfacente la condizione $P\pi = \pi$. Tale vettore è un autovettore di autovalore 1. Richiamiamo la definizione di autovalore e autovettore.

Definizione. Sia P una matrice $N \times N$. Un numero $\lambda \in \mathbb{C}$ è un *autovalore* di P se esiste un vettore non nullo $X \in \mathbb{C}^N$ tale che $PX = \lambda X$. Ogni vettore X di questo tipo è detto *autovettore* di P.

Richiamiamo anche il metodo per trovare autovalori e autovettori.

Proposizione. Sia P una matrice $N \times N$. Gli autovalori di P sono le radici del polinomio caratteristico $\det(\lambda I - P) = 0$, dove I è la matrice identità

$N \times N$. Gli autovettori di un autovalore λ sono le soluzioni non nulle del sistema lineare omogeneo $(\lambda I - P)X = 0$.

Il seguente teorema di Frobenius garantisce che per la matrice associata ad un grafo che modellizza il web si può sempre trovare una soluzione stazionaria.

Teorema. (Teorema di Frobenius) Consideriamo una matrice $N \times N$, $P = (p_{ij})$, matrice di transizione di un processo di Markov (nello specifico, $p_{ij} \in [0,1]$ per ogni i, j, e la somma degli elementi su ogni colonna è uguale a 1, ossia $\sum_{i=1}^{N} p_{ij} = 1$). Allora

1. $\lambda = 1$ è un autovalore di P.
2. Ogni autovalore λ di P soddisfa $|\lambda| \leq 1$.
3. Esiste un autovettore π di autovalore 1, le cui componenti sono tutte maggiori o uguali a zero. Senza perdere di generalità possiamo supporre che la somma delle componenti di π sia 1.

É giunto il momento di constatare la potenza di questo teorema. A tale scopo, aggiungeremo per semplicità l'ipotesi che la matrice P abbia una base di autovettori $B = \{v_1, \ldots, v_N\}$ e supponiamo che v_1 sia il vettore π del teorema di Frobenius. Per ogni v_i esiste λ_i tale che $Pv_i = \lambda_i v_i$. Consideriamo un qualsiasi vettore non nullo X tale che $X^t = (x_1, \ldots, x_N)$, dove $x_i \in [0,1]$ e $\sum_{i=1}^{N} x_i = 1$. Decomponiamo X nella base B:

$$X = \sum_{i=1}^{N} a_i v_i.$$

Una dimostrazione tecnica, che qui saltiamo, permette di provare che $a_1 = 1$. Ora, calcoliamo PX:

$$PX = \sum_{i=1}^{N} a_i P v_i = \sum_{i=1}^{N} a_i \lambda_i v_i,$$

dato che v_i è un autovettore di autovalore λ_i. E se iteriamo, otteniamo

$$P^n X = \sum_{i=1}^{N} a_i \lambda_i^n v_i.$$

Se tutti i λ_i per $i > 1$ verificano $|\lambda_i| < 1$, allora

$$\lim_{n\to\infty} P^n X = a_1 v_1 = \pi.$$

Questo è esattamente quanto è avvenuto nel nostro esempio!

Si osservi comunque che il teorema non garantisce che ogni matrice P, soddisfacente le ipotesi del teorema, abbia questa proprietà. Descriviamo le possibili patologie e i rimedi.

Possibili patologie

- L'autovalore 1 può essere una radice multipla del polinomio caratteristico $\det(\lambda I - P) = 0$.
- La matrice P può avere autovalori λ diversi da 1, con modulo uguale a 1.

Cosa facciamo in questi casi?

Una matrice di transizione di un processo di Markov che non presenta patologie si dice *matrice di transizione markoviana regolare*, ovvero

Definizione. Una matrice di transizione markoviana è *regolare* se

- L'autovalore 1 è una radice semplice del polinomio caratteristico $\det(\lambda I - P) = 0$.
- Tutti gli autovalori λ di P diversi da 1 hanno modulo minore di 1.

Si osservi che la maggior parte delle matrici P sono regolari. Quindi, se la matrice di transizione markoviana associata ad un web non è regolare, allora la strategia è quella di deformarla leggermente in una matrice regolare.

