La matematica dell'importanza

Federico Poloni[*]

Il problema del ranking In questo articolo, vogliamo descrivere un problema che ha un'importanza e un'utilità pratica sempre maggiori negli ultimi anni caratterizzati dalla crescita di internet e dalla disponibilità sempre maggiore di dati da analizzare.

Supponete di fare una ricerca su internet per trovare siti che parlano dell'ultimo film della Disney. Ci sono diversi risultati pertinenti: per esempio, la pagina ufficiale del film, le recensioni sui giornali, e le pagine dei fan che pubblicano opinioni, immagini e commenti. Quando digitiamo il titolo del film, i motori di ricerca moderni riescono con un'efficacia sorprendente a distinguere quali pagine sono più importanti per noi (per esempio, il sito ufficiale) e quali sono secondarie (per esempio, le pagine dei fan). Vi siete mai chiesti come fanno?

Per gli argomenti di maggiore rilevanza, come un film famoso, è possibile chiedere a una persona di selezionare manualmente le pagine; però, farlo per tutte le possibili ricerche è inimmaginabile. Per questo è necessario un metodo automatico, un algoritmo.

Proviamo a ragionare come un matematico, e cerchiamo di formalizzare il problema: **dato un insieme di pagine web che parlano di un certo argomento, trovare quali sono più importanti.** Dobbiamo innanzitutto dire cos'è una "pagina web". Sicuramente, essa contiene del testo e delle immagini. C'è un altro elemento però: i collegamenti (link) tra una pagina e l'altra. Se indichiamo con delle frecce i collegamenti da una pagina all'altra, otteniamo una struttura come quella della Figura 1, che i matematici chiamano *grafo*.

Diverse soluzioni Il secondo problema, più difficile, è quello di decidere cosa vuol dire "la più importante". Qui il matematico si ferma, e riconosce che a questa domanda non c'è una risposta univoca, automatica. **Definire l'impor-**

[*]Dipartimento di Informatica, Università di Pisa, federico.poloni@unipi.it.

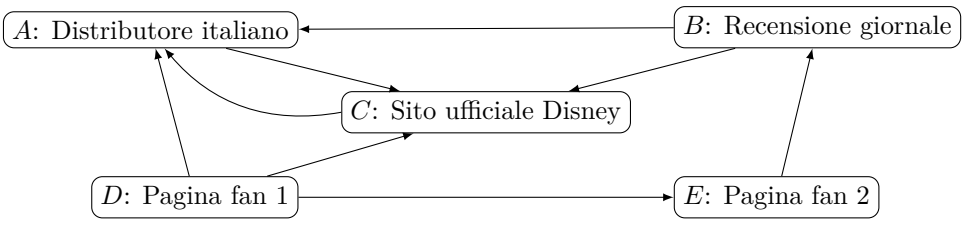

Fig. 1: Collegamenti tra una pagina e l'altra, esempio.

tanza è una scelta arbitraria. Una volta fatto, possiamo cercare dei modi di calcolarla, ma prima dobbiamo accordarci su una definizione.

Le prime possibilità che ci possono venire in mente sono le seguenti:

1. Le pagine 'migliori' sono quelle che contengono il nome del film nel titolo, oppure scritto più in grande.

2. Le pagine 'migliori' sono quelle che contengono più spesso il nome del film.

Riuscite ad individuare il grosso problema con criteri di questo tipo? Sono *facilmente manipolabili*: se un fan modifica il testo della sua pagina, riesce facilmente a conquistare il primo posto, scalzando i siti ufficiali. I motori di ricerca della prima generazione, negli anni '90, utilizzavano criteri di questo tipo, e abusi come questo erano all'ordine del giorno. Spesso pagine di 'spam' contenevano solo il nome di un film, ripetuto moltissime volte. Un criterio un po' migliore è il seguente:

3. Le pagine 'migliori' sono quelle verso le quali ci sono molti collegamenti.

Abbiamo introdotto un elemento nuovo: **sfruttare la struttura dei link** per ottenere un rating più affidabile. Manipolare questo criterio è un po' più scomodo, ma ancora fattibile: un fan deve semplicemente creare molte pagine fittizie, che nessuno mai guarderà, ma che contengono un link al suo sito.

