VERSO I MOTORI DI RICERCA DI PROSSIMA GENERAZIONE

Claudio Carpineto
Giovanni Romano

I motori di ricerca sono diventati sempre più potenti e aggiornati. Ora lo sforzo è cercare di renderli anche più intelligenti. Ma che cosa limita le prestazioni dei motori di ricerca attuali e quali problemi devono essere risolti per costruire quelli di prossima generazione? L'articolo discute questi temi e delinea le innovazioni che possiamo realisticamente attenderci nei prossimi anni.

1. UNA STORIA BREVE E DI SUCCESSO

I motori di ricerca costituiscono uno dei fenomeni tecnologici, culturali ed economici più importanti di questi ultimi anni e sono considerati da molti come la terza "killer application" di Internet dopo i browser web e la posta elettronica. Il loro successo è enorme e trasversale, e può essere osservato a vari livelli. Indagini di mercato recenti concordano che nell'80% dei casi gli utenti utilizzano un motore di ricerca per trovare i siti di interesse e si stima che ogni giorno vengano inviate collettivamente ai motori di ricerca cinquecento milioni di interrogazioni. Il mercato delle inserzioni a pagamento è in piena espansione e abbiamo visto tutti che la recente quotazione in borsa di Google ha rinverdito i fasti della bolla speculativa degli anni '90. Ancora, società specializzate esclusivamente nel migliorare il posizionamento nei risultati prodotti dai motori di ricerca stanno sorgendo un pò ovunque. Anche la lingua recepisce la novità. In inglese, "to google" è ormai un verbo di uso comune; significa "cercare su Internet".

Questi sistemi esercitano un monopolio di fatto in un settore delicato e cruciale come quello dell'accesso mondiale alle informazioni – non si dice che se non stai nella prima pagina dei risultati di Google è come se non esistessi? – e ormai vengono attentamente scrutinati anche riguardo alla trasparenza e "democraticità" dei risultati forniti. Gli interrogativi, giustificati dalla opacità delle tecnologie proprietarie, si vanno rafforzando alla luce dell'affermarsi di un modello di business basato sui collegamenti (link) sponsorizzati e dei forti fenomeni di concentrazione industriale cui stiamo assistendo.
Eppure all'inizio della loro breve storia molti analisti dissero che quello dei motori di ricerca sarebbe stato un fenomeno marginale. Era il 1993, e il web consisteva di poche centinaia di siti. Per contarli, Matthew Gray sviluppò il primo "web crawler", chiamato World Wide Web Wanderer. Questo fu il padre (o la madre) dei moderni motori di ricerca, mentre i nonni sono probabilmente da rintracciare in Archie e Veronica, di qualche anno precedenti. Questi ultimi non cercava-

no pagine web (il protocollo HTTP doveva essere ancora inventato) bensì file e testi, rispettivamente, da trasmettere con le due modalità che all'epoca la facevano da padrone su Internet, e cioè FTP e Gopher.
Il periodo immediatamente successivo all'introduzione del World Wide Web Wanderer fu il più ricco di novità. Limitandoci alle innovazioni nella logica di reperimento delle informazioni (per un resoconto storico più dettagliato si può consultare la pagina http://www.wiley.com/legacy/compbooks/sonnenreich/history.html) ecco che cosa successe. I crawler vennero rapidamente potenziati per estrarre URL e titoli, o usare brevi descrizioni fornite manualmente. Subito dopo WebCrawler indicizzò pagine intere, analogamente a quanto avviene oggi, ed Excite usò per primo l'analisi statistica delle occorrenze delle parole. Siamo nel 1994. Yahoo! costruisce il primo catologo web e più o meno nello stesso periodo c'è l'esordio di Lycos, che oltre a consentire la ricerca per parole contigue e ad ordinare i risultati per pertinenza, si caratterizza per le dimensioni: parte con 50000 pagine e nel giro di due anni arriva a contenerne 60 milioni. Il passo successivo più importante fu Altavista, debuttante nel dicembre '95, che accettava interrogazioni in linguaggio naturale e offriva la possibilità di contare i link entranti in un sito. Nel 1998 fu lanciato Google, che introdusse la "popolarità" dei siti come criterio chiave nella ricerca delle informazioni e si contraddistinse subito per la maggiore precisione dei suoi risultati. Google in qualche modo ha segnato un punto di svolta e sancito la maturità di uno stadio tecnologico, con una spinta alla omologazione dell'offerta. Successivamente, sono apparsi altri motori di ricerca di qualità elevata, come ad esempio AlltheWeb e l'ultimo Yahoo!, ma si ha l'impressione che non ci siano state radicali innovazioni tecnologiche.
In effetti, il miglioramento degli ultimi anni ha riguardato principalmente aspetti ingegneristici. Oggi, con decine di migliaia di computer in parallelo, si riescono a censire miliardi di pagine web, a seguirne gli aggiornamenti in modo sempre più puntuale e a rispondere a migliaia di interrogazioni simultaneamente. All'aumento di forza però non è corrisposto un pari aumento di intelligenza. Gran parte del web rimane *invisibile*, in particolare tutte le informazioni che vengono generate dinamicamente in seguito ad una interazione con l'utente, e il funzionamento basato su parole chiave non consente di capire la semantica delle interrogazioni e delle pagine *visibili*, con l'effetto di generare molti risultati ridondanti o palesemente inutili.
Nel seguito di questo articolo analizzeremo i principi di funzionamento dei motori di ricerca, per capire meglio le loro limitazioni intrinseche e i problemi che bisogna risolvere per spianare la strada ai sistemi di prossima generazione. Successivamente vedremo una serie di sviluppi recenti che, seppure a livello prototipale, vanno già oltre le funzioni e le prestazioni dei sistemi commerciali odierni.