Rimedio. Consideriamo la matrice $N \times N$, $Q = (q_{ij})$, con $q_{ij} = \frac{1}{N}$ per ogni i, j. Sostituiamo la matrice P associata al web con la matrice

$$P_\beta = (1 - \beta)P + \beta Q, \tag{1}$$

per valori piccoli di $\beta \in [0, 1]$. (Google ha usato il valore $\beta = 0.15$.) Si noti che la matrice P_β ha ancora elementi non negativi e che la somma degli elementi di ogni colonna è ancora uguale a 1, quindi essa è ancora una matrice di transizione markoviana. Il seguente teorema garantisce l'esistenza di un

piccolo valore di β per cui si pone rimedio alla patologia.

Teorema. Data una qualsiasi matrice di transizione markoviana P, esiste sempre un β positivo, piccolo a piacere, tale che la matrice P_β sia regolare. Sia π l'autovettore di 1 per la matrice P_β, normalizzato in modo che la somma delle sue componenti sia 1. Per la matrice P_β, dato un qualsiasi vettore X, tale che $X^t = (p_1, \ldots, p_N)$ con $p_i \in [0,1]$ e $\sum_{i=1}^{N} p_i = 1$, allora

$$\lim_{n\to\infty} (P_\beta)^n X = \pi. \tag{2}$$

Collegamento con il teorema del punto fisso di Banach

Una delle "vignette" successive tratta del teorema del punto fisso di Banach. Il teorema sopra può essere visto come una sua particolare applicazione. Il lettore può saltare questa sezione se non ha ancora letto l'altra "vignette". Infatti,

Teorema. Sia P una matrice di transizione markoviana regolare. Consideriamo $S = \{X | X^t = (p_1, \ldots, p_N), p_i \in [0,1], \sum_{i=1}^{N} p_i = 1\}$, con una distanza $d(X,Y)$ tra i punti adeguata (tale distanza dipende da P). Su S, consideriamo l'operatore lineare $L : S \to S$ definito da $L(X) = PX$. L'operatore L è una contrazione su S, ossia esiste $c \in [0,1[$ tale che per ogni $X, Y \in S$,

$$d(L(X), L(Y)) \leq c\, d(X,Y).$$

Allora, esiste un unico vettore $\pi \in S$ tale che $L(\pi) = \pi$.

Inoltre, dato un qualsiasi $X_0 \in S$, possiamo definire per induzione la successione $\{X_n\}$, dove $X_{n+1} = L(X_n)$. Allora, $\lim_{n\to\infty} X_n = \pi$.

Definizione di distanza d**.** La definizione di distanza d è sottile e può essere saltata. La includiamo per morivi di completezza per il lettore che vuole inoltrarsi nei dettagli. Ci limitiamo al caso in cui la matrice P è diagonalizzabile. Sia $B = \{v_1 = \pi, v_2, \ldots, v_N\}$ una base di autovettori. I vettori $X, Y \in S$ possono essere scritti nella base B:

$$X = \sum_{i=1}^{N} a_i v_i, \quad Y = \sum_{i=1}^{N} b_i v_i,$$

dove $a_i, b_i \in \mathbb{C}$. Allora definiamo

$$d(X,Y) = \sum_{i=1}^{N} |a_i - b_i|.$$

Dotato di questa distanza, S risulta uno spazio metrico completo, ossia ogni successione di Cauchy è convergente.

Tale teorema non solo garantisce l'esistenza di π, ma fornisce anche un metodo per costruirlo, come limite di una successione $\{X_n\}$. Abbiamo visto una dimostrazione di questa convergenza nel nostro esempio. Infatti, la j-esima colonna di P^n è il vettore $P^n e_j$, dove $e_j^t = (0, \ldots, 0, 1, 0, \ldots, 0)$ è il j-esimo vettore della base canonica. Naturalmente, nel nostro esempio, avremmo anche potuto trovare direttamente il vettore π risolvendo il sistema $(I - P)X = 0$ con matrice

$$I - P = \begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{3} & -1 & 0 \\ -1 & 1 & -\frac{1}{3} & 0 & -\frac{1}{3} \\ 0 & -\frac{1}{2} & 1 & 0 & -\frac{1}{3} \\ 0 & 0 & 0 & 1 & -\frac{1}{3} \\ 0 & 0 & -\frac{1}{3} & 0 & 1 \end{pmatrix}$$

