INGEGNERI DI BERGAMO N° 146 GEN/FEB 2000

Riceviamo e Pubblichiamo

MOTORI DI RICERCA

di Francesco Lanorte

Navigare in Internet è divertente, ma dopo il primo entusiasmo la maggior parte dei cybernauti perde interesse nel girovagare puro e semplice e cerca invece di sfruttare il prezioso strumento che ha a disposizione per reperire informazioni utili. E' un pò come entrare in un'enorme biblioteca: si sa che si possono trovare informazioni su QUALSIASI argomento, ma non si sa bene da dove cominciare.

Per questo motivo sono nati spontaneamente i cosiddetti **motori di ricerca**: i motori di ricerca sono la risorsa principale a disposizione dell'utente per la ricerca di informazioni su Internet. In chiave generale sono dei grandi archivi di dati che contengono informazioni dettagliate su un gran numero di pagine web. Sono alimentati da agenti software (gli spider, i crawler, gli scooter), che navigano sulla rete alla ricerca di pagine web. L'inserimento delle pagine web negli archivi dei motori di ricerca può avvenire in due modi: sia attraverso la registrazione manuale da parte dell'utente (ad esempio il responsabile del sito, comunemente detto webmaster), sia in modo automatico attraverso un particolare software che riesce a visitare milioni di siti web al giorno, inserendo le nuove pagine ed aggiornando le informazioni già censite. Attraverso tali funzioni, i motori di ricerca mantengono un archivio piuttosto aggiornato, anche se non sarà mai possibile classificare l'intero Web.

L'uso dei motori di ricerca è totalmente gratuito, così come la registrazione delle pagine negli archivi, nonostante i massicci investimenti necessari per offrire un simile servizio. La ragione sta nel fatto che nelle pagine presentate all'utente, compaiono degli annunci pubblicitari (i cosiddetti banner) attraverso i quali le società che gestiscono i motori di ricerca traggono i loro ricavi.

Esistono diverse tipologie di motori:

- I **search engine**
- Le **directories**
- I **metacrawler**
- I motori cosiddetti **parassiti**
- I motori **generalisti** che si distinguono dai motori **tematici**.

I **search engine**, come Altavista, Lycos ed Infoseek, censiscono le pagine di un sito attraverso due sistemi: a) add url, che permette ad ogni utente di segnalare la singola pagina del sito al motore di ricerca, specificandone l'indirizzo web; b) spider o crawler, software che scandagliano il web e acquisiscono tutte le pagine non ancora presenti negli archivi. I siti vengono censiti in base alla rilevanza delle parole contenute in ogni pagina evidenziando quelle più spesso ripetute, che si presume rappresentino l'argomento principale della pagina stessa.

Le **directories** sono dei siti strutturati per categorie ed argomenti organizzati ad albero (ad esempio Virgilio e Yahoo), dove viene riportato un unico riferimento per ogni sito, generalmente corrispondente alla home page.

Queste due categorie vengono spesso confuse tra loro. Occorre pertanto precisare che il motore di ricerca è un sistema utilizzato per rintracciare automaticamente le pagine web che contengono le parole chiave specificate dall'utente. La directory è invece un catalogo di siti suddiviso in categorie che l'utente consulta come un indice, alla ricerca del sito più appropriato.

I **metacrawler** sono siti che operano su più motori di ricerca contemporaneamente:
Net Sonar ad esempio cerca su Altavista, Lycos, Excite, Webcrawler e Yahoo.

I motori **parassiti** non sono dotati di un archivio proprio, forniscono risultati indicativi, utilizzano gli archivi di altri motori. Un esempio è Multiopac, servizio offerto dal CISI, che permette di consultare contemporaneamente più cataloghi bibliotecari diffusi in tutto il mondo.

I motori **generalisti** permettono di svolgere ricerche relative a tematiche differenziate, quelli tematici sono dotati di archivi circoscritti. Altavista (www.altavista.com) è un motore generalista. Amnesi è un esempio di motore tematico (aiuta l'utente a trovare il sito di cui non ricorda l'indirizzo esatto).

Se si vogliono reperire informazioni nel World Wide Web, occorre adottare opportune strategie di interrogazione, per non rischiare di ricevere un numero eccessivo di informazioni, talvolta inutili. Ci sono varie forme di ricerca che possiamo fare sulla rete e per ognuna di queste troveremo una risposta utilizzando un tipo di motore di ricerca differente.

Possiamo impostare il nostro percorso di ricerca in diversi modi:

per argomento generico:

se ad esempio si desidera trovare tutti i siti in cui si parla di musica. Fanno parte di questo tipo di motori di ricerca, denominati in genere directory, Yahoo e Virgilio. Su queste directory sono anche presenti degli spazi monografici e dei canali tematici, spesso corredati da servizi accessori (mappe, itinerari, ecc.).

con ricerche mirate:

le ricerche mirate permettono di individuare tipologie di documenti specifici, come ad esempio file grafici o di testo. Circoscrivere analiticamente la ricerca usando più di un termine, oppure estendere la ricerca con più motori è un primo passo da compiere.

Le ricerche sulla rete tendono ad essere sempre più complesse: l'aumento vertiginoso della quantità di informazioni e dei documenti disponibili, rende necessaria la nascita di strumenti di ricerca specifici. Esistono dei software specializzati che fungono da agente intelligente che analizza la rete non solo in base a parole chiave, ma a concetti sempre più perfezionati. Vi sono anche alcune aziende alle quali è possibile commissionare delle ricerche: la più nota si chiama HumanSearch.