2. PRINCIPI DI FUNZIONAMENTO

L'interazione fra l'utente e il motore di ricerca inizia con l'invio di una interrogazione, tramite *form* HTML. Il motore di ricerca utilizza le parole dell'interrogazione per cercare nei file indice che si è precedentemente costruito scaricando e analizzando tutte le pagine del web, quali pagine contengono quelle parole. Tali pagine vengono quindi ordinate per pertinenza utilizzando vari criteri, che essenzialmente si basano sul contenuto testuale delle pagine stesse e sulle informazioni rappresentate dai link sul web che puntano ad esse. Il risultato viene mostrato all'utente utilizzando una pagina HTML che contiene rappresentazioni condensate delle pagine più pertinenti. L'utente a questo punto può scegliere di scaricare una o più pagine intere che sono d'interesse o di inviare una nuova interrogazione al motore. Questo schema di massima è rappresentato nella figura 1. Nel seguito saranno descritte con maggiore dettaglio le tre funzioni chiave di un motore di ricerca, e cioè quelle per la raccolta (*crawling*) e indicizzazione (*indexing*) delle pagine e per l'ordinamento dei risultati (*ranking*).

2.1. Raccolta

Il primo stadio di un motore di ricerca è costituito dal *crawler* (chiamato anche spider, o web robot), il programma per raccogliere le pagine web e scaricarle in locale. Poichè non esiste l'equivalente di un elenco completo delle pagine web, l'unico modo per eseguire

