11. Web Information Retrieval

Come funzionano i motori di ricerca

L’utente inserisce una query. Il motore di ricerca cerca i termini della query nel suo vastissimo indice. L’indice contiene milioni di risultati, che vengono ordinati in base a una nozione di rilevanza e presentati all’utente come una lista ordinata. Programmi automatici (detti anche **spiders** o **crawlers**) scandagliano tutte le pagine del web per costruire tale indice.

Che cosa è rilevante?

Come può un motore di ricerca sapere cosa sta cercando veramente l’utente? Il concetto di **information need** rappresenta il desiderio di localizzare e ottenere informazioni per soddisfare un bisogno conscio o inconscio (secondo la definizione di Wikipedia). La **user query** è l’insieme di termini consecutivi formulati da un utente per esprimere tale bisogno informativo. L’**intent della query** rappresenta il compito, l’obiettivo o l’intenzione dell’utente espressa attraverso la query: la stessa query, se formulata da utenti diversi, può riflettere intenzioni differenti.

Ranking: PageRank di Google

Una nozione di rilevanza basata esclusivamente sulle frequenze dei termini non è sufficiente per ordinare miliardi di documenti. Vengono quindi applicate misure complesse per valutare la qualità, l’affidabilità e l’autorevolezza delle pagine web. Tra queste misure troviamo quelle basate sulle proprietà topologiche delle reti, come ad esempio il **PageRank** di Google, quelle che assegnano pesi differenti a seconda del campo in cui appare un termine (ad esempio titolo, sottotitolo, corpo del testo), e quelle che considerano aspetti semantici e proprietà spazio-temporali, come la **freschezza** di una pagina. Il PageRank crea un meccanismo circolare in cui i siti già popolari tendono ad acquisire sempre maggiore visibilità, generando un effetto "i ricchi diventano più ricchi", per cui pochi siti dominano le prime posizioni nei risultati.

Pagina dei risultati del motore di ricerca

Una **Search Engine Results Page (SERP)** mostra di default 10 risultati. In media, i primi 5 di questi risultati sono visibili senza dover effettuare uno scroll verso il basso. In questo contesto, si prende in considerazione il comportamento dell’utente rispetto alla **prima pagina** dei risultati dei motori di ricerca, ovvero i primi 10 risultati visualizzati.

Bias di posizionamento dei risultati

Quando cercano informazioni, gli utenti medi non valutano sistematicamente tutti i risultati, ma si limitano a cliccare sui primi risultati visualizzati. Studi sono stati condotti confrontando le reazioni degli utenti quando ricevevano le liste dei risultati in ordine normale rispetto a un ordine invertito.

Bias di visualizzazione e tracciamento oculare

La percentuale di visualizzazioni e i clic del mouse sono stati confrontati tra l'ordine normale e l'ordine invertito dei primi 10 risultati. Anche se i risultati mostrano una certa consapevolezza da parte degli utenti, la fiducia eccessiva negli algoritmi di ranking può avere un impatto negativo sui siti posizionati più in basso.

Impatto del posizionamento sul business

I risultati in prima posizione ricevono oltre 10 volte più clic rispetto a quelli in sesta posizione. Supponiamo che l'azienda A e l'azienda B si trovino rispettivamente in prima e in sesta posizione per una parola chiave di valore commerciale. B dovrebbe acquistare 10 volte il numero di clic che ottiene A per ottenere lo stesso traffico. Ma quanto costa un clic?

Valore economico del clic

Nella pubblicità online, il costo di un singolo clic, noto come Cost-Per-Click (CPC), è in media di $1. Tuttavia, su Google alcune parole chiave possono essere estremamente costose (parole chiave più costose su Google nel 2011):

Insurance: 54,91$
Mortgage: 47,12$
Attorney: 47,07$
Claim: 45,51$
Loans: 44,28$
Lawyer: 42,51$

Bing è ancora più costoso (parole chiave più costose su Bing nel 2015):

Lawyers: 109,21$
Attorney: 101,77$
Structured settlements: 78,39$

Web crawling

Il **web crawling** è il processo mediante il quale si raccolgono pagine dal Web. L'obiettivo è raccogliere rapidamente ed efficientemente il maggior numero possibile di pagine Web utili, insieme alla struttura dei collegamenti che le interconnettono.

