Chi cerca trova

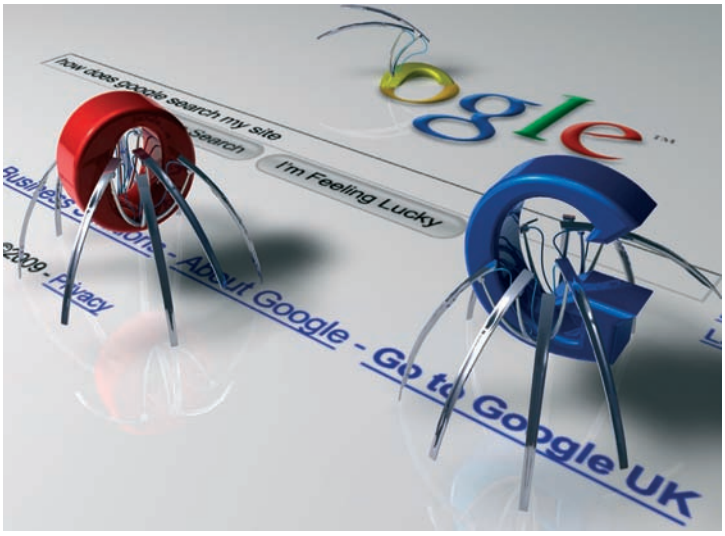
Evoluzioni dei motori di ricerca

di Federico Calzolari

Che cos'è il "dadaismo"? Come si cucina il risotto ai funghi? Di cosa tratta il film "Mon oncle d'Amérique" di Alain Resnais? Quasi tutti oggi, per trovare una risposta a queste domande, andrebbero semplicemente a cercarla in internet, in un motore di ricerca (come ad esempio Google), invece di cercarla nella propria libreria o rimanere senza una risposta.
È da quando, agli inizi degli anni '90, il World Wide Web ha iniziato a ospitare qualche milione di pagine, che è nata la necessità di provvedere alla realizzazione di un sistema per rendere pagine e documenti facilmente reperibili agli utenti finali. Altrimenti occorrerebbe conoscere con precisione l'indirizzo fisico del server e della pagina dove il documento è salvato.
È per questo che sono nati i primi indici, sotto forma di *directory*. Una *directory* è semplicemente un catalogo che, classificando le pagine web per argomenti trattati in categorie e sottocategorie tematiche, struttura lo scibile umano in una sorta di albero ramificato a struttura gerarchica. La categorizzazione in *directory* risulta essere uno strumento particolarmente utile per guidare l'utente nelle proprie ricerche. Partendo da una categoria, per passi successivi, l'utente può perfezionare e affinare la propria ricerca, dal momento che la specificità degli argomenti indicizzati aumenta all'aumentare della profondità di scansione della struttura.
È a metà degli anni '90 che nascono i primi motori di ricerca testuali: Yahoo!, Altavista, Google, dove è sufficiente digitare una parola, un pezzo di una frase, per trovare – con un'approssimazione dipendente dalla precisione della ricerca effettuata – le pagine che le contengono. È però solo dai primi anni del nuovo millennio che gli utenti si sono spostati in massa dall'utilizzo delle *directory* verso i nuovi motori di ricerca testuali.