Avremmo trovato che tutte le soluzioni sono della forma $(4s, \frac{16}{3}s, 3s, \frac{1}{3}s, s)^t$ per $s \in \mathbb{R}$. La soluzione la cui somma delle componenti è 1 è quindi π, dove

$$\pi^t = \Big(\frac{12}{41}, \frac{16}{41}, \frac{9}{41}, \frac{1}{41}, \frac{3}{41}\Big).$$

Calcolo pratico di una distribuzione stazionaria

Abbiamo individuato la semplice idea che sta alla base dell'algoritmo. Comunque, trovare la distribuzione stazionaria π, ossia un autovettore di autovalore 1 per la matrice P_β in (1), non è un compito banale quando la matrice ha miliardi di righe e colonne: sia il tempo computazionale, sia lo spazio in memoria per ottenerla rappresentano una vera e propria sfida. L'usuale metodo di eliminazione di Gauss risulta inutile in questo caso, sia a causa della dimensione del calcolo, sia perché richiede di dividere per piccoli coefficienti. Un algoritmo più efficace fa uso della proprietà (2) (si veda [2]). Questo fatto ci ricollega alla successiva "vignette" sul teorema del punto fisso di Banach che metterà in luce come la dimostrazione del teorema del punto fisso di Banach fornisca un algoritmo per costruire il punto fisso.

Infatti, iniziamo con X_0 tale che

$$X_0^t = \Big(\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}\Big),$$

e dobbiamo calcolare il $\lim_{n\to\infty} (P_\beta)^n X_0$. Solitamente, $(P_\beta)^n X_0$ per un n compreso tra 50 e 100 dà una buona approssimazione di π. Per induzione, calcoliamo $X_{n+1} = P_\beta X_n$. Anche questi calcoli risultano abbastanza lunghi. Infatti, per come è stata costruita, la matrice P_β in (1) non ha elementi nulli. D'altro canto, la maggior parte degli elementi di P è nulla. Quindi, dobbiamo decomporre il calcolo per approfittare di questo fatto, ossia

$$X_{n+1} = (1-\beta)PX_n + \beta QX_n.$$

A causa della forma speciale di Q, è semplice verificare che se X è un vettore in cui la somma degli elementi è 1, allora $QX = X_0$. Perciò, è sufficiente calcolare la successione

$$X_{n+1} = (1-\beta)PX_n + \beta X_0.$$

Conclusioni

Abbiamo presentato la parte pubblica dell'algoritmo PageRank di Google. Il lettore può già cimentarsi con semplici web e trovare i trucchi per migliorare il ranking della propria pagina personale aggiungendo link interni ed esterni in modo ottimale. Alcune parti private e più sofisticate continuano ad essere sviluppate. Alcune di esse consistono nel sostituire la matrice "neutra" Q in (1) con matrici che riflettano i *gusti* di chi naviga in internet. Altre assicurano che il ranking non sia troppo sensibile alle manipolazioni di coloro che cercano di migliorare il ranking delle proprie pagine.

Come conclusione generale, cosa abbiamo osservato? Un'idea semplice e intelligente ha portato ad una grandissima svolta nell'efficienza dei motori di ricerca e alla nascita di un impero commerciale. Nonostante l'implementazione sia di per sé un'impresa computazionale, l'idea di base richiede conoscenze di matematica "elementare", quali l'algebra lineare e la teoria delle probabilità. Questi strumenti relativamente standard, in particolare la diagonalizzazione delle matrici, hanno espresso pienamente il loro potere quando sono state usate al di fuori del loro "usuale" contesto. Non solo: abbiamo messo in evidenza le idee unificanti nella scienza, osservando che il teorema del punto fisso di Banach abbia applicazioni così lontane dalle sue origini.

Riferimenti bibliografici

[1] Eisermann, M. (2009) *Comment Google classe les pages webb*, http://images.math.cnrs.fr/Comment-Google-classe-les-pages.html.

[2] Langville, A. N & Meyer, C. D. (2005) *A Survey of Eigenvector Methods for Web Information Retrieval*, SIAM Review, Vol. 47, n. 1, pp. 135-161.

[3] Rousseau, C. & Saint-Aubin, Y. (2008) *Mathematics and technology*, SUMAT Series, Springer-Verlag (Esiste una versione francese del libro, pubblicata nella stessa serie.)