

Trucchi e consigli per trovare in Rete ogni cosa

La quantità di informazioni nel Web è veramente enorme e un'indicizzazione generica spesso non basta. Ecco i suggerimenti per ottenere il meglio dai quattro più importanti motori generalisti; più un elenco di siti che danno accesso al Deep Web, dove i crawler non arrivano.

► A cura di Sean Carroll

Nonostante le loro differenze, per tecnologia e interfaccia utente, i motori di ricerca su Internet oggi sono più correlati tra loro di quanto non lo fossero in passato. I motori tradizionali indicizzano il Web con programmi appositi, gli spider o i crawler, che scandagliano la Rete catalogando i dati di ogni pagina individuata o i metadati che descrivono il contenuto delle pagine. La frequenza di indicizzazione di questi programmi è ovviamente importantissima: il Web cresce e cambia molto rapidamente. Col tempo i crawler hanno migliorato continuamente le loro ►►

funzioni, molti di essi sono ora in grado di individuare e indicizzare informazioni contenute in un numero maggiore di formati di documento, per esempio i Pdf, che rappresentano una consistente parte del cosiddetto Web invisibile, difficilmente individuabile dai vecchi motori di ricerca e quindi poco raggiungibile dai navigatori.

La quantità di informazioni presenti in Rete è veramente enorme e una semplice strategia di indicizzazione generica non è una strada efficiente. Un motore di ricerca deve essere in grado di valutare le pagine in modo tale che i link più comunemente riconosciuti come utili appaiano per primi nell'elenco offerto al navigatore a caccia di informazioni. Ci sono diversi modi per riuscire nel compito, il più

conosciuto si basa sulla valutazione della "popolarità" del sito, si contano in pratica il numero delle altre pagine Web che puntano al sito valutato. È il modo in cui opera Google, primo motore ad adottare questo criterio di valutazione, seguito poi in varie forme anche dai suoi concorrenti.

I più critici sostengono che questa strategia alla fine contribuisca ad accrescere la popolarità di un sito, piuttosto che limitarsi a rilevarla, ma il successo di Google è un dato innegabile. Questo motore è diventato un marchio conosciuto da tutti e chiunque almeno una volta l'ha utilizzato. Persino Google, nonostante il favore riconosciuto dall'utenza, ha però molta strada per migliorare. L'ultima sfida è il Deep Web, i dati in Rete che non possono essere indiciz-

zati per due motivi: perché non esistono sotto forma di pagine statiche ma sono elementi dinamici creati in risposta all'interrogazione di un database, o perché sono irraggiungibili a causa di procedure di autenticazione che un crawler non è in grado di compiere. Questo tipo di informazioni sono spesso le più aggiornate: riviste, libri, recensioni, giornali. Per ottenerle bisogna cercare da soli i contenuti in un database e magari pagare per l'accesso.

Nelle pagine seguenti abbiamo esaminato quattro motori di ricerca "generalisti" più conosciuti, riportando una serie di suggerimenti per ottenere il meglio da ciascun di essi; un elenco di siti che fanno entrare nel Deep Web (vedi il riquadro "Ricerche di nicchia"); una serie di tool software o siti Web di supporto nella ricerca di informazioni.

Come migliorare la ricerca

> Meglio un motore di riserva

Se è vero che il mondo dei motori di ricerca è sempre più interconnesso, ciò non significa che tutti i siti diano gli stessi risultati. Due tecnologie di ricerca che accedono alle stesse informazioni possono dare risultati diversi se pesano i dati con metodi differenti. Quando non si riesce a trovare un'informazione che si pensa sia disponibile nel Web, è meglio provare a cercarla con un altro motore, senza arrendersi. In questo modo, almeno, ci si tiene aggiornati sui cambiamenti degli strumenti di ricerca.

> Conoscere il motore preferito

Ogni sito ha una barra per il browser o un'interfaccia in cui si inseriscono le parole da cercare, ciascuno con funzioni e particolarità individuali. Una volta che si è scelto un motore è bene prendersi il tempo per leggere le pagine di help ed esaminare le funzioni di ricerca più avanzate, le capacità che ha di raffinare le ricerche e qualsiasi altra opzione dedicata agli utenti più evoluti. Questo piccolo investimento di tempo verrà ve-

locemente ripagato. Vale la pena verificare ogni tanto queste funzioni, poiché i migliori motori continuano a introdurre novità e migliorie.

> Imparare il gergo

Conoscere la sintassi del linguaggio d'interrogazione booleano può aiutare a focalizzare le ricerche. Alcuni siti consentono l'uso degli operatori booleani già nelle maschere di ricerca principali; altri richiedono di selezionare le opzioni di ricerca avanzate. Si possono sempre migliorare i risultati di ricerca usando una combinazione di operatori booleani standard AND (che indica che entrambi i termini devono essere presenti nel risultato della ricerca) OR (uno o entrambi i termini devono essere presenti) e NOT (il termine che segue al not non deve essere presente nei risultati).

Accanto all'uso dei termini citati spesso si ricorre anche all'uso di parentesi e virgolette per raggruppare oggetti, come avviene nelle equazioni matematiche. Per esempio, se

si sta cercando David Copperfield (il libro di Charles Dickens e non il mago illusionista) e non ci si ricorda il nome dell'autore, si può impostare la ricerca così: "David Copperfield" AND book NOT magic.

Mentre AND, OR, NOT sono i classici operatori booleani, alcuni motori hanno una loro libera interpretazione dei termini. Si può trovare ANDNOT al posto di NOT, per esempio. Alcuni addirittura hanno dei moduli da compilare, già organizzati secondo la logica booleana, in cui basta inserire i termini in box separati e connessi tra loro da menu a tendina, le cui scelte sono gli operatori logici booleani.

> Ricerche avanzate

La maggior parte dei motori di ricerca, convenzionali e speciali, offrono metodi alternativi di ricerca nel Web. Un elenco di directories e una barra di ricerca (search bar) sono funzioni ormai standard sulle schermate di apertura delle pagine Web. I link per le opzioni di ricerca avanzate vanno

Ricerche di nicchia

> Asiaco

<http://search.asiaco.com>

Un indice ricercabile degli argomenti attinenti all'Asia che si trovano su Internet.

