# QUANDO IL SITO È INVISIBILE

**INTERNET SI ESPANDE COSÌ VELOCEMENTE CHE ANCHE I MIGLIORI MOTORI DI RICERCA RIESCONO A TROVARE SOLTANTO UN TERZO DELLE PAGINE. E LE ALTRE? PER SCOVARLE BISOGNA AVERE PAZIENZA E CONOSCERE QUALCHE INDIRIZZO SEGRETO. CHE «PANORAMA WEB» VI SVELA.**

di Fabio Metitieri
*(yukali@tin.it)*

**Chi esegue una ricerca su Google o Altavista**, e a volte addirittura chi usa Virgilio, è convinto di scandagliare tutto il Web. Niente di più falso. Forse non è più necessario ripetere che Virgilio non è un motore ma una directory, una ristretta selezione di siti scelti, classificati e recensiti da un team di redattori; è però senz'altro utile dare un'occhiata alle reali possibilità dei mitizzati Altavista e Google, e soprattutto a quello che i motori non riescono a trovare.

Il primo limite dei search engine è dato dalle dimensioni di Internet e dalla sua crescita sempre più rapida. Di ricerche attendibili e complete sul numero delle pagine e soprattutto sulla forma del web non se ne vedono più dal 2000, perché persino queste indagini ormai sono troppo lunghe, costose e difficili, e produrrebbero dei risultati destinati a invecchiare molto in fretta. Secondo le statistiche più recenti di Oclc, un'organizzazione non profit che si occupa di catalogazione e di archivi, i siti sono passati da 8.745.000 nel 2001 a più di 9 milioni nel 2002 (wcp.oclc.org/stats/size.html).

E i motori, a fronte di questo mare magnum di documenti, che cosa riescono a fare? Secondo l'osservatorio Search engine showdown di Greg R. Notess (www.searchengineshowdown.com/stats/sizeest.shtml), il popolarissimo Google (www.google.com) arriva a indicizzare 3 miliardi di pagine, mentre molti altri, come Hotbot o Msn search, superano di poco il miliardo.

Altrettanto severe sono le valutazioni del nostrano Motoridiricerca.it (www.motoridiricerca.it/tabella.htm), che riporta stime un po' più vecchie, di agosto del 2002. Il tanto apprezzato Google, insomma, arriva a vedere circa un terzo delle pagine esistenti, mentre molti dei suoi più grandi concorrenti arrivano a stento al 13 per cento del totale.

Le ragioni di queste basse percentuali sono molte. La prima, ovvia, è che gli spider dei motori, cioè i programmi che navigano per raccogliere le pagine da indicizzare, faticano a percorrere i miliardi di pagine del Web in tempi brevi. Ancora le analisi di Notess dicono che il tempo medio di aggiornamento di tutto l'archivio è di 3 settimane per Alltheweb, un mese per Google, e addirittura 4 o 5 mesi per Teoma e Wisenut. Incrementare il numero dei siti

**RISULTATI IN VENDITA**

La ricerca, in alcuni casi, è falsata dal «pay for placement», un termine usato per le diverse forme con cui i gestori dei motori vendono ai siti commerciali l'inserimento in archivio (pay for inclusion) delle loro pagine, o una loro indicizzazione più frequente (pay for indexing), o una loro buona posizione tra i risultati di una o più parole chiave (pay for ranking). Queste pratiche hanno già provocato alcune denunce e un richiamo formale ai gestori dei principali motori da parte della Federal trade commission statunitense. La situazione oggi è ancora poco chiara; di certo alcuni motori, come da sempre Google e da qualche mese Altavista, assicurano che tutti i risultati sponsorizzati sono presentati come tali, separati dagli altri, mentre qualcun altro continua ad avere comportamenti un po' ambigui. Looksmart (www.looksmart.com), per esempio, è stato contestato perché presentava i risultati divisi nelle sezioni «Featured listing», «Directory topics» e «Reviewed Web sites», dove solo la seconda lista, la Directory topics, rappresentava un lavoro privo di aggiustamenti per le aziende paganti. Adesso il nome del primo elenco è stato cambiato in «Sponsored listings», mentre per i Reviewed Web sites l'ambiguità resta. Anche le scelte di Altavista sono in forse, dato che nel prossimo aprile verrà acquistato da Overture (www.overture.com) uno dei search engine più dedicato al «pay for». Quello che è noto delle politiche di ranking adottate dai motori, in continua evoluzione, è pubblicato in diversi articoli sul Search engine watch (www.searchenginewatch.com).