Funzionamento base di un crawler

Un **crawler** (noto anche come spider) inizia da una serie di URL noti, detti *seed*. Questi vengono recuperati e analizzati, si estraggono gli URL a cui puntano, e gli URL estratti vengono inseriti in una coda (nota come *URL frontier*). Il crawler recupera quindi ogni URL nella coda e ripete il processo.

Requisiti di un crawler

Un crawler deve rispettare alcuni requisiti fondamentali. Deve garantire **robustezza**, evitando le cosiddette *spider trap*, ovvero situazioni in cui il crawler rischia di scaricare un numero infinito di pagine all'interno di un dominio specifico. Deve inoltre rispettare la **politeness**, ossia seguire le politiche dei server Web che regolano la frequenza con cui i crawler possono visitare i siti.

Robustezza

Il web crawling non è realizzabile con una sola macchina: tutte le fasi precedenti sono distribuite. Esistono inoltre pagine maliziose, pagine di spam e *spider trap*, anche generate dinamicamente. Anche le pagine non maliziose possono rappresentare una sfida: la latenza e la larghezza di banda verso server remoti può variare, e molti webmaster impongono linee guida specifiche. È importante anche definire fino a che punto si debba esplorare la gerarchia degli URL di un sito, e gestire il problema di mirror del sito e pagine duplicate.

Politeness

La **politeness esplicita** consiste nelle specifiche fornite dai webmaster su quali porzioni del sito possono o non possono essere scansionate, ad esempio attraverso il file `robots.txt`. La **politeness implicita** impone di non sovraccaricare un sito con troppe richieste, anche in assenza di specifiche formali.

robots.txt

Il file `robots.txt` è un protocollo per fornire agli spider (i "robots") un accesso limitato a un sito Web, introdotto originariamente nel 1994. Maggiori informazioni sono disponibili su www.robotstxt.org. Il sito Web comunica ciò che può (o non può) essere scansionato. Per un server, si crea un file chiamato `robots.txt`, che specifica le restrizioni di accesso. Questo file contiene un insieme di regole che i client devono seguire.

Esempio di robots.txt

Un esempio:

```
User-agent: *
Disallow: /yoursite/temp/
User-agent: searchengine
Disallow:
```

Per tutti i user-agent (cioè i nomi dei client) viene proibito l'accesso alla directory `/yoursite/temp`. Per i client chiamati "searchengine" non è vietato nulla: tutto è quindi accessibile.

llms.txt

È stata proposta la standardizzazione di un file dedicato per fornire linee guida ai LLM (Large Language Models) su come interpretare i contenuti di un sito Web durante l'inferenza. Maggiori dettagli sono disponibili su https://llmstxt.org. Si tratta di un file in formato Markdown da aggiungere ai siti Web per fornire contenuti adatti all'elaborazione da parte dei LLM. Il file, denominato `llms.txt`, è leggibile sia dagli esseri umani che dai LLM, e consente metodi di elaborazione fissi (ad esempio tramite parser e regex).

Esempio di llms.txt

```
# Title
> Optional description goes here
Optional details go here

## Section name
- [Link title](https://link_url): Optional link details

## Optional
- [Link title](https://link_url)
```

Rage Against the Machine

Cloudflare ha sviluppato **AI Labyrinth** per mitigare l'impatto dello scraping non autorizzato del web da parte di bot che raccolgono dati per l'addestramento di intelligenze artificiali. Il sistema genera pagine esca create dall'IA che appaiono autentiche, ma sono progettate per fuorviare i bot e consumarne le risorse. Monitorando le interazioni con queste pagine esca, Cloudflare è in grado di rilevare e inserire in blacklist gli scraper maligni con elevata precisione.

