Web-search: Reti e informazione

Trascrizione da Slide

Introduzione

Il materiale descritto in queste lezioni costituisce i Capitoli 13 e 14 del testo. Nella serie di lezioni corrispondenti a questi lucidi ci interessiamo al problema di individuare all'interno di una rete ed estrarre da essa informazioni rilevanti ad una data richiesta e lo faremo nel caso in cui la rete sia il Web.

Il Web è una rete diversa dalle reti fino ad ora considerate in questo corso:

- Le precedenti erano costituite da individui collegati da relazioni o da dispositivi di calcolo collegati mediante dispositivi fisici (reti wireless).
- Il Web è una rete costituita da documenti collegati mediante hyperlink.

Naturalmente, utilizzeremo ancora il grafo come modello di rete e, questa volta, lo strumento sarà l'algebra lineare.

1 Il World Wide Web

È un'applicazione progettata per condividere informazione su Internet fra il 1989 e il 1991, da Tim Berners-Lee, il cui progetto è basato su una architettura client-server:

- I documenti sono resi disponibili, sotto forma di pagine web, memorizzandoli in zone pubblicamente accessibili di taluni computer (server).
- L'accesso a tali pagine avviene mediante un'applicazione (browser) eseguita dal client e che accede agli spazi pubblici dei server.

Il principio logico fondazionale del web è l'ipertesto, nel quale l'informazione è organizzata in una struttura di rete:

- I documenti che costituiscono il web sono nodi di un grafo diretto.
- Si tratta, perciò di una organizzazione non lineare.

L'ipertesto è una implementazione del concetto di memoria associativa che ha una lunga storia...

1.1 Memoria associativa

Esistono, in effetti, numerosi precursori dell'ipertesto:

- Le citazioni: quando nei libri o negli articoli si citano altri libri o articoli che contengono, ad esempio, informazioni supplementari.
- I *crossing reference*: i riferimenti incrociati, che permettono di collegare argomenti diversi mediante relazioni di vario tipo.
- Esempio (catena di collegamenti): Nash Equilibrium → Game Theory → John Forbes Nash → A Beautiful Mind (film) → Apollo 13 (film) → Ron Howard → NASA → Conspiracy Theories → RAND.

Questa catena di collegamenti ricorda un po' il flusso di coscienza, ossia il modo in cui avvengono le associazioni mentali, o le mappe concettuali che molti studenti sperimentano nella preparazione di tesine.

1.2 MEMEX

Vannevar Bush, in un articolo del 1945 ("As we may think") provò a immaginare in che modo la nascente tecnologia dell'informazione poteva modificare il modo in cui le informazioni vengono memorizzate e con il quale si può accedere ad esse.

Egli osservò che:

- I metodi tradizionali per rappresentare l'informazione (libri, biblioteche, dizionari enciclopedici) sono sostanzialmente lineari (collezione di dati elencati sequenzialmente).
- Il modo in cui l'essere umano pensa e memorizza (e accede a) le sue esperienze è associativo (inizia a pensare a una cosa, che gliene fa venire in mente una seconda, e così via).

Bush immaginò anche un prototipo: MEMEX, una versione digitalizzata della conoscenza umana nella quale i dati sono collegati da link associativi. Bush è stato citato esplicitamente da Tim Berners-Lee nel suo progetto del Web.

2 Dal Web al Web 2.0

Inizialmente, il Web era costituito da un insieme di pagine al cui interno erano (possibilmente) presenti alcuni hyperlink ciascuno dei quali collegava quella pagina ad un'altra pagina (pagine "statiche"). Sono appunto gli hyperlink che implementano l'idea di memoria associativa.

Con il tempo, il Web si è evoluto:

- Sono state progettate pagine che permettevano di eseguire azioni (sottomettere una richiesta, effettuare un pagamento...). I link contenuti in queste pagine permettono di concludere le azioni e sono detti transazionali, per distinguerli dai link navigazionali.

- Sviluppo di strumenti per la creazione e il mantenimento di contenuti condivisi (Wikipedia), per il mantenimento di dati personali on-line (web-mail) e per la connessione di individui, invece che di documenti (social networks).