## Ecco chi ne trova di più

| motore | indirizzo | pagine dichiarate (in milioni) | pagine riscontrate (in milioni) |
|---|---|---|---|
| Google | www.google.com | 3.083 | 3.033 |
| Alltheweb (o Fast) | www.alltheweb.com | 2.112 | 2.106 |
| Altavista | www.altavista.com | 1.000 | 1.689 |
| Wisenut | www.wisenut.com | 1.500 | 1.453 |
| Hotbot (usa Inktomi) | www.hotbot.com | 3.000 | 1.147 |
| Msn search | www.msnsearch.com | 3.000 | 1.018 |
| Teoma | www.teoma.com | 500 | 1.015 |

**PAGINE DA DICHIARARE?**
**La tabella riassume le analisi sui principali motori di ricerca effettuate dall'osservatorio di Greg R. Notess, famoso consulente americano sui temi della Rete. Si evidenzia una certa discrepanza fra il numero di pagine indicizzate dichiarato dai singoli motori e quelle effettivamente riscontrate.**

da visitare aumenterebbe ancora questi tempi, già troppo lunghi per molte ricerche.
Un secondo motivo è la scelta deliberata di «scremare» i risultati ottenuti dagli spider per eliminare le pagine ritenute poco importanti. In realtà, gli utenti dei motori, come ha evidenziato uno studio recente di Altavista, sono molto pigri, guardano solo i primi 10 o 20 risultati ottenuti dalla ricerca e non usano quasi mai più termini di ricerca o gli operatori booleani (AND, OR e NOT). Motori come per esempio Altavista e Inktomi, quindi, cercano di mantenere i propri archivi più «snelli» di quelli di Google, per superarlo con la qualità dei primi link elencati in ciascuna ricerca.
È la dura lotta per il «ranking», l'ordinamento migliore dei risultati, che spesso riduce le dimensioni degli archivi dei motori. Secondo alcuni è meglio non indicizzare tutto, ma scegliere le risorse più interessanti e passarle più di frequente. Così Altavista prevede aggiornamenti settimanali per i siti ritenuti più importanti e quotidiani per molti dei siti più popolari. Per alcune pagine non profit, editoriali o istituzionali, i suoi spider girano quattro volte al giorno e per la ricerca delle notizie vi è un aggiornamento ogni quarto d'ora su 2.700 fonti prescelte, da tutto il mondo. Inktomi, appena acquistato da Yahoo!, preferisce invece affiancare a un indice generale, grande ma molto filtrato, degli archivi più piccoli divisi per area geografica.
Ma alcuni siti Web sono di fatto irraggiungibili. Gli spider navigano seguendo i link presenti nelle pagine e passando da una all'altra; una ricerca condotta da Compaq, Ibm e Altavista già nel 2000 (non ve ne sono di più recenti) aveva evidenziato la presenza in Rete di molte pagine non collegate alle altre e dunque difficilmente raggiungibili. Il webspazio assomiglierebbe a un cravattino a farfalla. Il nodo rappresenta i siti molto connessi (il 30 per cento del totale), la parte sinistra della farfalla un insieme di pagine «di origine» (24 per cento), la parte destra un insieme di pagine «di arrivo» (24 per cento) e i lacci pagine «disconnesse» (22 per cento). Il nucleo centrale può essere navigato con facilità, grazie a un grande numero di collegamenti. L'ala sinistra contiene invece pagine che permettono di raggiungere il nucleo centrale ma che non sono raggiungibili da esso. Al contrario, l'ala destra del cravattino può essere raggiunta facilmente dalla parte centrale ma non ha molti link che riportino a essa. Le pagine disconnesse, infine, sono tagliate fuori dal nucleo centrale e sono collegate solo tra loro, in quella che potrebbe essere definita la periferia del Web, sconosciuta ai motori.
Ma non è finita qui. Anche se una pagina può essere raggiunta, non sempre può essere indicizzata correttamente. Esiste un «Web invisi-