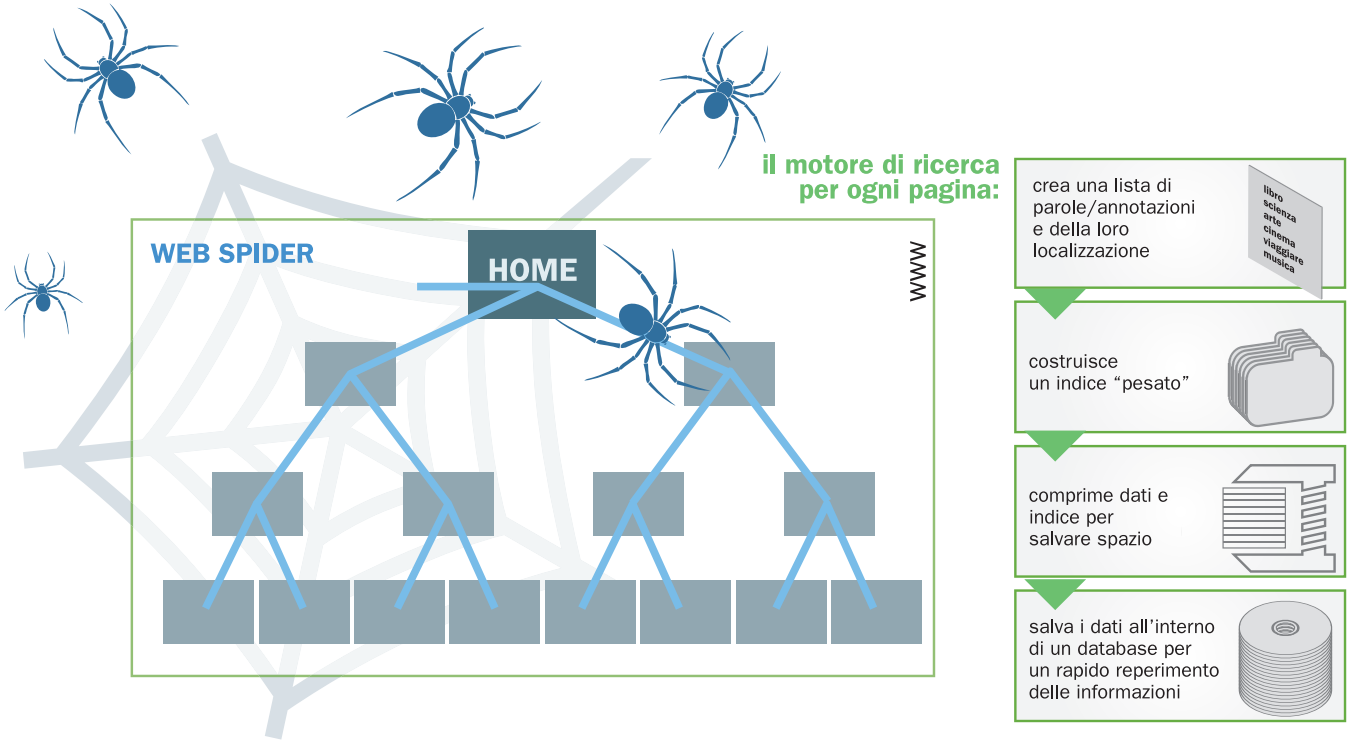
a.
Un motore di ricerca è una buona alternativa a passare ore a cercare risposte alle proprie curiosità in una biblioteca o in libreria.

Il lavoro di un motore di ricerca si può in linea di massima suddividere in tre fasi: scansione del web, analisi e catalogazione dei dati, risposta alle richieste degli utenti.
La fase di scansione di tutte le pagine raggiungibili in rete è un'operazione oggi relativamente facile da eseguire attraverso l'utilizzo di *spider* (anche chiamati robot o *crawler*). Uno *spider* è un sistema completamente automatizzato capace di scandagliare in profondità tutto il web seguendo i link presenti sulle pagine da cui transita, salvando in locale le pagine visitate e tenendo traccia dei puntatori alle pagine stesse.
Estremamente più complessa, tanto da essere al centro di casi di spionaggio industriale, è la fase di analisi e indicizzazione, dove sistemi derivati dalla linguistica computazionale catalogano le pagine per contenuti e per parole chiave, organizzandole e indicizzandole all'interno di un database di dimensioni gigantesche. La risposta alle richieste degli utenti, in termini di tempo impiegato e precisione, è quello che fa, alla fine, la differenza tra un motore di ricerca e un altro. Comprensibile quindi il perché di tanta segretezza in tutto quel sistema di hardware, software e algoritmi che presiedono al sistema di archiviazione, indicizzazione e recupero delle informazioni.
Nel 1998 nasce Google. Il nome deriva dal termine Googol, che in matematica indica un numero enorme, esprimibile con un 1 seguito da 100 zeri. Ma dato che al momento



b.
La scansione di tutte le pagine raggiungibili in rete è un'operazione realizzabile attraverso l'utilizzo di uno *spider*.

c.
Schema di funzionamento di un motore di ricerca.