> AskERIC

www.askeric.org

The Educational Resources Information Center (ERIC) un sistema di informazioni pubblico creato dall'Education Department degli USA.

> Ask Jeeves Kids

www.ajkids.com

Ogni bambino dovrebbe avere questo maggiordomo ad aiutarlo nelle ricerche su Internet (debitamente filtrate ma purtroppo solo in lingua inglese).

> Biography Resource Center

www.galegroup.com/BiographyRC

Permette di fare ricerche sulle persone in base alla data di nascita o di morte, la nazionalità, l'etnia, il tipo di lavoro che svolge o il genere (maschile, femminile).

> CiteSeer

<http://citeseer.nj.nec.com/cs>

Qualcuno ha mai citato quel documento nascosto che avete scritto sulla nanotecnologia? Qui potrete scoprirlo. Infatti vengono ricercati non solo i documenti a carattere scientifico indicizzati ma anche le citazioni che essi contengono.

Cool4Kids

www.cool4kids.com

Questo motore per soli bambini è diseg-

gnato sul Progetto di una directory aperta per bambini e teenager (Kids and teen Open Directory Project) e conta 17.549 link.

> EBizSearch

<http://gunther.smeal.psu.edu/index.html>

Qui si possono fare ricerche generiche sul Web come di articoli accademici e commerciali sulle varie tematiche dell'e-Business.

> eLibrary

<http://ask.elibrary.com>

Archivi ricercabili di libri, articoli, quotidiani, trascrizioni, immagini e mappe. Anche dopo il periodo di prova di sette giorni, si possono avere gratis degli abstract di articoli, ma i file completi sono a pagamento.

> Philosophy Research base

www.erraticimpact.com

C'è un po' di filosofia ovunque in questo sito, ma il cuore è dato da un motore di ricerca che trova i libri su argomenti filosofici in partnership con il sito di Amazon.com.

> FindArticles

www.findarticles.com

Un utile motore di ricerca consultabile gratuitamente che indicizza gli articoli pubblicati da oltre 300 fonti diverse.

GPO Access

www.access.gpo.gov/su_docs/multidb.html

Tutto sulla pubblica amministrazione americana con accesso a database multipli per consultare statistiche, pubblicazio-

ni e la storia delle varie amministrazioni.

> HighWire

<http://highwire.stanford.edu>

Si possono cercare oltre dodici milioni di articoli completamente indicizzati in più di 4.500 riviste a carattere medico, gli abstract, come minimo, sono sempre gratuiti. E oltre 360 siti di giornali offrono contenuti gratuiti (anche se relativi ai numeri passati).

> Hoover's Online

www.hoovers.com

Questo servizio in abbonamento a pagamento rende disponibili oltre 6.000 pubblicazioni per il mondo business disponibili tramite Factiva, una società di Dow Jones & Reuters. Si possono fare ricerche per parole chiave, nome di azienda e simboli.

> IncyWincy

www.incywincy.com

Lo spider di IncyWincy scandaglia il cosiddetto Web invisibile.

> JSTOR

www.istor.org

I ricercatori universitari strablieranno. Questa collezione di archivi contiene riviste accademiche piene di immagini e bene organizzate, che spaziano dalla botanica al mondo degli affari. È necessario autenticarsi tramite un'istituzione accademica.

> LexisNexis Academic

<http://web.lexis-nexis.com/universe>

Un database full-text a pagamento di news, informazioni finanziarie, legali e amministrative.

usati per definire, limitare o espandere i termini di ricerca. Molti siti guidano attraverso l'intero processo con i loro file di aiuto. Si possono usare gli operatori logici booleani (con i motori che li supportano), oppure si possono sfruttare i moduli per le ricerche speciali. Questi ultimi in certi casi consentono di abilitare filtri per le parole derivate o tronche (per esempio cercando color* per avere tra i risultati le parole: colore, colori, e colorato). I risultati si possono poi raffinare o presentare in un certo formato. Per esempio, le preferenze avanzate personalizzate di AllTheWeb permetto-

no di usare ricerche già preparate per migliorare i risultati, riformulando l'interrogazione iniziale, così come funzioni di autocompletamento (per suggerire modi di completare frasi parziali di ricerca) e di integrazione con le news. Le funzioni di ricerca avanzate di Google sono facili da gestire e consentono di mettere filtri a parole, domini, siti, lingue, oltre che ai filtri per i bambini.

> Attenzione alle parole

I motori di ricerca non si sottraggono alla legge del "Garbage in, Gar-

bage out": se si inserisce spazzatura si ottiene spazzatura. Ciò significa che i motori usano i termini che avete inserito per svolgere le ricerche e se commettete errori di battitura i risultati saranno privi di senso e poco utili. Quando ottenete risultati strani controllate l'ortografia.

> Sbagliare i termini

Il Web è pieno di errori ortografici, in qualche caso può anche non essere una cattiva idea "sbagliare" volontariamente una parola. Ovviamente dipende dal tipo di ricerca

che si sta compiendo. Se per esempio cercate "DAN ricombinante" al posto di "DNA ricombinante" difficilmente troverete materiale per una ricerca sulla genetica, ma se cercate "Avril Lavigne" al posto di "Avril Lavigne" i risultati si faranno vedere. Provare per credere.

> Ridurre i risultati

Restringere il campo di ricerca è utile se i risultati sono troppo numerosi e poco coerenti. In questo caso basta dare un'occhiata a qualche link tra quelli non pertinenti e cercare una parola chiave comune da escludere per restringere i risultati. Molti motori consentono di escludere alcuni

risultati servendosi di un'interfaccia avanzata o con l'uso di operatori booleani, altri attraverso simboli come il segno meno davanti alla parola chiave da escludere.

> Scegliere l'ampiezza nella Rete

Se non si riescono a trovare buoni risultati con i motori che normalmente si utilizzano più spesso, perché non provare con i metasearch? Sono strumenti che analizzano i risultati ottenuti da una serie di motori. Detto molto francamente la metaricerca non è la nostra strada preferita. I metamotori non hanno alle spalle la lunga storia dei sistemi classici, e dal momento che aggregano risultati

provenienti da siti differenti, spesso hanno il difetto di presentare troppe informazioni con un'interfaccia che non è di grande aiuto. Cosa ancora più importante essi non possono tradurre il linguaggio d'interrogazione nel formato specifico richiesto da ciascun sito.

