Il web può essere visto come un'ingente collezione non strutturata e distribuita. I dati presenti all'interno di questa collezione sono eterogenei (testi, immagini, suoni e video).

Strumenti per la ricerca sul web

Esistono diverse modi per identificare informazioni rilevanti sul web:

- ricerca diretta dato un particolare URL;
- ricerca mediante link, ossia navigando;
- servizi web per la ricerca.

I principali servizi web per la ricerca sono rappresentati da:

- motori di ricerca, i quali indicizzano una porzione di documenti web e permettono all'utente di formulare query e reperire indirizzi di pagine web pertinenti;
- portali web, oltre a mettere a disposizione un motore di ricerca, classificano per argomento i documenti web e forniscono un'interfaccia per la navigazione del catalogo dei documenti;
- sistemi a supporto del commercio elettronico, come ad esempio i recommender systems.

Evoluzione dei motori di ricerca

L'evoluzione dei motori di ricerca può essere divisa in tre macrofasi.

1. La prima generazione di motori di ricerca utilizzava solamente i dati testuali presenti all'interno delle pagine, analizzando parametri come la frequenza delle parole.
2. Con la seconda generazione si è invece passati all'analisi di dati specifici per il web, analizzando aspetti come la connettività di una pagina (i link), i link effettivamente utilizzati dagli utenti o il modo in cui gli utenti fanno riferimento ad una particolare pagina.
3. La terza generazione, quella attuale e ancora in fase sperimentale, è invece caratterizzata dal tentativo di soddisfare "la necessità oltre la query".

Tra le novità introdotte da quest'ultima generazione le più interessanti sono:

- analisi semantica,
- spostamento del focus sulle necessità dell'utente piuttosto che sulla query espressa,
- determinazione del contesto,
- interazione con l'utente,
- integrazione di ricerca e analisi del testo.

Struttura di un motore di ricerca

Ad alto livello, la struttura di un motore di ricerca è del tutto analoga a quella di un sistema di IR classico.

Un motore di ricerca è composto da una componente che consente all'utente di esprimere le proprie necessità informative attraverso la formulazione di una query, una componente di indicizzazione, una componente di matching e un componente che si occupa della visualizzazione dei dati.

La novità rispetto agli IRS classici è rappresentata dai differenti contenuti trattati dai sistemi. Nel IRS web i contenuti sono rappresentati dall'intero insieme delle pagine web pubbliche. Questa fondamentale differenza rende necessaria l'introduzione di una nuova componente, un sistema di raccolta dei documenti, detto anche di gathering.

Come detto, le pagine web corrispondono ad un documento nell'IR tradizionale l'insieme delle pagine pubblicamente indicizzabili corrispondono alla collezione nell'IR tradizionale. Da questo insieme di pagine sono escluse tutte quelle che richiedono autorizzazioni, le pagine dinamiche, ...

Problematiche nell'IR su web

Le problematiche nel campo dell'IR su web sono molteplici. Dal punto di vista dei dati:

- la grande quantità dei dati porta a problemi di scalabilità;
- la forte distributività dei dati porta a problemi di affidabilità delle sorgenti e delle connessioni;
- la volatilità dei dati rappresenta un altro problema, il 40% dei documenti cambia ogni mese e questo rende necessario un continuo aggiornamento dei dati e degli indirizzi;
- la ridondanza dei dati a livello fisico e a livello semantico è fonte di problemi dal punto di vista dell'efficienza (il 30% dei dati presenti sul web è duplicato);
- la qualità dei dati rappresenta un ulteriore problema, sul web sono infatti presenti grandi quantità di dati non validi, obsoleti o contenenti errori; infine
- l'eterogeneità dei dati porta a problemi di gestione e compatibilità dei diversi media, formati, strutture, lingue e alfabeti utilizzati sul web.

Anche dal punto di vista dell'interazione con l'utente le problematiche sono diverse. Come specificare le proprie necessità informative? Come presentare grandi quantità di documenti reperiti?

Crawler

La prima fase del funzionamento di un motore di ricerca è quella di raccolta dei documenti da indicizzare. Esistono due modalità per lo svolgimento di questa fase:

1. Le pagine web vengono fornite direttamente al motore di ricerca dai proprietari di queste;
2. il motore di ricerca è dotato di un agente software, detto crawler, che attraversa il web per identificare pagine nuove o aggiornate e spedire al server che si occuperà di indicizzarle. Il crawler naviga sul web usando come punti di partenza URL noti per essere punti di accesso interessanti. La visita comincia da queste pagine e continua percorrendo i link identificati all'interno delle pagine.

Delle linee guida per il comportamento dei crawler possono essere fornite dal proprietario di un sito web specificando un file alla radice del web server che ospita il sito. Questo file consente di indicare quali

contenuti che possono essere indicizzati (Robots.txt). I crawler possono scegliere di ignorare tali indicazioni ma potrebbero essere bloccati.