Evoluzioni così rilevanti che si è parlato di avvento del Web 2.0.

3 La struttura del Web

Anche il Web "navigazionale" può essere modellato come un grafo, in particolare come un grafo diretto:

- I nodi sono le pagine.
- Esiste un arco diretto (u, v) se la pagina u contiene un hyperlink alla pagina v (diciamo che u punta a v).

Osserviamo che i nodi non hanno coscienza degli archi entranti: una pagina Web non ha modo di conoscere le pagine che la puntano!

In [Broder et al., 2000] viene studiata la struttura del grafo del web. È stato rilevato che esso aveva una struttura *bow-tie* (a farfalla) contenente una (unica) SCC gigante. La struttura comprende:

- SCC (Strongly Connected Component): 56 milioni di nodi.
- IN: 44 milioni di nodi.
- OUT: 44 milioni di nodi.
- Tendrils: 44 milioni di nodi.
- Tubes.
- Disconnected components.

4 Ricerca e classificazione

La ricerca di documenti è un problema antico, ma tradizionalmente avveniva in ambienti specifici (cercatore e cercato erano dello stesso ambiente, es. avvocati, medici). La difficoltà era causata dalla scarsità di documenti disponibili.

L'avvento del Web ha cambiato sostanzialmente la questione:

- Il Web è un'immensa collezione di documenti appartenenti alle categorie più disparate.
- Chiunque cerca nel Web (lo scienziato, l'avvocato, il medico, il letterato).
- La difficoltà è ora la **sovrabbondanza** di documenti disponibili.

4.1 Ricerca non specializzata

Tradizionalmente le parole chiave erano in numero limitato e con significato ben definito. Nel Web, il numero di parole chiave è potenzialmente illimitato e la stessa parola può avere tanti diversi significati (polisemia). Esempio: la parola "albero" può essere un membro del regno vegetale, un albero genealogico, o un grafo privo di cicli.

4.2 La necessità di sfoltire le alternative

Per un individuo, non sapere dove cercare un documento o sapere di doverlo cercare in una collezione di milioni di documenti è all'incirca la stessa cosa. Affinché uno strumento di web search sia efficace, è necessario "sfoltire" l'immensa mole di pagine, eliminando quelle meno rilevanti e proponendo una lista in ordine di rilevanza non decrescente (ranking).

5 Link Analysis

Le caratteristiche interne a una pagina non consentono, da sole, di capire quanto una pagina sia rilevante. Gli hyperlink sono uno strumento utile a questo scopo.

Analogia: Quando prepariamo una tesi, se un articolo in bibliografia è citato dalla maggioranza degli altri articoli, esso sarà probabilmente molto rilevante.

Generalizzando:

- Dopo aver individuato un insieme di pagine significative rispetto a una richiesta.
- Consideriamo maggiormente rilevanti le pagine che sono molto puntate da altre pagine nell'insieme.

La rilevanza di una pagina è calcolata tramite analisi dei link che la coinvolgono.

6 HITS: Hubs and Authorities

Il metodo *Hyperlink-Induced Topic Search* (HITS), o *Hubs and Authorities*, proposto da Kleinberg, considera due indici per ogni pagina:

- **Authority (a):** esprime la rilevanza della pagina ai fini della ricerca (quanto è puntata).
- **Hub (h):** esprime l'autorevolezza della pagina nel conferire autorità alle pagine alle quali punta.

Possiamo pensare che:

- Il valore di autorità di una pagina i (a_i) sia il numero di pagine attinenti che puntano ad essa.
- Il valore di hub di una pagina i (h_i) sia il numero di pagine attinenti alle quali essa punta.

Sia M la matrice di adiacenza del sottografo attinente alla ricerca (n pagine):

$$M[i,j] = \begin{cases} 1 & \text{se } i \to j \\ 0 & \text{altrimenti} \end{cases}$$

Allora:

$$a_i = \sum_{1 \le j \le n} M[j,i] \quad \text{e} \quad h_i = \sum_{1 \le j \le n} M[i,j]$$

Raffiniamo l'idea:

- L'autorità è tanto più elevata quanto più autorevoli sono le pagine che la puntano (hanno alto valore di hub).
- L'hub è tanto più elevato quanto più rilevanti sono le pagine a cui punta (hanno alto valore di authority).