Anche se i siti di ricerca individuali sono in continuo miglioramento, nessuno è perfetto e nessuno è in grado di indicizzare l'intera Internet. Se non si è riusciti ad avere buoni risultati concedete un tentativo a un sito di metaricerca come Dogpile, IxQuick, MetaCrawler, o Mamma.com, o uno strumento come Copernic (qui recensito). HotBot, che in genere non viene considerato un sito di metaricerca,

<segue>

> MergentOnline

www.mergentonline.com

Un'integrazione di database come l'Industry Manual di Moody, gli archivi EDGAR (Electronic Data Gathering, Analysis and Retrieval) e Company DataDirect (Usa e international).

> NatureServe Explorer

www.natureserve.org/explorer

Qui si possono cercare informazioni su piante, animali e ambienti ecologici relativamente alle zone degli Stati Uniti e del Canada.

> The On-Line Encyclopedia of Integer Sequences

www.research.att.com/~njas/sequences

Questo è un sito strano. Provate a digitare una serie di numeri e vi sarà dato il razionale che c'è dietro ciascun numero. Abbiamo provato "8549176320" - i numeri erano in ordine alfabetico in base alle lettere iniziali dell'alfabeto inglese - e il sito ha dato la risposta esatta.

> PublicLibraries.com

www.publiclibraries.com

Tutte le biblioteche online: quelle pubbliche, di Stato, universitarie, presidenziali, ma sempre relative agli Stati Uniti d'America.

> PubMed

www.ncbi.nlm.nih.gov/entrez/query.fcgi

Questo servizio della Biblioteca Nazionale di Medicina degli Stati Uniti fornisce ac-

cesso a più di dodici milioni di citazioni a carattere medico, a partire da metà degli anni Sessanta in poi.

> O'Reilly Network Safari Bookshelf

<http://safari.oreally.com>

Contiene le versioni elettroniche di centinaia di libri tecnici editi da un editore prestigioso di pubblicazioni informatiche, che coprono oltre 20 categorie, dal mondo Business al linguaggio Xml.

> S&P Netadvantage

www.netadvantage.standardpoor.com

Un buon posto per cercare informazioni sui trend degli investimenti, dei mercati e delle società in gioco. Tuttavia occorre un'autenticazione con password e user ID anche solo per accedere al sito.

> SearchBooks.com

www.searchbooks.com

Un indice a testo pieno degli e-book, cioè i libri elettronici.

> SearchEdu.com

www.searchedu.com

Ricerche all'interno del dominio .edu.

> Search Engine Colossus

www.searchenginecolossus.com

Per fare ricerche su tutti i search engine del mondo nella lingua che desiderate.

> SearchGov.com

www.searchgov.com

Ricerche all'interno del dominio .gov.

> SearchMil.com

www.searchmil.com

Ricerche nel dominio .mil.

> SpeechBot

<http://speechbot.research.compaq.com>

Un inventivo della Hewlett Packard ha ideato un sito di ricerca che indicizza 15.590 ore di trasmissioni sul Web, comprensive di audio e di trascrizioni.

> 10K Wizard

www.tenkwizard.com

Fa ricerche nella banca dati EDGAR della SEC (Secure Exchange) per trovare gli archivi in tempo reale.

> VolunteerMatch

www.volunteermatch.org

"Get out. Do good". Già il logo dice tutto. Si può cercare per area tematica, interessi e programmi.

> World News Connection

<http://wnc.fedworld.gov>

Un servizio a pagamento di aggregazione di news estere gestito da Dipartimento del Commercio USA. Volete scoprire le ultime sulle ruberie in Uganda? Questo è il posto giusto.

> Yahoo!igans!

www.yahooligans.com

Uno dei migliori siti di ricerca orientati ai bambini e un eccellente portale anche per i più giovani.



ha una nuova interfaccia accattivante che offre una specie di metaricerca laterale. Senza dover riscrivere la propria interrogazione si fa un clic sui pulsanti presenti per ottenere i risultati della ricerca in AllTheWeb, Google, Inktomi o Teoma.

> **Donde esta?**

Se la ricerca è in una lingua specifica o la si vuole limitare alla propria realtà locale o si ha bisogno della traduzione, sono a disposizione una quantità di strumenti pronti ad aiutare. Molti motori, primo tra tutti Google, consentono di impostare le preferenze in modo da avviare la ricerca nella lingua prescelta. Ci sono anche siti di ricerca specifici per paese e filtri che restringono la ricerca in base alla lingua e alla nazione.

> **Le proprie origini**

Search Engine Colossus (www.searchenginecolossus.com) ha una directory di 195 paesi e 38 territori.

Una ricerca sulla Svizzera ha prodotto qualche dozzina di link ai motori di ricerca locali, incluso un sito in lingua giapponese. Chi è a contatto con il mondo della scuola può trovare risorse utili anche in www.iecc.org per entrare in contatto con gli studenti di tutto il mondo.

> **Vedere l'invisibile**

Molti pensano di poter trovare qualsiasi informazione digitale, utilizzando i motori di ricerca. Non è vero. Esiste un'intera categoria di pagine che non hanno link che puntano verso di esse o che non sono indicizzate perché contengono dati come immagini, file musicali e Pdf, che spider e robot non indicizzano (molti motori convenzionali negli ultimi anni hanno però fatto progressi in questo campo). La parte più ampia del cosiddetto Deep Web consiste di database che gli spider non possono indicizzare. Uno spider può leggere l'indirizzo di un database ma non può decifrarne il contenuto perché le pagine dinamiche vengo-

no create solo come risposte temporanee alle interrogazioni degli utenti. Molti database, poi, richiedono una user id e una password che complica il lavoro a spider e robot. Si può accedere a questi database attraverso portali specializzati nel fornire accesso al Deep Web. Tra questi vi sono CompletePlanet (www.completeplanet.com) e InfoMine (infomine.ucr.edu).