Voti più e meno importanti Ci serve invece un criterio che tenga conto che non tutti i collegamenti hanno lo stesso valore: un link proveniente dal sito della Disney 'vale' molto di più che non uno proveniente dal blog di un fan. In uno slogan: *una pagina è importante se viene linkata da pagine importanti.* Questa definizione è ricorsiva: apparentemente, per calcolare l'importanza di una pagina dobbiamo sapere l'importanza di tutte le pagine da cui partono collegamenti a lei, e così via. In realtà, vedremo che per calcolare questo punteggio basta risolvere un sistema di equazioni lineari. Vediamo di formalizzare la cosa: chiamiamo $x_A, x_B, \ldots, x_E$ l'"importanza' delle pagine in Figura 1. Il sito B contiene collegamenti ad A e C; possiamo immaginare che esso 'voti' per A e C, o che 'distribuisca' la sua importanza equamente tra di essi:

$$x_C = \tfrac{1}{2}x_B + \ldots \text{altri termini}\ldots, \qquad x_A = \tfrac{1}{2}x_B + \ldots \text{altri termini}\ldots.$$

Analogamente, il D dividerà equamente la sua 'importanza' tra i tre siti verso cui ha link, e A la darà tutta al sito C. Complessivamente, le equazioni sono

$$\begin{cases} x_A = \frac{1}{2}x_B + x_C + \frac{1}{3}x_D, \\ x_B = x_E, \\ x_C = x_A + \frac{1}{2}x_B + \frac{1}{3}x_D, \\ x_D = 0, \\ x_E = \frac{1}{3}x_D. \end{cases} \tag{1}$$

Da sole, esse non sono ancora sufficienti a determinare univocamente l'importanza: dati $x_A, x_B, \ldots, x_E$ che soddisfano il sistema, possiamo moltiplicare tutto per due (o per tre, per quattro...) e ottenere un'altra soluzione altrettanto valida. Aggiungiamo una condizione per fissare il totale:

$$x_A + x_B + x_C + x_D + x_E = 1. \tag{2}$$

Il sistema formato da (1) e (2) ha una sola soluzione: $x_A = x_C = \frac{1}{2}, x_B = x_D = x_E = 0$. Cioè, *le uniche pagine importanti sono A e C*.

Possibili modifiche Vi soddisfa questo risultato? Se la risposta è no, dobbiamo **cambiare la definizione di importanza**. Per esempio, per evitare zeri nella soluzione, possiamo assumere che solo una certa percentuale dell'importanza, diciamo l'85%, venga distribuita in questo modo; il restante 15% viene suddiviso in parti uguali, in modo che nessuno resti a zero. Le equazioni così modificate diventano

$$\begin{cases} x_A = (1-\alpha)\frac{1}{5} + \alpha\left(\frac{1}{2}x_B + x_C + \frac{1}{3}x_D\right), \\ x_B = (1-\alpha)\frac{1}{5} + \alpha x_E, \\ x_C = (1-\alpha)\frac{1}{5} + \alpha\left(x_A + \frac{1}{2}x_B + \frac{1}{3}x_D\right), , \\ x_D = (1-\alpha)\frac{1}{5}, \\ x_E = (1-\alpha)\frac{1}{5} + \alpha\frac{1}{3}x_D. \end{cases} \tag{3}$$

dove abbiamo lasciato scritto esplicitamente il parametro $\alpha = 0.85$. Il sistema formato da (2) e (3) ha soluzione $x_A = 0.43$, $x_B = 0.063$, $x_C = 0.43$, $x_D = 0.030$, $x_E = 0.038$. Ora la soluzione sembra molto più ragionevole! Cattura il fatto che il fan 2 è più importante del fan 1, e che la recensione è più importante di entrambi. Ora siamo pronti per usare questo metodo su problemi più grandi che non uno con cinque sole pagine. Un momento, però...

Metodi di soluzione Come avete risolto i sistemi qui sopra (se avete provato a farlo da soli)? Molto probabilmente avete eliminato una variabile dopo l'altra, oppure avete sommato e sottratto tra loro le equazioni. Questi metodi richiedono un numero di operazioni che cresce come *il cubo del numero di pagine*. Se avessimo 10 pagine invece di 5, calcolarne l'importanza non richiederebbe il doppio del tempo, ma *otto volte tanto*. I numeri diventano difficili da gestire molto presto, anche per un calcolatore. Un computer moderno richiede circa un millisecondo per risolvere un sistema di 100 equazioni lineari in 100 incognite. Anche a questa velocità, però, calcolare l'importanza di *tutte* le pagine che parlano di un certo argomento è difficile: se abbiamo 100.000 pagine, servono 11 giorni. E se invece ne abbiamo 1.000.000?