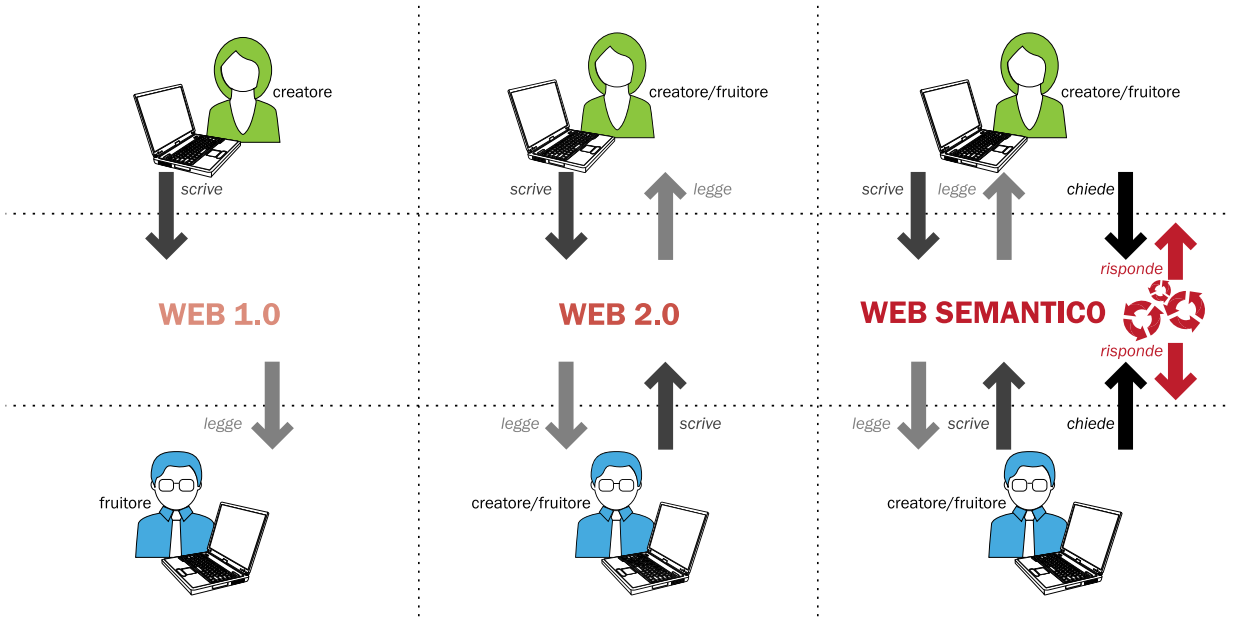
d.
Sergey Brin (in alto) e Larry Page (sotto), fondatori di Google, assieme a Eric Schmidt (sulla loro sinistra), terzo uomo del Consiglio di Amministrazione della società dal 2001.

della registrazione il dominio era già assegnato, Larry Page e Sergey Brin, i fondatori del motore di ricerca, ripiegarono su quello che di lì a pochi anni avrebbe segnato una rivoluzione nell'ambito del web: Google. Oggi è il motore di ricerca più utilizzato, con una percentuale di affezionati superiore al 65% a livello mondiale e con picchi che superano il 90% in molti stati – in Italia, ad esempio, Google viene utilizzato dal 93% degli utenti. Ad oggi Google indicizza circa 50 miliardi di pagine, su un totale stimato di circa 80 miliardi di pagine ospitate in tutto il web.
Cos'è che ha portato Google a un predominio pressoché assoluto del mercato? Il *PageRank*. Si tratta di un algoritmo, inventato dal matematico italiano Massimo Marchiori per un suo prototipo di motore di ricerca e poi adottato da Page e Brin, capace di assegnare un peso a una determinata pagina in funzione della rilevanza della pagina stessa in relazione ai termini cercati. In questo modo, le pagine frutto di una specifica ricerca non hanno più tutte lo stesso peso, ma sono classificate in funzione della pertinenza con quanto cercato dall'utente. Sono circa 200 i fattori che contribuiscono al peso di una specifica pagina e al suo posizionamento nel motore di ricerca; in primis, il grado di popolarità del sito che la ospita – determinato tendenzialmente dal numero di link che la puntano, a loro volta pesati dal livello di popolarità del sito di origine.
Una delle ultime sfide lanciata nell'ultimo anno da Google è la ricerca per immagini, un'alternativa – in alcuni casi indispensabile – alla ricerca testuale. L'utente può iniziare la sua ricerca da un'immagine che ha sul proprio computer, andando a scandagliare la rete, attraverso un motore di ricerca *ad hoc*, in cerca di immagini simili. Quindi, non è più strettamente necessario partire da una parola, ma è sufficiente un'immagine, una foto, magari scattata con il proprio cellulare, per sapere il nome del monumento davanti al quale ci si trova, e scoprire magari che, anche se illuminata di verde in onore del *Saint Patrick's Day* (il 17 marzo, ndr), si tratta della Torre di Pisa. Tale sistema, integrato con un qualunque dispositivo Gps per la geolocalizzazione (oggi disponibile su praticamente qualunque dispositivo mobile, dal cellulare alla macchina fotografica), attraverso l'integrazione di informazioni testuali, visive e geografiche, apre scenari fino a ieri inimmaginabili: da informazioni dettagliate sul luogo e sui dintorni di dove ci troviamo, a possibili interazioni con altri utenti nella stessa area geografica.
Sebbene il *PageRank* abbia rivoluzionato il funzionamento dei motori di ricerca, l'interesse da parte di questi ultimi è oggi indirizzato alla *profilazione* degli utenti (una classificazione dell'utenza in funzione dei dati raccolti), al fine di rendere il risultato delle ricerche pressoché *ad personam*, sia dal punto di vista dei risultati, sia dal punto di vista della pubblicità associata. È infatti la pubblicità, oramai inscindibilmente legata ad ogni pagina che riporti i risultati di una ricerca, la prima e spesso unica fonte di guadagno per un motore di ricerca. Pubblicità mirata in funzione delle ricerche di ogni singolo utente: è questa oggi la base dell'economia del web 2.0, quello fatto non solo di web e motori di ricerca, ma anche e soprattutto di social network. Sarà probabilmente dall'unione di motori di ricerca e social network, in modi ad oggi probabilmente ancora inimmaginabili ai più, che nasceranno un domani prodotti più indirizzati alle esigenze di ogni singolo utente. Qualcuno già chiama questa terza rivoluzione della rete *web 3.0*.