> **Biblioteche online**

Le biblioteche restano sempre una grande fonte di informazioni, anche in quest'era online. Ma oggi molti degli indici usati in passato in versione cartacea, sono disponibili online. E cosa ancora più importante, lo sono anche le riviste e le recensioni dei quotidiani che senza dubbio molti di voi avranno usato in passato per le ricerche scolastiche. Molte biblioteche accademiche e pubbliche stanno convertendo gradatamente l'accesso alle loro opere dalla carta all'online, almeno negli Stati Uniti. E il futuro delle versioni stampate di queste pubblicazioni è in seria discussione. Come le biblioteche hanno pagato in passato per avere libri e periodici, oggi esse pagano un prezzo per la versione online corrispondente. Per questa ragione, l'accesso a tali documenti è ristretto e per poterli consultare occorre un permesso. Molte biblioteche forniscono un accesso remoto ai loro utenti tramite indirizzi Ip o username e password. Solitamente gli utenti devono configurare

I motori di ricerca italiani

I principali motori di ricerca italiani si avvalgono di engine internazionali (Google, AltaVista, Yahoo!, Inktomi), e quelli sviluppati in Italia si contano sulla punta delle dita. Il più conosciuto è il motore di ricerca Arianna, realizzato nel 1996 nei laboratori Olivetti di Pisa per il portale di Italia Online, che arriva a indicizzare fino a 44 milioni di pagine Web, 8 milioni d'immagini, 55.000 file video e un milione di file Mp3. Per le ricerche nel web Arianna si avvale di Google. Virgilio ha sviluppato in proprio le directory e ha recentemente introdotto i servizi di segnalazione dei siti a pagamento. I siti che pagano per comparire nelle ricerche sono riportati all'inizio dell'elenco dei risultati, sotto la dicitura "link sponsorizzato". Per le ricerche nel Web ricorre anch'esso a Google. Un'altra tecnologia italiana era quella di Janas, utilizzato da Tiscali prima che diventasse un network europeo. Oggi al suo posto c'è Fast. MSN.it ha sviluppato internamente una serie di directory aggiornate localmente da

un team di persone, a cui affianca la tecnologia di Inktomi. Per il resto ci sono altre tecnologie, come il Trovatore che indicizza fino a 40 milioni di pagine Web italiane, o il motore pay per click di Godado, ma ad esse manca la massa critica per competere con i "grandi" a livello internazionale.

Quali tecnologie usano i motori italiani

Italy	Ricerche nel Web locale	Ricerche nel www	Directory	Pay For Insertion	Pay per click
Google.it	Google	Google	ODP		Adwords
Virgilio.it	Google	Google	Virgilio *		
Arianna.iol.it	Arianna	Google		Own / Fast	
MSN.it	Inktomi	Inktomi / Lycos	MSN	Inktomi / Lycos	OV
Altavista.it	Altavista	Altavista	Looksmart *	Altavista	OV
Yahoo.it	Google	Google	Yahoo.it *		Spotting
Tiscali.it	Fast	Fast		Fast	
Lycos.it	Fast	Fast	Lycos.it *	Fast	OV/ES
Supereva.it	Google	Google	Supereva		Godado

Rielaborazione su fonte di Webmasterworld.com

La tabella riporta i motori italiani elencati in ordine di traffico generato

OV = Overture
ES = Espotting
* = inserimento dei siti a pagamento

Come abbiamo valutato i siti

Tutti i motori esaminati producono risultati congruenti, abbiamo quindi privilegiato come criteri di valutazione la facilità d'uso, la ricchezza di funzioni e la flessibilità. Un punteggio molto alto per l'**interfaccia** significa assenza di annunci a pagamento non pertinenti e di altri elementi estranei, oltre che risultati facili da scremare e un sistema di aiuto efficace. Un sito che riporta una **buona flessibilità** di ricerca è un sito che consente di specificare l'oggetto della ricerca utilizzando funzioni come gli operatori booleani, i caratteri jolly e ri-

cerche di prossimità.

Alcuni siti permettono di specificare dove si vuole ricercare. In questo caso si parla di opzioni di **ricerca mirate** a specifici domini o siti, o all'interno di URL o nei titoli e di ricerche limitate a certi tipi di file o di lingue. Abbiamo valutato i siti anche in base alle **opzioni di miglioramento dei risultati**, come la possibilità di rifinire la ricerca con nuove parole chiave o di visualizzare le versioni memorizzate nella cache del motore di ricerca e di associare a un particolare risultato pagine simili.

	Interfaccia	Flessibilità di ricerca	Ricerche mirate	Miglioramento dei risultati	Punteggio
AlltheWeb	●●●	●●●●	●●●●●	●●●	●●●●
Aiol	●●	●●	●	●●	●●
Google	●●●●●	●●●	●●●●●	●●●●	●●●●
MSN	●●●	●●●	●●●●	●●●	●●●
Yahoo!	●●●	●●●	●●●●	●●●	●●●

●●●●● = eccellente
 ●●●● = molto buono
 ●●● = buono
 ●● = discreto
 ● = scarso

il loro browser e fornire il loro codice d'iscrizione alla biblioteca per avere accesso ai database da casa. Alcune istituzioni consentono anche ai non iscritti di utilizzare le loro risorse online; altri (specialmente le biblioteche universitarie) restringono l'accesso o l'uso ai clienti registrati. Infine molte biblioteche hanno pagine Web che descrivono le risorse online disponibili, dando un'idea degli argomenti e delle aree coperti dalle loro opere online.

I database delle biblioteche sono creati da editori ben conosciuti e sono valutati molto attentamente prima di essere acquistati dalle varie istituzioni bibliografiche.

> Utilizzare più di una parola chiave

Quasi un terzo di tutte le ricerche sono interrogazioni basate su un termine solo. Anche se i motori sono diventati abbastanza intelligenti da accettare molte ricerche basate su un'unica parola chiave, in questi casi essi tendono a mettere per primi i risultati più popolari. Per esempio nei nostri test una delle ricerche che abbiamo condotto con ciascun motore utilizzando solo una keyword è stata

la parola Godiva, (ceravamo il marchio dei famosi cioccolatini). Ma se noi avessimo voluto cercare informazioni sull'origine anglosassone del nome Lady Godiva e sulla sua storia? Digitando Godiva nei campi di ricerca di Google si ottengono solo risultati pertinenti ai cioccolatini, ma se si digita il nome per esteso Lady Godiva ecco che si trova la pagina con il risultato cercato, cioè chi era Lady Godiva e l'origine del suo nome (Godgifu). A seconda della parola utilizzata si ottengono migliaia di risultati da una ricerca per singola parola chiave. Aggiungendone più di una non si limita necessariamente il numero dei risultati tra i quali cercare, ma si possono evidenziare i risultati più importanti tra i primi.