L'ordine in cui gli URL vengono visitati da un crawler è importante ed esistono diverse strategie per definirlo.

- Breadth first, vengono visitate prima tutte le pagine raggiungibili con un link dalla pagina corrente. Questa modalità è adatta per siti web che trattano argomenti in relazione tra loro. La copertura risultante sarà ampia ma superficiale.
- Depth first, viene visitato il primo link indentificato nella pagina corrente e si continua con questo procedimento fino a che non si incontra una pagina senza link.

Un buon schema di ordinamento dei risultati dei motori di ricerca dipende anche dall'ordine con cui si visitano le pagine migliori. Questo è quello che viene fatto dall'algoritmo PageRank.

Tipologie e focused crawler

Esistono diverse tipologie di web crawler:

- tradizionali, visitano l'"intero" web per rinnovare gli indici del motore di ricerca di cui fanno parte;
- periodici, visitano porzioni di web e aggiornano sottoinsiemi dell'indice;
- incrementali, selezionano parti del web e modificano in modo incrementale gli indici;
- focused, visitano solamente le pagine relative ad un dato argomento.

Un focused crawler è caratterizzato dalla presenza di due componenti: un classificatore e un distillatore di pagine hub. Il classificatore si occupa di determinare la rilevanza delle pagine e dei link uscenti da ogni pagina rispetto agli argomenti specifici del crawler. Il distillatore ha invece il compito di identificare le pagine hub, le pagine non rilevanti che risultano comunque interessanti in quanto contenengono molto out-link a pagine rilevanti. Un focused crawler visita le pagine basandosi sui punteggi di rilevanza determinati dal suo classificatore e dal suo distillatore.

Indicizzazione

In fase di indicizzazione possono essere tracciati anche dati ausiliari come: data di indicizzazione, autore, lingua e tipo di pagina web.

Le tecniche di indicizzazione e compressione possono ridurre la dimensione del file inverted di circa il 30% della dimensione del testo, una percentuale che può essere ulteriormente ridotta applicando tecniche come l'eliminazione delle stopwords. Questo significa che per mantenere un indice relativo a 100 milioni di pagine sono necessari circa 15GB di spazio su disco.

Interfaccia utente

Molti motori di ricerca utilizzano delle brevi descrizioni del contenuto delle pagine web, dette snippet, per

dare un'idea del contenuto dei documenti reperiti all'utente. Gli snippet vengono mantenuti insieme all'indice.

Il linguaggio di interrogazione utilizzato dai motori di ricerca è generalmente quello Booleano, integrato con la possibilità di ricerca per frasi.

Spesso le liste dei termini vengono valutate in maniera tale da porre per primi nel ranking i documenti che contengono tutti i termini della query e via via quelli che ne contengono un numero minore.

I risultati vengono tipicamente presentati come ordinati in funzione della rilevanza alla query anche se spesso i motori di ricerca consentono l'utilizzo di altre funzioni di ordinamento. In generale, la funzione di rilevanza utilizzata per l'ordinamento dei risultati è la componente cruciale dal punto di vista della qualità percepita durante l'utilizzo di un motore di ricerca.

Un'aspetto peculiare dell'IR su web è rappresentato dall'impossibilità di misurare il recall. Questo perché il numero delle pagine rilevanti può essere molto elevato anche per query molto semplici.

Links

La differenza principale tra IR tradizionale e IR su web riguarda la presenza di web link, i quali rappresentano una relazione tra pagine connesse.

In seguito vengono presentate alcune definizioni utili legate ad i web link.

- Documento sorgente: la pagina web che contiene il link
- Testo ancora: il testo associato al link visibile all'utente durante la navigazione
- Documento target: la pagina web riferita dal link
- In-link della pagina p: link da una pagina web alla pagina p
- Out-link della pagina p: link dalla pagina p a una pagina web

Nell'ambito dell'IR su web si compiono le seguenti assunzioni riguardo alla presenza e all'utilizzo di link:

1. l'inserimento di un link in una pagina è un suggerimento dell'autore di approfondire l'argomento visionario il documento target;
2. se due pagine sono connesse da link è più probabile che trattino gli stessi argomenti di pagine che non sono legate;
3. un testo ancora descrive il documento target.

A partire da queste assunzioni alcuni IRS utilizzano i link che puntano ad una pagina come una misura della sua popolarità.

Modelli di ranking e popolarità

Esistono due principali tipologie di criteri di ranking:

- basati sul contenuto delle pagine, tra i quali i più diffusi sono il modello booleano e quello vettoriale;

- basati sull’analisi dei link, tra cui PageRank e HITS, i quali agiscono determinando la qualità delle pagine web;
- esistono anche criteri che combinano i due criteri precedenti.