**I BREVETTI**
**Introvabili con i motori, i brevetti sono raccolti in archivi specializzati ad accesso gratuito. Brevetti di tutto il mondo (30 milioni) si trovano su Esp@cenet (it.espacenet.com), per iniziativa dell'Ufficio europeo brevetti; quelli statunitensi sono invece all'indirizzo www.uspto.gov/patft/index.html. Per ridere sulla possibile assurdità delle registrazioni, c'è invece il sito Patently absurd (www.around.com/patent.html).**

I LINK SEGRETI A BIBLIOTECHE E OPERE D'ARTE

## Tra volumi e musei virtuali

Tra le risorse non accessibili con i motori di ricerca vi sono senza dubbio tutte le schede sui libri posseduti dalle biblioteche, i cui cataloghi (Opac) sono ormai largamente disponibili in Rete. Importantissimo e molto ricco (conta circa 5,5 milioni di titoli) è il servizio bibliotecario nazionale (Sbn), con indirizzo http://opac.sbn.it. Due buone liste di Opac (italiani e internazionali) si trovano sul sito dell'Associazione italiana biblioteche (www.aib.it); sempre in italiano si può leggere *Biblioteche in Rete. Istruzioni per l'uso*, di Fabio Metitieri e Riccardo Ridi (Laterza, 2002, 275 pagine, 16 euro). Anche per le opere d'arte la ricerca con i motori produce spesso troppo rumore e pochi indirizzi utili. Meglio consultare Adam (adam.ac.uk), l'Art, design, achitecture and media information gateway, oppure gli archivi specializzati sui musei online, come il sito Musei virtuali internazionali (www.muvi.org), promosso dall'Unione europea, o il Museum register (www.museumregister.com).

bile» o «deep Web» costituito da file in formati non analizzabili dai motori. I file nel formato Pdf di Adobe, per esempio, o le immagini, o i siti che utilizzano funzioni dinamiche, con pagine costruite «al volo» utilizzando i dati estratti da un archivio. Oltre a questi, le animazioni, le amate e odiate pagine in Flash, i file audio e video, e i file in formato compresso (come gli zip) o i programmi eseguibili. E infine i data base, gli archivi di dati, ai quali si può accedere solo usando uno specifico linguaggio di interrogazione, e le molte pagine protette da un meccanismo di iscrizione. Per trovare informazioni su queste risorse, gli spider e gli indexer possono utilizzare le pagine che le collegano e che, non sempre, le descrivono. Per i Pdf un sito dimostrativo è gestito da Adobe (searchpdf.adobe.com), ma le funzioni per la loro lettura e indicizzazione ormai sono presenti anche in Google e in Altavista. Per le immagini il cammino è più difficile. Dal 2000 sono stati implementati dei software in grado di riconoscere il «pattern», l'aspetto, di una foto o di una figura, classificandola e trovando altre immagini con lo stesso aspetto. Per il futuro si attendono i programmi in grado di classificare i file audio e video, ma per ora tutte queste tecniche sono troppo pesanti per essere utilizzate a tappeto dai motori.

In ciò che è Web invisibile cercano diversi motori o directory specializzati. Profusion di Intelliseek (www.profusion.com) per alcuni argomenti, per esempio medicina, mappe o geografia, propone indicazioni su come completare l'indagine all'interno di speciali archivi e risorse scandagliati ad hoc dal motore di Intelliseek. La stessa Intelliseek aveva già realizzato Invisible Web (www.invisibleweb.com), una directory che raccoglie le risorse di circa 10 mila data base inaccessibili ai motori di ricerca tradizionali. L'invisible Web è anche in parte l'obiettivo della directory dei «Reference» realizzata da Lycos (dir.lycos.com/Reference).