> Valutare se un'informazione è valida

Ci sono alcuni tool che sono in grado di dire molto sull'informazione trovata con un motore e che permettono di capire se essa sia valida o meno. Un buon criterio è controllare chi possiede il sito che il motore ha indicizzato, un'operazione che si può fare andando su www.whois.sc, oppu-

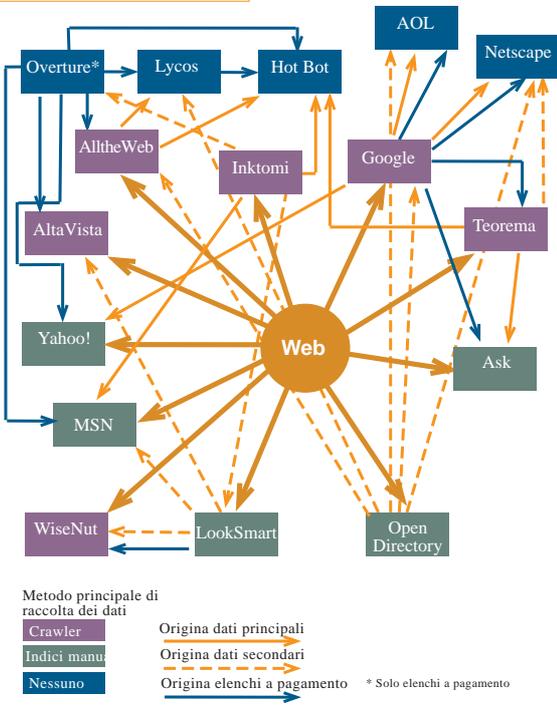
re sul più conosciuto www.whois.org. Scoprire chi c'è dietro un sito è un trucco prezioso, specie quando si cercano informazioni finanziari. In alcuni casi si possono considerare anche servizi di ricerca a pagamento come DQM2 Deep Query Manager di BrightPlanet che aiuta a trovare, classificare e gestire le informazioni.

> Uno sguardo al futuro della ricerca su Internet

Un motore di ricerca esegue i suoi compiti in un modo puramente meccanico. Ricerca il testo in un documento, tiene conto di quanto spesso le parole sono ripetute e quindi applica regole per la valutazione (ranking). Un buon motore di ricerca ha un'interfaccia facile da usare, deve documentare un enorme numero di pagine Web e aggiornare i link frequentemente. Ciò che però non può fare è capire il significato di una ricerca a un livello umano. Non può fare associazioni tra parole e argomenti nel modo in cui lo facciamo noi e quindi perde precisione nel recuperare le informazioni. Questo è il motivo per cui ci siamo abituati a fare acrobazie con i vocaboli e poi ci siamo accontentati di pochi utili link invece di nuotare in un mare di migliaia se non milioni di risultati del tutto inutili. Potrà migliorare questa situazione? Di fatto la risposta a questa domanda ha un nome: Latent Semantic Indexing (LSI). Si tratta di una tecnologia di ricerca che sfrutta un algoritmo per valutare le associazioni tra parole. Affiancando la tecnologia LSI a quelle già esistenti si possono ottenere livelli più alti di comprensione dei documenti consentendo anche la convergenza di numerosi indici di diversi database.

Come i motori tradizionali, la tecnologia LSI esamina le parole contenute nei documenti come opposti a parole di uso comune come *and* e *the*. Documenti contenenti molte parole simili vengono considerati semanticamente vicini, mentre quelli con poche parole uguali sono valutati semanticamente lontani. LSI quindi

Le ricerche nel Web



mappa per così dire la collocazione dei documenti all'interno di uno spazio multidimensionale, con una dimensione per ogni parola indice, raggruppando i documenti semanticamente vicini. Le trasformazioni matematiche rendono questo spazio accessibile e il risultato è una ricerca che trova pagine correlate anche se queste non condividono le stesse parole chiave.

Per chi vuole essere aggiornato su questa nuova area di sviluppo della ricerca ecco una serie di indirizzi utili:

- http://javelina.cet.middlebury.edu/lsa/out/cover_page.htm
- <http://lsi.research.telcordia.com/>
- www.psych.nmsu.edu/pfoltz/cois/filtering-cois.html

> Non si è soli quando si cerca sul Web

La ricerca come qualsiasi altra attività legata al computer e a Internet può essere tracciata. Un hacker può ricostruire le ricerche sul Web da

una quantità di fonti. Alcune, come l'history del browser, possono essere facilmente cancellate (sebbene distruggere determinate ricerche può richiedere più di una semplice cancellazione della cache). Altre tracce, come i log dei server in un'azienda, se si utilizzano i motori dal posto di lavoro, o i log del proprio Internet provider e della stessa società che fa il motore di ricerca, non sono facilmente cancellabili.

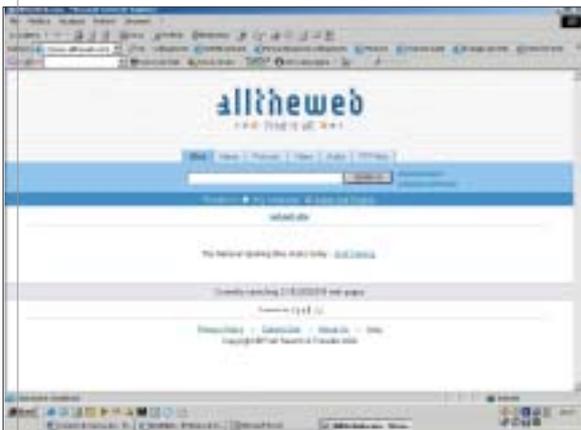
Anche le biblioteche si trovano ad affrontare la versione aggiornata di un problema vecchio: cosa fare in merito alle richieste di informazioni sulle abitudini dei prestiti dei clienti provenienti dall'autorità pubblica. Per esempio quando l'applicazione della legge richiede informazioni su una ricerca condotta al computer da un cliente della biblioteca. Questo non è un problema per gli utenti occasionali che non devono produrre alcun tipo di identificazione, ma se si usa una password per collegarsi in remoto bisogna considerare che tipo di informazioni personali si lasciano dietro di sé.