Il ranking di una pagina è una combinazione di rilevanza tematica e grado di popolarità.

La popolarità di una pagina web (popularity o authority) è una funzione di due valori: il suo grado di in-link e di out-link.

Esistono due diverse metodologie di valutazione della popolarità di una pagina:

- indipendenti dalla query, operano un’analisi globale, simulando un attraversamento casuale del web e calcolando il grado di probabilità di raggiungere la pagina (es. PageRank)
- dipendenti dalla query, operano un’analisi locale, focalizzando query generiche che reperiscono troppe pagine e identificando tra queste le pagine autorevoli (es. HITS).

PageRank

Come accennato nel paragrafo precedente, l’algoritmo PageRank simula la navigazione casuale di un utente sul web. Una pagina ha un PageRank alto se la somma dei PageRank dei suoi in-link è alta. Per questa ragione, una pagina avrà una PageRank particolarmente alto se questa possiede molti in-link oppure se ha pochi in-link ma questi hanno un PageRank alto.

Il PageRank di una pagina è la probabilità che l’utente sia su quella pagina in un dato istante. Viene calcolato utilizzando una formula ricorsiva: inizia con un qualunque insieme di valori e itera fino a convergere

$$PR(a) = K_d + K(1-d)\sum_{i=1,n}\frac{PR(a_i)}{C(a_i)}$$

dove K è un fattore di normalizzazione, d dipende dal sistema, a è la pagina puntata dalle pagine a_i con $i \in \{1, \dots, n\}$ e $C(a_i)$ rappresenta il numero di outlink della pagina a_i.

PageRank è un processo modellato da catene di Markov, in cui le pagine web rappresentano gli stati. Inizialmente l’utente si troverà su una pagina p; ad ogni passo procede

1. o verso una pagina web casualmente selezionata con probabilità d, detta fattore di damping
2. o verso una pagina legata da un link alla pagina corrente con probabilità $1 - d$.

L'attività di retrieval di sistemi che utilizzano questo algoritmo sarà quindi basata su una combinazione di rilevanza tematica (RSV) e valore di PageRank.

HITS

HITS, Hypertext Induced Topic Search o Connectivity Analysis Approach, nasce dalla frustrazione dovuta

all'inefficacia di query generiche. Query generiche tendono infatti a reperire moltissimi documenti, rendendo necessaria la definizione di nuove misure di qualità che consentano di distinguere le pagine più autorevoli.

La soluzione proposta da HITS è quella di reperire le pagine più autorevoli in risposta a query generiche. Questo fine fatto identificando, per alcuni argomenti:

- pagine autorevoli che contengono informazioni rilevanti, ossia sono buone sorgenti di contenuto;
- pagine hub che puntano a pagine utili, ossia che sono buone sorgenti di link.

L'intuizione alla base di questa sistema è che l'autorevolezza di una pagina non dipenda solamente dal numero di in-link, poiché anche le pagine popolari hanno questa caratteristica.

Le migliori pagine autorevoli hanno in-link provenienti da buone pagine hub; le migliori pagine hub hanno out-link a buone pagine autorevoli. In altri termini, viene modellato un mutuo rafforzamento tra pagine hub e pagine autorevoli.

L'algoritmo alla base di HITS si divide in tre fasi fondamentali:

1. inizializza l'insieme delle pagine R_q che sono reperite da una query;
2. espandi l'insieme delle pagine con l'insieme B_q, ossia l'insieme delle pagine che puntano o sono puntate dalle pagine in R_q;
3. classifica le pagine in B_q come autorevoli, A_q, o hub, H_q.

Il calcolo dei gradi di autorità e di hub di una pagina, rispettivamente $A(p)$ e $H(p)$, viene effettuato mediante una propagazione iterativa:

1. inizializza $A(p)$ e $H(p)$;
2. ripeti fino al verificarsi della condizione di stop;
    1. calcola il peso di autorità $A(p)$, se p è puntata da molte pagine con valori alti di H, allora deve ricevere un valore alto di A,
    2. calcola i peso di hub di $H(p)$, se p punta molte pagine con valori alti di A, allora deve ricercare un valore alto di H;
3. normalizza i pesi calcolati.

Meta motori di ricerca

I meta motori di ricerca sono delle interfacce che permettono di inviare contemporaneamente la stessa query a più motori di ricerca. A seguito di un'interrogazione, i risultati prodotti da ognuno di questi vengono fusi all'interno di un'unica lista ordina da presentare all'utente. L'assunzione alle base di questo tipo di motori di ricerca è che la ricerca da parte di più sistemi possa trovare informazioni migliori di quelle che potrebbe trovare un sistema singolarmente. Ovviamente, l'efficacia di un meta motore di ricerca dipende fondamentalmente dall'efficacia del suo criterio di fusione delle liste prodotte dai singoli motori di ricerca.