URL frontier

Le pagine vengono aggiunte alla *URL frontier* secondo le seguenti strategie. Nella **strategia breadth first**, data una pagina Web nella URL frontier, si aggiungono tutte le pagine collegate dalla pagina corrente. Questo metodo garantisce una copertura ampia ma superficiale. Nella **strategia depth first**, invece, data una pagina Web nella URL frontier, si segue il primo link presente nella pagina corrente fino a raggiungere la prima pagina priva di collegamenti.

Ricerca sul web

Esistono migliaia di miliardi di pagine sul Web, ma la maggior parte non è particolarmente interessante. Supponiamo di voler visitare il sito di eBay, ma di non sapere che l'URL è `www.ebay.com`. Ci sono milioni di pagine Web che contengono il termine "eBay", e potrebbero esserci siti che menzionano la parola "eBay" con maggiore frequenza rispetto al sito ufficiale stesso. È quindi necessaria una nozione di **popolarità**, oltre che una nozione di **rilevanza**.

Recupero dell'informazione sul Web

Rispetto ai motori di ricerca testuali tradizionali, i sistemi di recupero dell'informazione sul Web costruiscono il ranking combinando almeno due evidenze di rilevanza: il grado di corrispondenza di una pagina, ovvero il **content score**, e il grado di importanza della pagina, ovvero il **popularity score**. Il **content score** può essere calcolato utilizzando uno dei modelli di recupero dell'informazione già descritti. Il **popularity score**, invece, può essere calcolato analizzando la struttura dei collegamenti ipertestuali tra le pagine indicizzate, attraverso uno o più modelli di analisi dei link. La presenza di link rappresenta una forma di attribuzione di autorità verso alcune pagine? Questo può essere utile ai fini del ranking?

Analisi semplice dei link

I collegamenti ipertestuali sono potenti fonti di autenticità e autorità. Un principio base, chiamato **The Good, The Bad and The Unknown**, si basa su una logica iterativa semplice: i nodi "buoni" non collegheranno mai nodi "cattivi"; se punti a un nodo "cattivo", allora anche tu sei considerato "cattivo"; se invece un nodo "buono" punta a te, allora sei "buono".

Analisi delle citazioni

La frequenza delle citazioni è una stima della popolarità di un ricercatore. La **frequenza di accoppiamento bibliografico** (bibliographic coupling) indica che articoli che co-citano gli stessi articoli sono tra loro correlati. L'indicizzazione delle citazioni viene utilizzata come strumento per la valutazione delle riviste scientifiche. Una domanda fondamentale è: da chi viene citato un determinato autore? (Garfield, 1972). Una prima anticipazione del concetto di **PageRank** è presente già nel lavoro di Pinski e Narin negli anni '70, i quali si chiedevano: quali riviste scientifiche sono realmente autorevoli?

PageRank

La tecnica **PageRank** per l'analisi dei collegamenti assegna un punteggio numerico compreso tra 0 e 1 a ogni nodo nel grafo del web. Il punteggio PageRank di un nodo dipende dalla struttura dei link all'interno del grafo del web. Data una query, un motore di ricerca calcola per ogni pagina web uno **score composito** che combina centinaia di caratteristiche (come la **similarità coseno**) insieme al punteggio PageRank. Questo score composito viene utilizzato per generare un elenco ordinato di risultati in risposta alla query.

Il navigatore casuale

Consideriamo una navigatrice casuale, Alice, che inizia un **cammino casuale** sul web, partendo da una pagina. Alice è estremamente annoiata e vaga senza meta tra le pagine web. Il suo browser ha un pulsante speciale "**sorprendimi**" in alto, che, quando cliccato, la porta a una pagina web casuale. Ogni volta che una pagina si carica, Alice sceglie se cliccare su un link casuale presente nella pagina oppure usare il pulsante "sorprendimi". Alice è così annoiata che intende continuare a navigare sul Web in questo modo per sempre.