I search Engine

> AllTheWeb
Fast Search & Transfer ASA

www.alltheweb.com
Punteggio: ●●●●●

Ciò per cui è famoso AllTheWeb è l'ampiezza del suo indice, ma più grande non vuole dire necessaria-



mente migliore, e una ricerca veloce può essere una ricerca affrettata. Diversi dei nostri tentativi di interrogazione hanno prodotto risultati irrilevanti, riportando siti individuati in base a parole chiave nei loro metatag. Va detto però che le numerose opzioni avanzate di AllTheWeb e la sua abilità nel rifinire i risultati di ricerca sono di aiuto agli utenti più esperti.

Il sito è ben personalizzabile con opzioni sofisticate come le ricerche per pagine che puntano a un determinato sito, o nelle Url o nei titoli delle pagine e le ricerche limitate dalla dimensione della pagina web. C'è un box di verifica per le ricerche per frasi esatte e un'opzione di personalizzazione è quella di rendere accessibile la ricerca avanzata da menu a tendina già nell'interfaccia principale del programma.

Le funzioni avanzate non sempre seguono le convenzioni: per esempio si usano le parentesi per indicare l'operatore booleano OR già nella pagina principale (solitamente compare nella pagina della ricerca avanzata). Fortunatamente la pagina di aiuto è chiara ed estesa. AllTheWeb analizza anche la frase di ricerca utilizzata, reimpostandola con le virgolette ed eliminando parole come l'articolo *the* per ottenere migliori risultati. Si può disattivare tale funzione se lo si desidera, attraverso il pulsante Customize, ma noi abbiamo trovato che essa migliora considerevolmente i risultati delle query effettuate con il linguaggio naturale.

AllTheWeb dispone di etichette di ricerca specifiche che compaiono come menu pop up ovunque, dedicate alle news, alle immagini, ai file

Le toolbar di ricerca

Con l'aiuto di semplici programmi aggiuntivi e di strumenti specifici per il Web scovare le informazioni diventa più semplice. Per esempio le barre con gli strumenti di ricerca, add-on per i browser, danno accesso ai principali motori in modo molto immediato. Le applicazioni di metaricerca interrogano diversi motori in una volta. E le applicazioni di ricerca visuale forniscono un metodo intuitivo per individuare i concetti visualizzando i risultati sotto forma di mappe invece che di elenchi di indirizzi.

Diversi siti, incluso Google, Teoma e Yahoo! offrono toolbar di ricerca personalizzate. Ciascuna aggiunge una serie di strumenti al browser e box di testo dove compilare le interrogazioni per il motore di ricerca al quale si appoggiano.

> Google Toolbar è il migliore del gruppo. Consente di avviare due tipi di ricerche. Il pulsante Ricerche nel Web (Search Web) vi dà le pagine frutto di una ricerca generica, mentre il pulsante Search Site (Cerca nel sito) limita la query al dominio in cui vi trovate.

Le funzioni di ricerca nel sito spesso fanno un lavoro migliore di tanti altri motori internazionali. Basta selezionare lo strumento evidenziatore e Google fa risaltare le ultime frasi cercate all'interno di ciascuna pagina Web. Basta fare clic su un termine di ricerca e Google lo trova subito all'interno della pagina corrente. Si possono tradurre siti in lingua stra-

niera e un menu a tendina permette di saltare ai subsiti di Google, come Google News dove si possono cercare le ultime notizie

Google, <http://toolbar.google.com>. Punteggio: ●●●●●

> Teoma Search Bar offre un piccolo gruppo di strumenti che non sono disponibili con Google. Per esempio se si capita per caso su un sito che si ritiene molto interessante è possibile inviare al proprio indirizzo e-mail l'indirizzo di quella pagina. Con un altro pulsante si può consultare il dizionario on-line Merriam-Webster. La Search Bar di Teoma non è però in grado di limitare le ricerche per dominio o pagine tradotte. <http://sp.ask.com/docs/teoma/toolbar>. Punteggio: ●●●●●

> Molto di più di una toolbar di ricerca, Yahoo! Companion dà accesso a Yahoo! e-mail, offre un account su Yahoo! Shopping, l'accesso al sito di Yahoo! News e molto altro ancora. Si può trasformare l'applicazione in una toolbar per navigare in Yahoo! Finance oppure configurarla solo per le ricerche nel Web. Questa configurazione di Yahoo! Companion non può tradurre le pagine in altre lingue, ma offre comunque molte delle opzioni disponibili in Google. Per esempio la si può utilizzare per consultare on-line l'American Heritage Dictionary <http://companion.yahoo.com>. Punteggio: ●●●●●

video e audio e Ftp. Le ricerche nelle news sono personalizzabili selezionando tra le varie opzioni disponibili: internazionali, statunitensi, locali, business, e così via.

Quando AllTheWeb trova file multimediali, ne visualizza i link in fondo alla pagina dei risultati; noi non abbiamo mai trovato per caso un link audio, anche quando le ricerche sotto l'etichetta audio hanno prodotto

numerosi risultati.

AllTheWeb consente di aggiungere un pulsante di ricerca a Internet Explorer, una barra laterale in Netscape e un pannello con l'hotlist nel browser Opera. I risultati di AllTheWeb non sono però straordinari. Nelle prove che abbiamo eseguito il programma si è mostrato inadeguato nel trovare le home page e anche le ricerche in linguaggio naturale hanno dato problemi. In generale abbiamo ottenuto i migliori risultati con interrogazioni specifiche e basate su più parole.

> Google

www.google.com

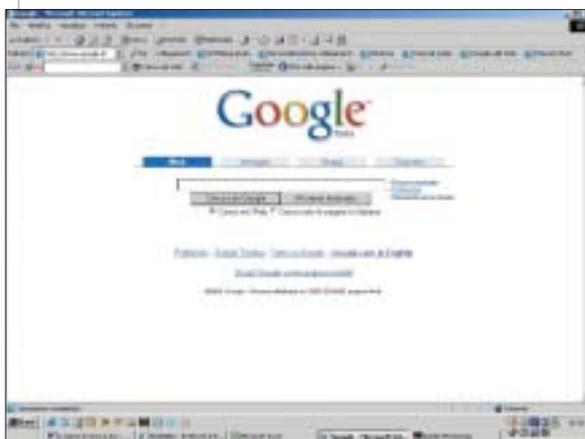
Punteggio: ●●●●●



Il talento di Google nel trovare qualsiasi cosa stiate cercando ha qualcosa di misterioso. Nelle nostre prove Google ha dato costantemente buoni risultati; non stupisce che fornisca i risultati delle sue ricerche a molti altri motori. Ci siamo sorpresi, però, del fatto che Google non abbia distaccato in modo così marcato i suoi

concorrenti, pur restando sempre in testa al gruppo nelle funzioni di miglioramento delle ricerche.