Non sempre è possibile e conveniente commissionare ricerche, per cui dobbiamo tenere presenti alcune regole che ci permettono di trovare, con economia di tempo, le informazioni che stiamo cercando. Quando le normali interrogazioni dei motori di ricerca danno risultati troppo vaghi, quindi poco utili, c'è soltanto una possibilità: ridurre il campo d'indagine, imponendo condizioni più complesse e restrittive. Solitamente ci si collega al sito, si inserisce la parola desiderata e in breve compare l'elenco di tutte le pagine presenti nell'indice che contengono il termine introdotto. Spesso però il numero di documenti che rispettano la condizione impostata è troppo elevato e l'entusiasmo iniziale si tramuta in frustrazione: avere 10.000 risposte è quasi come non averne nessuna. Come fare ? Supponiamo di effettuare una ricerca su Altavista per ottenere informazioni riguardanti l'ingegneria civile. Utilizzando come motore di ricerca Altavista e digitando alternativamente le parole ingegneria e civile si ottengono rispettivamente 127.735 e 172.635 documenti. Si tratta di quantità che risultano troppo elevate per essere utili. Possiamo effettuare una prima scrematura limitando la ricerca soltanto ai documenti scritti in italiano. Così facendo, il numero di risposte scende rispettivamente a 114.180 (ingegneria) e 118.205 (civile), ma è ancora troppo elevato. Bisognerebbe poter richiedere i documenti che contengono entrambi i termini: si può esprimere questa condizione separandoli con il simbolo + (ingegneria+civile = circa 4432 documenti di risposta). Se non ci importa nulla dei trasporti, possiamo scartare i file che contengono questa parola inserendola preceduta dal simbolo – e da uno spazio (ingegneria+civile –trasporti = circa 3690 documenti di risposta). Se poi desideriamo consultare solo quelle pagine che sono state modificate nell'ultimo mese dovremo inserire la condizione nell'apposita mascherina, ottenendo solo 6 documenti di risposta. Più specifica è la richiesta, più è facile che il risultato comprenda solo documenti significativi. Racchiudendo tra virgolette due o più parole ("Ingegneria dei Trasporti"), si restringe il campo di ricerca ai documenti che contengono quella esatta sequenza di caratteri.

Spesso gli agenti software che navigano su Internet alla ricerca di nuove pagine, si limitano a rintracciare le parole chiave in qualsiasi documento appaia, anche il meno rilevante. L'ordine in cui sono elencati i risultati, inoltre, è in genere determinato solo dal numero di volte in cui una parola chiave compare: un criterio solo quantitativo, inadeguato a soddisfare le nostre richieste.

Tuttavia i webmaster, per rendere più facilmente intercettabili dai motori le loro pagine, possono allegare dei "campi invisibili" a ciascuna pagina web: il linguaggio HTML può contenere dei comandi (chiamati meta tag) mediante i quali il webmaster specifica il titolo, il sommario e le parole chiave che meglio descrivono la pagina in questione. I meta tag non sono visibili ma vengono letti e indicizzati dai motori di ricerca. Per vederli, si deve visualizzare il codice sorgente della pagina. Le pagine che contengono i meta tag sono più facilmente reperibili e questo aiuta sia chi effettua le ricerche sia l'autore. Spesso le parole chiave che vengono utilizzate da chi cerca per specificare l'argomento cui è interessato, non sono presenti in una pagina. Usando un plurale anziché un singolare, un sinonimo o un termine vago, si rischia, in assenza di meta tag, di non trovare pagine preziose per la propria ricerca. Ma non sempre i meta tag si rivelano utili. Sempre più spesso capita che facciano la loro comparsa fra i risultati pagine che, oltre a non contemplare le parole chiave, non sembrano avere alcun nesso con quanto cercato. Ciò è da attribuire ai cosiddetti spammer, coloro che sfruttano il web a scopo di lucro senza rispettare la netiquette (il codice deontologico cui un buon navigatore dovrebbe attenersi). Una delle strategie più utilizzate dagli spammer consiste nell'attirare visitatori inserendo nei tag parole chiave di uso comune, come "download", "software", "free". Gli spammer inseriscono le parole chiave ingannatrici nel testo oltre che nei meta tag, ma esse sono ugualmente invisibili perché scritte nel medesimo colore della pagina. L'unico modo per scoprirle è selezionare tutto il testo della pagina.

I motori di ricerca sono quindi lo strumento fondamentale per la promozione e la ricerca di un sito web. Non a caso, tra i primi dieci siti più visitati in assoluto su Internet, sei sono motori di ricerca, con medie che superano anche i dieci milioni di accessi al giorno. Per fare un esempio Yahoo! (Indirizzo web per l'Italia www.yahoo.it) oltre ad essere il motore di ricerca più utilizzato è attualmente uno dei siti web più frequentati in assoluto, con oltre 300 milioni di pagine visitate nel mondo al giorno (dati aggiornati al settembre 1999). Nato nel 1993 dall'idea di due studenti dell'università di Stanford, Jerry Yang e David Filo, Yahoo!, che impiega oltre 600 persone, prevede di fatturare 190 milioni di dollari nel 1999, dei quali l'80% deriva dalla vendita di spazi pubblicitari.