Procedimento iterativo:

$$a_i^{(k+1)} = \sum_{1 \le j \le n} M[j,i] h_j^{(k)}$$

$$h_i^{(k+1)} = \sum_{1 \le j \le n} M[i,j] a_j^{(k+1)}$$

In forma matriciale (vettori colonna):

$$a^{(k+1)} = M^T h^{(k)}$$

$$h^{(k+1)} = M a^{(k+1)}$$

Partendo da $h_i^{(0)} = 1$ per ogni $1 \le i \le n$.

Combinando le espressioni:

$$h^{(k)} = M a^{(k)} = M M^T h^{(k-1)}$$

Induttivamente:

$$h^{(k)} = (M M^T)^k h^{(0)}$$

Analogamente per $a^{(k)}$.

6.1 Convergenza di HITS

Le espressioni sono somme di termini non negativi, quindi i valori tendono a crescere. Possiamo parlare di convergenza solo in presenza di opportuna normalizzazione.

Teorema: Esistono un valore $c \in \mathbb{R}^+$ e un vettore $z \in \mathbb{R}^n$ non nullo tali che, comunque si scelga un vettore $h^{(0)}$ a coordinate positive:

$$\lim_{k \to \infty} \frac{h^{(k)}}{c^k} = z$$

Per dimostrarlo, richiamiamo nozioni di algebra lineare:

- La matrice MM^T è reale e simmetrica.
- Una matrice reale e simmetrica ha n autovettori e n autovalori reali; gli autovettori formano una base ortonormale per $\mathbb{R}^n$.
- MM^T è semidefinita positiva ($\forall x \in \mathbb{R}^n : x^T (MM^T) x \ge 0$).
- Se è semidefinita positiva, gli autovalori sono non negativi.

Dimostrazione (sintesi): Siano $z_1, z_2, \ldots, z_n$ autovettori di MM^T (base ortonormale) e $c_1 \ge c_2 \ge \cdots \ge c_n \ge 0$ i corrispondenti autovalori. Esprimiamo $h^{(0)}$ come combinazione lineare: $h^{(0)} = \sum_{i=1}^n q_i z_i$. Allora:

$$h^{(k)} = (MM^T)^k h^{(0)} = \sum_{i=1}^n q_i c_i^k z_i$$

Dividendo per c_1^k:

$$\frac{h^{(k)}}{c_1^k} = q_1 z_1 + q_2 \left(\frac{c_2}{c_1}\right)^k z_2 + \cdots + q_n \left(\frac{c_n}{c_1}\right)^k z_n$$

Se $c_1 > c_2$, per $k \to \infty$, i termini con $(c_i/c_1)^k$ tendono a 0. Il limite è $q_1 z_1$. Resta da dimostrare che $q_1 \neq 0$ (ovvero $h^{(0)} \cdot z_1 \neq 0$) per qualsiasi $h^{(0)}$ positivo. (La dimostrazione prosegue mostrando per assurdo che non può essere nullo se $h^{(0)}$ ha coordinate positive e z_1 è autovettore principale).

Conclusione: Se $c_1 > c_2$ (situazione tipica), il metodo HITS converge sempre ad un vettore parallelo all'autovettore z_1, indipendentemente dai valori iniziali (purché positivi). La valutazione dipende solo dalla matrice M.

6.2 HITS: Considerazioni

HITS valuta la pagina in due ruoli: authority (rilevanza) e hub (attitudine a dare rilevanza). Questo metodo si presta bene a modellare situazioni in cui le pagine sono partizionate in due sottoinsiemi "semanticamente" distinti (es. e-commerce: siti come eBay puntano ai venditori, ma i venditori non si puntano tra loro perché concorrenti).