Le opzioni di personalizzazione vanno dalla lingua al filtro dei contenuti, al numero di risultati per pagina, ma non si possono aggiungere all'interfaccia principale né i menu a tendina di ricerca avanzata, né i box di controllo. In compenso la pagina per la ricerca avanzata dispone di svariate opzioni di ottimo livello, incluse le ricerche booleane, le ricerche per formati di file, data e dominio Internet. Ci sono anche un piccolo gruppo di ricerche non elencate tra quelle specializzate riportate nel menu principale, come le ricerche per pagine simili e per termini trovati nel titolo, nell'indirizzo Internet o le ricerche di testo in una pagina, o in altre pagine che puntano allo stesso Url. Tra le ricerche specializzate ci sono: Immagini, Newsgroup (c'è un immenso archivio di forum di discussione dei gruppi usenet) Directory (i contenuti sono organizzati per argomenti, come fa Yahoo!) e le News.



<segue>

> Se si cambia continuamente motore di ricerca una tool bar specifica di un sito non è di molto aiuto, in questo caso c'è bisogno di un'applicazione di metaricerca. Copernic Agent Basic è un tool scaricabile gratuitamente che dà accesso ad oltre 90 motori di ricerca e altre fonti di informazioni presenti in rete. Quando abbiamo eseguito i nostri test tuttavia tra di esse non abbiamo trovato Google per un problema tecnico. Copernic intende rendere di nuovo operativo Google ma non ha detto quando. Le 90 fonti attualmente offerte sono divise in 10 categorie, ciascuna copre un diverso tipo d'informazione. Una categoria chiamata The Web, per esempio, contiene 17 motori di ricerca, inclusi Aol, Msn e Yahoo!; cliccando su di essa e digitando una parola chiave s'interrogano simultaneamente tutti e 17 i motori. Con una seconda categoria si possono cercare i siti Web in un singolo paese straniero, scelto all'inizio, altre forniscono accesso a indirizzi e-mail e newsgroup. Le ultime sei categorie girano nel Web alla ricerca di una varietà di articoli. Copernic raccoglie i suoi risultati in una pagina singola, eliminando i link doppi. Con un clic si possono rimuovere i link non più attivi dall'elenco, salvare le pagine sull'hard disk per poterle consultare con calma off-line o inviare per e-mail gli indirizzi che si vogliono tenere da parte.

Ci sono due versioni a pagamento di questo tool: Copernic Agent Personal (29,95 dollari) e Copernic Agent Professional (79,95 dollari) che danno accesso a oltre mille fonti d'informazioni basate su Web per 120 categorie.

www.copernic.com. Punteggio: ●●●●●

> Un altro strumento che opera seguendo gli stessi principi di Copernic è FirstStop WebSearch, ma la versione base del prodotto, meglio nota come Standard Edition, dà accesso a solo 18 motori di ricerca in due categorie (WebSearch e News). La Deluxe Edition (24,95 dollari per l'acquirente finale) contiene i link a oltre 50 motori divisi in cinque categorie. La Visual Edition (79,95 dollari) fa un passo in più rispetto a Copernic, mostrando esempi visivi delle pagine dei siti già nei risultati. Nessuno di questi prodotti offre così tante funzioni di posticerca come Copernic, ma ciò significa anche che l'interfaccia di FirstStop è meno ingombrante e più facile da usare.

www.firststopwebsearch.com. Punteggio: ●●●●●

> In tema di metaricerche, strumenti come Grokker e KartOO offrono un tipo completamente diverso di interfaccia di ricerca e visualizzano i risultati non come elenchi sequenziali ma come mappe. Groxis non ha ancora perfezionato la versione definitiva di Grokker, al momento disponibile in una versione beta al prezzo di 99,95 dollari. Un prezzo che può sembrare eccessivo per un'edizione di collaudo. Viene descritto dalla società produttrice come un tool di personal data-mining, ovvero di gestione dei dati personali, piuttosto che come strumento di ricerca vero e proprio. Il programma raggruppa le informazioni in modo intelligente per aiutare a capire i concetti correlati a una determinata frase di testo.

La preview di Grokker fornisce un nuovo front end per tre diversi motori di ricerca: Northern Light (il cui proprietario, Divine, è fallito per bancarotta, e quindi questo non risulta di grande utilità), Open Directory Project (www.dmoz.com) e Teoma. Basta scegliere uno dei tre e inserire una frase da ricercare. Grokker organizza i risultati in categorie basate sulla rilevanza. Una ricerca sullo spam, per esempio, è stata suddivisa in *spamming*, *anti-spam*, e *spam-laws*. Grokker visualizza queste categorie come cerchi colorati su uno sfondo nero. Cliccando su una categoria si evidenziano delle subcategorie che a loro volta includono altre sotto-categorie e così via. Grokker richiede un po' di abitudine all'uso, ma può essere un modo intuitivo per fare ricerche per concetti.

www.groxis.com. Punteggio: ●●●●●

> KartOO è un altro servizio on-line di metaricerca che visualizza i risultati come mappe tematiche. Inserite una frase di ricerca e il programma crea una mappa in cui gli argomenti (le parole chiave) e i siti Web sono presentati in ordine d'importanza. Spostandosi con il mouse su ciascun argomento sulla mappa vengono mostrati i siti correlati, mentre muovendosi su un sito viene mostrata la parola chiave rilevante. Se incontrate un sito che vi piace potete fare clic con il mouse sulla sua icona per aprirlo in una finestra separata del browser. Se vi muovete con il mouse su un argomento, compaiono i segni più e meno. Il segno più corrisponde a un approfondimento, il segno meno vi porta indietro da quell'argomento a tutti gli altri disponibili. KartOO può apparire più intuitivo di Grokker ma questa è una questione di gusti personali.

www.kartoo.com. Punteggio: ●●●●●

Google migliora i risultati di ricerca controllando che le parole siano digitate in modo esatto e offrendo un aiuto a proposito. I termini più comuni come *what*, *of*, e *is* vengono ignorati. Si può forzare il motore a riconoscere una parola anteponendo a essa il segno +. La pagina di aiuto che è un po' troppo estesa per un visitatore frettoloso, illustra scorciatoie per rendere le ricerche più efficaci, incluse quelle nei titoli o nelle Url, guardando ai siti correlati o solo nelle pagine cache. Infatti

pressoché tutti i risultati di Google includono l'opzione di visualizzare le versioni delle pagine memorizzate nella cache del motore, che tornano utili quando un link non funziona più o un sito non è più disponibile. Google fornisce anche la traduzione delle pagine in alcune lingue straniere.