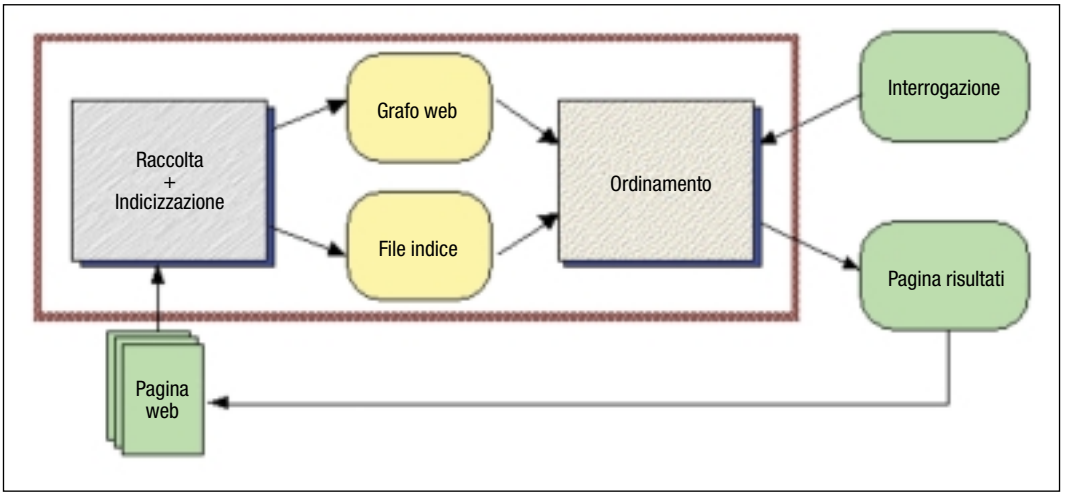

FIGURA 1
Schematizzazione funzionale di un motore di ricerca

questa operazione è quello di sfruttare i link fra di esse. Si parte da un insieme iniziale di pagine e si scaricano tutte le pagine raggiungibili dall'insieme seguendo i link, evitando di visitare quelle già viste. Si può fare o una visita in ampiezza, o in profondità, scegliendo dei criteri di terminazione legati alla profondità di visita o al numero totale di pagine raccolte. In linea di principio si tratta quindi di un programma molto semplice da realizzare, riassumibile addirittura in un singolo comando UNIX (*wget*).
In realtà la costruzione di un crawler su scala industriale, in grado di scaricare una frazione significativa del web in tempi ristretti (ordine di decine di migliaia di pagine al secondo), deve confrontarsi con una serie di problemi tecnici, alcuni di soluzione relativamente semplice, come l'estrazione e la normalizzazione delle URL e l'eliminazione efficiente delle URL già visitate, altri più complessi, come l'ottimizzazione della risoluzione delle URL negli indirizzi IP e la gestione di connessioni HTTP multiple a un server senza sovraccaricare quest'ultimo di richieste o saturare la sua banda. Per i lettori interessati ai problemi sottostanti alla ingegnerizzazione su larga scala di un crawler si rimanda all'eccellente capitolo su Web Crawling in [4]; per gli scopi di questo articolo è sufficiente capire il principio di base.

2.2. Indicizzazione

Dopo aver scaricato una pagina, il suo contenuto viene elaborato per estrarne il testo e inserirlo in un file indice che associa a ciascuna parola nella collezione le pagine in cui quella parola compare. La preparazione del file indice avviene dietro le quinte, ma è altrettanto importante delle successive e forse più suggestive elaborazioni, perché sono le parole immagazzinate nel file indice quelle che in ultima analisi determinano i documenti candidati a far parte dei risultati di un'interrogazione.
L'estrazione del testo varia da sistema a sistema in dipendenza di una serie di ragioni. Innanzitutto la presenza di errori di formattazione può influenzare in modi differenti il risultato prodotto dagli analizzatori sintattici delle pagine. In secondo luogo, i marcatori HTML possono essere trattati in modo differenziato, ad esempio stabilendo di ignorarne alcuni come "comment" e "alt text" e di annettere un'importanza particolare ad altri, come "title". Un ulteriore livello di variabilità può essere rappresentato dalla decisione di gestire anche pagine in formati non testuali, come per esempio *pdf* o *ppt*. Il testo estratto dalle pagine viene successivamente segmentato in parole singole, ignorando la punteggiatura e tenendo conto anche dei caratteri non alfanumerici.
A questo punto, l'insieme delle parole contenute in una pagina può essere ulteriormente normalizzato utilizzando una "stop list" (cioè togliendo le parole a basso contenuto informativo come articoli, preposizioni e pronomi) e lemmatizzando le parole (per esempio per assimilare plurale e singolare di uno stesso sostantivo o le diverse forme di un verbo). Questo trattamento è prassi in molti sistemi di "information retrieval", principalmente al fine

di aumentare il numero di documenti pertinenti che vengono recuperati dal sistema, ma è meno diffuso nei motori di ricerca per il web, in parte perchè ciò precluderebbe, di fatto, la gestione di alcune interrogazioni (come quella relativa al gruppo musicale "The Who"), in parte perchè per altre interrogazioni, anche comuni, ci sarebbe una probabile penalizzazione della precisione dei risultati. Infatti, l'enorme abbondanza di sigle e abbreviazioni di natura tecnica e commerciale presenti sul web può facilmente confondere un motore di ricerca con lemmatizzatore nel caso in cui ci sia coincidenza con qualcuna delle varianti linguistiche associate all'interrogazione (per esempio "sock" e "SOCKS", o "ides" e "IDE").
Il file indice è l'equivalente dell'indice analitico di un libro, che ci consente di trovare subito l'argomento cercato senza sfogliare le pagine. Per ogni parola, il file indice contiene la lista delle pagine in cui quella parola compare con la posizione relativa. Quest'ultima informazione è essenziale nel caso in cui si voglia sfruttare la prossimità delle parole dell'interrogazione nelle pagine ai fini dell'ordinamento. In pratica, il file indice viene implementato utilizzando delle strutture dati che consentono un accesso veloce a ciascuna parola, ad esempio utilizzando una tavola "hash" per associare alle parola una chiave in uno spazio d'indirizzamento diretto ridotto, ed esso viene di solito partizionato per pagine o per parole per parallelizzare gli accessi e diminuire i tempi di risposta durante l'elaborazione delle interrogazioni [1]. Nel file indice, per ciascun termine comparirà, oltre alla lista delle pagine in cui compare e alla posizione nella pagina, un peso della parola in ciascuna pagina. Il significato di questo peso è descritto nella sezione seguente.