7 PageRank

In altre situazioni (es. articoli di ricerca), sono le pagine rilevanti a decretare direttamente la rilevanza delle pagine alle quali puntano. Il PageRank (da Larry Page) modella queste situazioni. Si basa sull'analisi dei soli link entranti.

Concetto del Fluido: Assume che nella rete sia presente una unità di fluido, inizialmente distribuita equamente ($1/n$ per nodo). Iterativamente, ciascun nodo distribuisce equamente il fluido fra i suoi vicini. Detto ω_j il numero di archi uscenti dalla pagina j:

$$f_i^{(k+1)} = \sum_{1 \leq j \leq n : j \to i} \frac{f_j^{(k)}}{\omega_j}$$

Teorema: Se il grafo è fortemente connesso, allora esiste ed è unico il $\lim_{k\to\infty} f^{(k)}$. Il vettore limite f^* è una configurazione di equilibrio:

$$f_i^* = \sum_{1 \leq j \leq n : j \to i} \frac{f_j^*}{\omega_j}$$

Se il grafo non è fortemente connesso, il fluido tende a concentrarsi nelle regioni del grafo dalle quali non si può uscire ("vicoli ciechi" o *sinks*).

8 PageRank Scalato

Per evitare che il rank si concentri nei vicoli ciechi, si modifica il procedimento. Fissato un parametro $s \in [0, 1]$:

- Una frazione s del rank è distribuita lungo gli archi uscenti.

- La parte rimanente $(1 - s)$ è distribuita uniformemente fra tutti i nodi.

Formula:

$$r_i^{(k+1)} = \left(\sum_{1 \leq j \leq n : j \to i} s \frac{r_j^{(k)}}{\omega_j} \right) + \frac{1-s}{n}$$

In forma matriciale $r^{(k+1)} = Nr^{(k)}$, dove:

$$N[i,j] = \begin{cases} \frac{s}{\omega_j} + \frac{1-s}{n} & \text{se } j \to i \\ \frac{1-s}{n} & \text{altrimenti} \end{cases}$$

8.1 Analisi Spettrale

Cerchiamo un vettore di equilibrio $r^* = Nr^*$. Questo significa che r^* deve essere un autovettore di N con autovalore 1.

Teorema di Perron (semplificato): Se A è una matrice reale $n \times n$ a elementi positivi, allora A ha un autovalore $c \in \mathbb{R}^+$ strettamente maggiore del modulo di ogni altro autovalore, e l'autovettore corrispondente è unico e a elementi positivi.

La matrice N del PageRank scalato è positiva e stocastica per colonne (la somma su ogni colonna è 1). **Teorema:** Se una matrice è stocastica, ha un autovalore massimo pari a 1. Quindi esiste un unico vettore di equilibrio r^* (PageRank).

9 PageRank e Random Walks

Il PageRank ha un'interpretazione legata al *random walk* (cammino casuale).

- **Random Walk standard:** Si sceglie un nodo a caso, poi si sceglie uniformemente un arco uscente, e così via. La probabilità di trovarsi nel nodo i al passo k corrisponde alla formula del PageRank semplice.
- **Scaled Random Walk:** Da un nodo u, con probabilità s si segue un arco uscente, con probabilità $(1 - s)$ si salta a un nodo qualsiasi del grafo (teletrasporto).

La probabilità di trovarsi nel nodo i al passo k del random walk scalato è data da:

$$P(sw_i^k = 1) = \sum_{j \in V : (j,i)} \frac{s}{\omega_j} P(sw_j^{k-1} = 1) + \frac{1-s}{n}$$

Che coincide con la formula del PageRank scalato $r_i^{(k)}$.

10 Modern Web Search

L'enorme crescita del Web (in dimensioni e contenuti) ha richiesto la revisione delle tecniche. Si è passati dalla Link Analysis pura a tecniche che combinano Link Analysis e analisi del contenuto testuale (es. *anchor text*). Viene considerata anche la frequenza di apertura delle pagine (click-through).

Poiché esistono interessi commerciali a comparire nelle prime posizioni, è nata un'industria (SEO) per alterare il rank "naturale". Questo induce i motori di ricerca a rivedere continuamente e tenere segreti i propri algoritmi.