GLOSSARIO

## Indici, classifiche o archivi?

◆**Directory:** sono archivi con siti selezionati e commentati da una redazione di esperti, che li classifica per argomento secondo schemi ad albero. In Italia la più nota è Virgilio; a livello internazionale Yahoo!. Sono molto più piccole dei motori: una delle più grandi, la Open directory (www.dmoz.org), classifica 3,8 milioni di siti (per le directory si parla di siti e non di pagine) La directory di Virgilio considera invece circa 100 mila siti italiani.

◆**Motori di ricerca:** i grandi motori come Altavista e Google hanno dei programmi chiamati spider o crawler che setacciano la Rete in continuazione, raccogliendo le pagine che trovano. L'archivio viene quindi organizzato in base a parole chiave da programmi chiamati indexer.

◆**Indici Web retrospettivi:** le pagine Web hanno una vita media di 6 settimane; la loro conservazione per i posteri è curata dalla Wayback machine (www.archive.org), che dal 1996 tiene in archivio le vecchie versioni dei siti Web che indicizza, salvati ogni 6 mesi. Un altro progetto di questo tipo è stato avviato nel 2002 dalla British library.

◆**Meta indici:** Permettono di interrogare in un colpo solo più motori o directory. I risultati, ovviamente, sono caratterizzati da molto rumore e da numerosi risultati duplicati. Il più famoso di questi è Metacrawler (www.metacrawler.com).

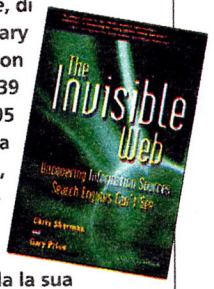
Se quanto si cerca ha qualche attinenza con una lingua o con un Paese, è senz'altro consigliabile sfruttare i motori di ricerca locali, come in Italia Arianna (arianna.libero.it) e Il Trovatore (www.iltrovatore.it). Lo stesso vale per ricerche legate a una disciplina specifica, anche se i motori e le directory specializzati per argomento sono ancora pochi. Un elenco di motori classificati per Paese o per settore si trova sul Search engine watch, all'indirizzo www.searchenginewatch.com/links.

La ricerca su Web con i grandi motori è molto meno perfetta di quanto si possa pensare e per un determinato campo, oltre al passaparola tra ricercatori e appassionati, che ai tempi di Internet funziona ancora molto bene, occorre usare più strumenti e non fidarsi ciecamente soltanto di Google e di Altavista.

**A pag. 81 la nuova rubrica di consigli sull'uso dei motori di ricerca.**

**UN LIBRO PER SAPERNE DI PIÙ**

Il testo più autorevole e completo uscito finora su questi argomenti è *The Invisibile Web. Uncovering information sources search engines can't see*, di Chris Sherman e Gary Price (Information today, 2001, 439 pagine, 29,95 dollari). Uscito a dicembre del 2001, questo manuale risente un po' del tempo passato, ma resta valida la sua parte didattica su come cercare in Rete senza diventare vittime di falsi miti sui motori. La directory di risorse invisibili riportata nel libro è disponibile anche online, gratuitamente, all'indirizzo www.invisible-web.net.




## Attenzione, c'è chi trucca le ricerche

**Si chiama Spamdexing ed è una pratica per ingannare i motori adottata da molti siti commerciali che vogliono risultare sempre in testa alle classifiche.**

Lo Spamdexing è una deprecabile pratica adottata da ormai quasi tutti i webmaster dei siti commerciali; in sostanza, si cerca di ingannare gli indexer dei motori (i programmi che ordinano e indicizzano le pagine raccolte dagli spider) con diversi trucchi in modo da risultare sempre in testa alle liste di risultati, anche in corrispondenza di parole chiave che non rispecchiano il contenuto delle proprie pagine. L'esistenza dello spamdexing è il motivo per cui nessun motore, neppure il trasparentissimo Google, rivela i precisi criteri in base a cui effettua il ranking, criteri che in ogni caso vengono modificati molto spesso.