Nelle nostre ricerche comunque Google si è sempre distinto: sia quando si trattava di trovare home page specifiche sia quando c'erano ricerche complesse con più termini.



Ciò che lo rende superiore agli altri è che i risultati delle pagine web sono posti sempre in alto alla pagina, tranne rare eccezioni. I link sponsorizzati infatti sono espressi sotto forma di piccoli box di colore diverso in alto alla pagina dei risultati, o vengono posizionati ai lati dove non danno fastidio, sebbene il loro grado di pertinenza sia sempre molto elevato.

> MSN Microsoft. Corp.

www.msn.com

Punteggio: ●●●●●

MSN sta guadagnandosi lentamente il nostro VIP: è in grado di restringere le ricerche a domini particolari, ad alcuni tipi di file, aree geografiche e idiomi. Inoltre corregge automaticamente gli errori di ortografia o di battitura ed è esperto nelle interrogazioni in linguaggio naturale. I risultati di MSN però non sempre sono aggiornati e non si possono raffinare le ricerche con operatori specializzati già dalla pagina di ricerca principale. MSN non offre neppure pagine cache o traduzioni. I suoi risultati sono organizzati in cinque modi diversi: Populars Topics (le ricerche generiche fatte dagli utenti di MSN), Featured Sites, cioè i siti consigliati dagli editor della redazione di MSN; i siti sponsorizzati (Sponsored Sites), la Web Directory (un indice simile a quello di Yahoo!) e le Pagine Web (prese dal Web in generale). MSN visualizza

za solo i siti sponsorizzati più rilevanti. Quando abbiamo provato a cercare *McDonald's+ Wireless*, per avere informazioni sugli hot-spot del servizio di accesso a Internet in wi-fi offerto dalla catena di fastfood, MSN non ha mostrato i siti che vendono hardware wireless. Sfortunatamente le categorie sono sempre visualizzate con l'ordine sopra citato, con i siti generici seppelliti sotto altri risultati. E' discutibile anche la scarsa differenziazione con cui sono segnalati i siti sponsorizzati: non è abbastanza chiara e si finisce per scorrere i link sponsorizzati mentre si stanno leggendo i risultati. Un problema quest'ultimo che Google ha risolto in modo molto più gradevole.

Nelle interrogazioni con linguaggio naturale (in inglese) MSN si comporta meglio dei motori che basano la loro tecnologia su Google. Nella stragrande maggioranza delle ricerche, almeno uno dei primi dieci risultati era un link non funzionante e in qualche caso abbiamo avuto più di una riserva anche su quelli attivi. Ad esempio quando abbiamo cercato "bed and breakfast" + "New England" il sommario che MSN dava in uno dei risultati diceva "Bernice Chesler's Bed & Breakfast in New England Web Site is now closed". Quando si inseriscono le parole chiave nella maschera di ricerca che si trova nell'home page di MSN si possono usare le doppie virgolette per ricercare frasi specifiche. Ma per altri operatori come il segno meno (per escludere pagine contenenti date parole) o gli asterischi (come i caratteri jolly alla fine delle parole) bisogna andare nella pagina delle ricerche avanzate. E ad ogni modo bisogna passare di qui anche se si vuole restringere le ricerche per dominio, tipo di file, regione e idioma. Perché allora non renderli già disponibili nell'home page?

> Yahoo!

www.yahoo.com

Punteggio: ●●●●●

Yahoo! ultimamente ha migliorato in modo considerevole il suo sistema di ricerca e ora fornisce diversi strumenti per definire meglio le interrogazioni. I risultati di ricerca dal Web in generale sono forniti da Google e vengono raggruppati nella stessa pagina in cui compaiono i risultati della directory di Yahoo!; in passato invece i risultati delle ricerche nel Web erano riportati in una pagina diversa. Yahoo! può essere la scelta migliore per quanti preferiscono fare ricerche utilizzando la sua Directory, viene visualizzata sopra i risultati delle ricerche nel web generico e che spesso contiene risultati più pertinenti. Per contro qualcuno potrebbe non apprezzare l'essere spinto verso i siti sponsorizzati. Quando Yahoo! mostra i risultati delle ricerche prima cerca di far combaciare la ricerca effettuata col motore di ricerca con le categorie della directory e posiziona i link sponsorizzati sopra i risultati di Google. Quando si fa una ricerca su prodotti e servizi i link sponsorizzati possono anche rivelarsi utili, ma se si stanno cercando semplicemente informazioni, i siti sponsor possono fuorviare. Per esempio quando abbiamo inserito la parola chiave 1394, come criterio di ricerca per l'home page del 1394 Trade Organization, Yahoo! ci ha portati ad essa, dopo aver elencato tre siti che vendevano computer.

Grazie a una nuova pagina di ricerca avanzata ora è possibile restringere le ricerche a domini particolari, aree geografiche e idiomi linguistici. Per esempio si può impostare una ricerca basata sulla data ultima di aggiornamento dei siti oppure si possono cercare solo le pagine che mostrano una certa parola chiave nelle loro Url. Yahoo! comunque ha ancora molta strada da fare: manca di alcune capacità di ricerca che abbiamo potuto apprezzare in altri motori, come l'uso dei caratteri jolly e la possibilità di avviare ricerche booleane già all'interno del box di ricerca principale. A breve però dovrebbe essere rilasciata una nuova versione del motore. ■