Un metodo diverso, approssimato ma più efficiente, viene dal dare a queste equazioni un'interpretazione diversa. Supponiamo di guardare un utente annoiato che naviga tra le pagine seguendo ogni volta un link scelto a caso, tutti con la stessa probabilità. Dopo che ha seguito questa procedura per molto tempo, qual è la probabilità di trovarlo su una pagina piuttosto che su un'altra? Per esempio, chiamiamo x_C la probabilità di trovarlo su C. Perché questo succeda, ci sono tre possibilità: o la pagina precedente era la A, che succede con probabilità x_A; o era nella pagina B (e in questo caso però visto che B ha due link uscenti c'è solo $\frac{1}{2}$ di possibilità che finisca in C), oppure era nella pagina D (e in questo caso ha $\frac{1}{3}$ di possibilità di finire in C). Quindi, abbiamo l'equazione $x_C = x_A + \frac{1}{2}x_B + \frac{1}{3}x_D$. Ma questa equazione l'abbiamo già incontrata in (1), e così quelle delle altre pagine: **le quantità $x_A, x_B, \ldots, x_E$ calcolate più sopra rappresentano le probabilità di trovare un 'utente casuale' in ognuna delle pagine**. L'equazione (2) ci dice per l'appunto che queste probabilità devono sommare a uno. In altre parole, più una pagina è importante, più è facile finirci seguendo collegamenti a caso, il che è ragionevole.

Questo suggerisce un altro metodo di soluzione: simuliamo il comportamento di questo navigatore casuale. Partiamo da una delle pagine, per esempio D. Poi, con un generatore di numeri casuali, scegliamo un link tra quelli presenti sulla pagina, per esempio quello verso E, e seguiamolo. Poi scegliamo un link a caso tra quelli presenti su E (in questo caso uno solo), e continuiamo così. Teniamo traccia della percentuale del tempo che abbiamo passato su ognuna delle pagine, e al crescere di k, queste quantità tenderà con grande probabilità alla soluzione del sistema (1)+(2). Un metodo leggermente più efficiente è questo: partiamo al passo 0 con probabilità uguali di essere su ogni pagina: $[p_A(0), p_B(0), p_C(0), p_D(0), p_E(0)] = [\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}]$. Al passo 1, non è complicato calcolare che seguendo i collegamenti le probabilità di essere su ogni pagina sono $[p_A(1), p_B(1), p_C(1), p_D(1), p_E(1)] = [\frac{11}{30}, \frac{1}{5}, \frac{11}{30}, 0, \frac{1}{15}]$. Notate che non è possibile essere nella pagina D, perché non ci sono collegamenti che puntano ad essa. Al passo 2, le probabilità sono $[\frac{7}{15}, \frac{1}{15}, \frac{7}{15}, 0, 0]$, e al passo successivo arriviamo alla soluzione vera del sistema (1)+(2).

Teletrasporto Riusciamo ad attribuire un significato simile alle equazioni (3)? Sì, e anche questa volta la soluzione è molto elegante: esse corrispondono alle probabilità di incontrare in una certa pagina un navigatore che con probabilità α segue un collegamento, oppure altrimenti (con probabilità $1-\alpha$) si trasferisce su una pagina scelta a caso uniformemente tra tutte. Provate a scrivere le equazioni corrispondenti a questo comportamento e a verificarlo.

Costo Quante operazioni sono necessarie per calcolare le probabilità al passo $t+1$, date quelle al passo t? Le equazioni sono identiche a quelle di (1) (o (3)), solo che le quantità a sinistra dell'uguale si riferiscono al passo $t+1$, quelle a destra al passo t: per esempio,

$$p_C(t+1) = p_A(t) + \tfrac{1}{2}p_B(t) + \tfrac{1}{3}p_D(t).$$

C'è un addendo per ogni link *entrante* nella pagina C. Se ci sono k addendi, dobbiamo fare k moltiplicazioni e $k-1$ addizioni. Quindi il numero di operazioni in un 'passo' di questo algoritmo è $(2\ell - 1)n$, dove n è il numero di pagine e ℓ è il numero medio di link entranti in ogni pagina. Per esempio, nel grafo in Figura 1, in A e C arrivano tre link, in B ed E uno, in D nessuno, quindi $\ell = 8/5 = 1.6$ collegamenti.