Definiamo più formalmente il comportamento di Alice. Alice naviga seguendo questo algoritmo:

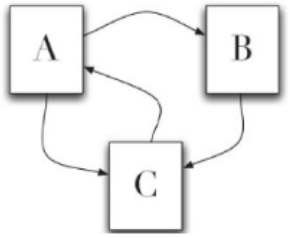
1. Scegli un numero casuale r tra 0 e 1
2. Se $r > \lambda$ allora:
   clicca il pulsante "sorprendimi"
3. Se $r \leq \lambda$ allora:
   clicca un link casuale nella pagina corrente
4. Ripeti dall'inizio

Grazie al pulsante "sorprendimi" di Alice, possiamo garantire che prima o poi raggiungerà **ogni pagina** su Internet.

Ora supponiamo che, mentre Alice sta navigando, tu entri nella stanza e guardi casualmente la pagina web sul suo schermo. Qual è la probabilità che stia guardando il sito di eBay? Quella probabilità corrisponde al **PageRank** di eBay.

Calcolo del PageRank

Il calcolo del **PageRank** corrisponde alla ricerca della **distribuzione di probabilità stazionaria** di un cammino casuale sul grafo del Web. Un cammino casuale è un caso particolare di **catena di Markov**, in cui lo stato successivo dipende unicamente dallo stato attuale.



Se il web è composto dalle 3 pagine mostrate in figura (A, B, C), il PageRank di C dipende dal PageRank di A e B secondo la formula:

$$PR(C) = \frac{PR(A)}{2} + \frac{PR(B)}{1}$$

Il PageRank conferito da un collegamento in uscita è uguale al punteggio PageRank del documento diviso per il numero di link in uscita. Si inizia assumendo che i valori di PageRank per tutte le pagine siano uguali, quindi si itera il calcolo. Dopo alcune iterazioni, i valori convergono ai seguenti:

$PR(C) = 0{,}4$
$PR(A) = 0{,}4$
$PR(B) = 0{,}2$

Uso del PageRank in Google

Il **PageRank** è oggi solo uno dei tanti fattori che determinano il punteggio finale di una pagina Web in Google. Fa parte di un sistema di ranking molto più ampio, che si ritiene tenga conto di oltre 200 diversi "**segnali**" (variabili di ranking), tra cui:

Caratteristiche linguistiche, come frasi, sinonimi, errori ortografici.
Caratteristiche della query, legate alla lingua o a termini/frasi in tendenza.
Caratteristiche temporali: ad esempio, per query legate a notizie, è preferibile restituire documenti indicizzati di recente, mentre per query fattuali è meglio rispondere con pagine più stabili e "resilienti".
Caratteristiche di personalizzazione, legate alla cronologia delle ricerche dell'utente, al comportamento e al contesto sociale.

Hyperlink-Induced Topic Search (HITS)

In risposta a una query, invece di restituire un elenco ordinato di pagine che soddisfano la query, il modello HITS cerca **due insiemi di pagine interconnesse**:

Le **hub pages** sono buoni elenchi di collegamenti su un determinato argomento, ad esempio "La lista di Bob dei link sul cancro".
Le **authority pages** sono pagine che compaiono ripetutamente nei buoni hub per quel tema.

Questo metodo è più adatto per query su argomenti ampi, piuttosto che per query di localizzazione di una singola pagina. Fornisce una visione più ampia dell'opinione comune.

Hub e authority

Una buona **hub page** per un argomento punta a molte pagine autorevoli su quell'argomento.
Una buona **authority page** per un argomento è quella a cui puntano molte buone hub.
Questa è una definizione circolare, ma può essere trasformata in un calcolo iterativo.

Ricerca semantica

Il termine "**information retrieval**" è lo standard, ma nella pratica tradizionale non è del tutto corretto. In genere, si ottiene un recupero di documenti, ma il resto del lavoro è lasciato all'utente.

La **ricerca semantica** consiste nel fare ricerche su grafi di conoscenza strutturati, invece che su semplici testi. Alcuni esempi includono:

- **Google Knowledge Graph**
- **Facebook Graph Search**
- **Satori** di Bing
- Strumenti come **Wolfram Alpha**