[as] approfondimento

Dal cercare al chiedere

1.
Evoluzione dal web 1.0 al web semantico. L'interazione con la rete e i motori di ricerca diventa sempre più "umana".



Ancora a livello di chimera è il *web semantico* (termine coniato dallo stesso Tim Berners-Lee), di cui si parla oramai da più di un decennio. La maggior parte dei motori di ricerca già oggi è capace di catalogare in modo totalmente automatico le pagine per contenuti tematici e comprendere più o meno correttamente una richiesta effettuata in linguaggio umano. Resta tuttavia cosa ancora pressoché impossibile da realizzare con le tecnologie attuali la possibilità per l'utente di effettuare ricerche per significato, un qualcosa che vada oltre le parole scritte nel testo. Questo è l'obiettivo ultimo del web semantico.
Una concreta interazione tra motori di ricerca e web semantico potrebbe essere costituita da ricerche del tipo "qualcosa di simile a" o "qualcosa che abbia a che fare con", dove la ricerca della singola parola nel testo non può soddisfare i criteri di ricerca. La ricerca nel web per parole chiave può ovviamente produrre falsi positivi per l'utente a causa di *polisemie* (alla stessa parola corrispondono più significati). Allo stesso tempo possono esistere falsi negativi, a causa di testi che trattano l'argomento richiesto ma utilizzando sinonimi delle parole chiave cercate. Il web semantico si propone di risolvere questi problemi attraverso l'utilizzo di strutture dati che contengano oltre alle parole di un testo anche informazioni, sotto forma di metadati, che ne specifichino il contesto semantico. Manca però ancora la capacità di collegare i vari termini tra loro in funzione di relazioni che intercorrano tra gli stessi (compito della *ontologia*).
Una tale ristrutturazione dei dati, con l'utilizzo di ontologie, renderebbe il significato di un testo accessibile non solo a un umano, ma anche a una macchina. Questo potrebbe portare un motore di ricerca a capire, ad esempio, se il termine "vodka" indichi un liquore o la marca di un tostapane, e al tempo stesso trovare in rete pagine e documenti che trattino non solo di "vodka", ma anche di altri liquori o alternative per un piatto flambé.

Biografia
Federico Calzolari, esperto di Grid computing, lavora presso la Scuola Normale Superiore di Pisa e collabora con l'Infn e il Cern di Ginevra. È stato riconosciuto da Google e accreditato da stampa e media come autore del primo exploit degli algoritmi di ranking di Google ed elencato dal Corriere della Sera nel 2009 tra i "20 Italiani che cambiano l'Italia".

Link sul web

www.google.com

www.yahoo.com

http://www.segnalidivita.com/motoridiricerca/