Potrebbe preoccuparvi questa quantità ignota ℓ: ci sono pagine che ricevono molti collegamenti, per esempio non è irragionevole che la pagina di un film famoso venga raggiunta da centinaia di migliaia di link. Però questa elegante osservazione, che un matematico chiamerebbe 'teorema', vi dimostra che la media non può essere troppo alta:

Teorema 1. *Il numero medio di link entranti è uguale al numero medio di link uscenti da una pagina.*

Dimostrazione. Invece di pensare in termini di medie, pensiamo in termini di totali: ogni freccia ha una 'coda' e una 'punta', quindi il numero totale di frecce che escono da una pagina è uguale al numero totale di frecce che entrano in una pagina. □

Nel nostro esempio, da A, C, E esce un link solo, da B due, e da D tre: in media, 1.6 collegamenti, esattamente come previsto dal teorema. Anche per

pagine molto lunghe, il numero di collegamenti presenti su una pagina non supererà qualche decina, quindi ℓ non può essere troppo grande.

Un ultimo dettaglio: quanti passi di questo algoritmo sono necessari per ottenere una buona approssimazione delle 'importanze' vere? È difficile rispondere, anche per un matematico, ma per matrici grandi algoritmi di questo tipo di solito sono più efficienti degli quelli 'diretti' che richiedono n^3 operazioni. Nel caso del sistema (2)+(3), per esempio, già al passo 3 abbiamo quattro cifre significative esatte.

passo	A	B	C	D	E
1	0.3417	0.2000	0.3417	0.0300	0.0867
2	0.4139	0.1037	0.4139	0.0300	0.0385
3	0.4344	0.0627	0.4344	0.0300	0.0385
4	0.4344	0.0627	0.4344	0.0300	0.0385
5	0.4344	0.0627	0.4344	0.0300	0.0385

Hubs and authorities Una variante di questo metodo è un modello chiamato *hubs and authorities*: esso riconosce che ci sono anche pagine che pur non essendo 'buone' contengono link a pagine di buona qualità. Per esempio, una pagina di recensioni conterrà link a buoni film, mentre un buon film magari conterrà solo link a film dello stesso produttore, non necessariamente buoni. Ad ogni pagina quindi associamo due punteggi, che chiamiamo y e z (quindi $y_A, z_A, y_B, z_B, \dots$). I punteggi y indicano quanto una pagina è buona come 'recensione', e i punteggi z indicano quanto sono buoni i suoi contenuti. Una pagina ha un punteggio y più alto quando contiene link a pagine con z alto, e ha un punteggio z alto quando viene puntata da pagine con y alto. Sapreste scrivere delle equazioni per calcolare questi punteggi?

L'importanza dell'importanza Algoritmi come questi sono diventati molto popolari negli ultimi anni; il primo indice a farne uso nell'ambito dei motori di ricerca è stato il *pagerank* di Google, che è stato uno dei fattori che ne hanno determinato il successo. Analisi automatiche di questo tipo compaiono in sempre più siti: i siti di e-commerce vogliono trovare i prodotti più adatti da consigliare ai clienti, i produttori di pubblicità vogliono trovare gli *ad* più rilevanti da mostrarci...

Il ruolo del matematico Trovare il modo migliore di calcolare questi indici è un problema che interessa matematici e informatici a diversi livelli: bisogna trovare buone definizioni di concetti come 'importanza', saperli calcolare in modo efficiente, e dimostrarne le proprietà. Per esempio, un problema che va individuato e risolto è capire come vanno modificati gli algoritmi precedenti se da una pagina non esce nessun link. Oppure: provate a vedere cosa succede se rimpiazzate il grafo di Figura 1 con uno in cui ci sono due insiemi di pagine separati, senza collegamenti tra l'uno e l'altro. Dovreste notare che il sistema (1)+(2) ha infinite soluzioni, mentre il sistema (2)+(3) continua a funzionare. Un matematico è in grado di dimostrare questi risultati e spiegare il comportamento del metodo. Infine, non è poi facile dimostrare che l'algoritmo che abbiamo descritto qui sopra produce sempre una soluzione con tutti gli x positivi o nulli, che è fondamentale per poter assegnare loro il significato di 'importanza'.

Insomma, anche nella scienza del web e dei *big data* (algoritmi per lavorare con grandi quantità di dati) il ruolo del matematico è fondamentale.