## Ricerche in Rete

# Chi cerca meglio TROVA prima (e di più)

**I motori sono molto conosciuti ma usati male. Ecco qualche idea per cambiare abitudini.**

di Fabio Metitieri
*(yukali@tin.it)*

**Il problema centrale in qualsiasi ricerca** su un archivio dati o su un search engine è quello di ottimizzare sia il richiamo sia la precisione.
Il richiamo è la percentuale dei documenti pertinenti che si riescono a trovare rispetto al totale di tali documenti che sono in archivio. Se per esempio si ricerca in una banca dati bibliografica che contiene 1.000 schede su altrettanti libri, e se solo 100 di questi libri corrispondono all'argomento voluto, allora una ricerca che ottiene 200 risultati con 80 dei quali pertinenti ha un richiamo dell'80 per cento. La precisione invece è la percentuale dei documenti pertinenti ritrovati rispetto al totale di quelli ottenuti dalla ricerca. In questo esempio, la precisione è del 40 per cento, dato che solo 80 dei 200 titoli trovati sono interessanti. Per i motori che scandagliano Internet la logica è la stessa. Quando si conosce poco l'argomento su cui si eseguono le ricerche, è consigliabile iniziare con interrogazioni più generiche, per poter fare un primo esame di quello che esiste nell'archivio e delle parole chiave associate alle pagine cercate. In un secondo tempo si lanciano ricerche più precise. Al contrario, se l'argomento è già ben conosciuto, si possono utilizzare subito delle parole chiave più mirate, per poi completare la ricerca con altri termini più generici, controllando se si sono perse alcune pagine importanti. Questa «modulazione» delle ricerche può essere fatta scegliendo parole chiave più o meno precise e restrittive, ma gli strumenti più importanti per modificare il richiamo e la precisione sono i cosiddetti operatori booleani, quali AND (per includere due o più parole), OR (per includere una o l'altra) NOT (per escluderne una). La ricerca perfetta, tuttavia, quella con il 100 per cento di precisione e il 100 per cento di richiamo, non esiste quasi mai.
Occorre quindi tentare più volte e soprattutto non leggere frettolosamente solo i primi 10 risultati della lista.

### INDICI DI GRADIMENTO

Zeitgeist, ovvero lo spirito dei tempi, è la rubrica di classifiche nella quale Google raccoglie per argomenti le parole chiave più usate per le ricerche. Un indice interessante può anche essere il numero delle pagine ritrovate in Rete su un determinato argomento. Nella politica italiana, Silvio Berlusconi surclassa tutti, con 123.000 pagine. Per batterlo, a sinistra occorre riunire Massimo D'Alema (40.800), Francesco Rutelli (27.900), Piero Fassino (24.400), Sergio Cofferati (21.500) e Fausto Bertinotti (13.200), che soltanto insieme arrivano a 127.800 pagine. Usando Altavista invece di Google, cambiano i numeri ma i rapporti rimangono quasi invariati: in questo caso la sinistra potrebbe fare a meno solo di Bertinotti.

### NOVITÀ IN VISTA PER I GRANDI

**Il mondo dei motori** è in continuo fermento, soprattutto in questi tempi di crisi. Per primo è arrivato l'annuncio che Yahoo! (www.yahoo.com) ha acquistato il software e gli archivi di Inktomi (www.inktomi.com). Il portale di Yahoo! ultimamente si appoggiava a Google (www.google.com), mentre in passato aveva utilizzato Altavista (www.altavista.com); l'acquisizione di Inktomi gli permetterà di usare un motore che a differenza degli altri due non è per nulla concorrente con il suo portale. Inktomi, infatti, senza mai aprire un proprio portale ha sempre fornito la sua tecnologia ad altri, come Hotbot (www.hotbot.com), Microsoft search (www.msn.com) e in Italia fino a poco tempo fa a Kataweb (www.kataweb.it). Nello scorso febbraio anche Kataweb si è convertito a Google. A inizio del 2003 è stato annunciato che Altavista sarà acquistato da Overture (www.overture.com). Per ora è difficile dire se Overture, specializzato nella vendita dei risultati delle ricerche (con le formule dette di «pay for»), modificherà l'attuale «oggettività» dei risultati che si ottengono con Altavista. Il general manager di quest'ultimo search engine, Kevin Eyres, giura però che gli indici non verranno inquinati in alcun modo dalle sponsorizzazioni.