2.3. Ordinamento

Quando viene elaborata una interrogazione, il file indice consente di accedere velocemente a tutti i documenti in cui sono presenti i termini della interrogazione. Poiché anche per le interrogazioni più specifiche ci possono essere migliaia di pagine coinvolte, è necessario ordinare i risultati per pertinenza. Il meccanismo fondamentale per l'ordinamento dei risultati, mutuato dalle ricerche fatte nel settore dell'"information retrieval", consiste nell'assegnare ad ogni termine di ciascun documento un peso (durante la fase di indicizzazione), e nel calcolare poi a "run-time" il punteggio di pertinenza di ciascun documento a fronte di una interrogazione sommando i pesi dei termini dell'interrogazione che sono presenti nel documento.
Schematicamente, il peso di un termine in un documento dipende da quanto il termine caratterizza il documento in oggetto, da quanto lo discrimina rispetto agli altri documenti e dalla lunghezza del documento stesso. Questo modello a tre componenti è alla radice di vari modelli di pesatura sviluppati nella comunità di "information retrieval", a partire dal classico modello a spazio di vettori [6], in cui il peso di una parola è proporzionale alla frequenza della parola nella pagina e inversamente proporzionale alla frequenza delle pagine in cui compare la parola e alla lunghezza della pagina.
Recentemente, sono state sviluppate tecniche più efficaci, principalmente utilizzando formule probabilistiche derivate dal teorema di Bayes o seguendo il paradigma della modellazione statistica del linguaggio, adoperato anche nel riconoscimento del parlato. Un terzo approccio, denominato divergenza dalla casualità (deviation from randomness), è stato sviluppato e sperimentato dagli autori di questo articolo insieme al collega Giambattista Amati della Fondazione Ugo Bordoni. Esso ha ottenuto risultati lusinghieri nelle ultime edizioni di TREC (Text REtrieval Conference) e CLEF (Cross Language Evaluation Forum), due forum scientifici internazionali dedicati alla sperimentazione e alla valutazione di prototipi innovativi per la ricerca delle informazioni. La sua idea essenziale consiste nel calcolare i pesi come una funzione inversa della probabilità che l'occorrenza di un termine in un documento segua una distribuzione casuale (per i lettori interessati ad approfondire l'argomento si rimanda alla consultazione degli atti TREC e CLEF). È importante notare come tutte queste funzioni di pesatura si basino su semplici statistiche relative alle pagine gestite dal sistema e possano quindi essere calcolate in modo estremamente efficiente.
Il meccanismo di ordinamento appena descritto si basa sulla similarità fra l'interrogazione e il contenuto testuale della pagina. Il

secondo criterio fondamentale per riuscire a filtrare ed ordinare in modo più efficace l'enorme quantità di pagine teoricamente pertinenti è basato sulla struttura del web, a prescindere dal contenuto testuale delle pagine. L'osservazione chiave è che certi siti web sono oggettivamente più importanti o "popolari" di altri e che un indice significativo della loro popolarità è costituito dal numero di pagine che puntano ad essi, così come nel mondo scientifico il prestigio di un articolo è misurato dal numero di articoli che lo citano. In realtà, per misurare la popolarità di un sito, contano sia la quantità dei siti che puntano ad esso sia la popolarità di questi ultimi, con una definizione evidentemente ricorsiva.
Per venire a capo di quest'ultima si procede così. Inizialmente si assegna uno stesso punteggio di popolarità a ciascuna pagina. Poi si calcola un punteggio aggiornato, dove sostanzialmente il punteggio di una pagina A è ottenuto sommando i punteggi delle pagine che puntano ad A divisi per il numero di link uscenti da ciascuna di esse. A questo punto, si itera il procedimento considerando ad ogni passo i punteggi correnti, fino a quando i punteggi non cambiano più. Questa è l'idea di base di PageRank, l'algoritmo utilizzato da Google. La formula esatta e un esempio molto dettagliato con i risultati delle singole iterazioni sono disponibili all'indirizzo http://www.whitelines.nl/html/google-page-rank.html. A questo punto ciascun sito (o pagina appartenente ad un sito) riceve un punteggio che misura la sua popolarità, e che può essere usato per influenzare la sua posizione nella lista dei risultati.
Per esattezza scientifica bisogna sottolineare che PageRank è stato preceduto da modelli analoghi che sono probabilmente anche migliori da un punto di visto teorico, come HITS di Kleinberg e Hyperinformation di Marchiori; Google è stato però il primo a mostrare la bontà dell'idea in un sistema commerciale su larga scala. Recentemente, il concetto di popolarità è stato raffinato con buoni risultati da Teoma, il quale calcola una popolarità relativa all'argomento dell'interrogazione invece che quella assoluta del sito.
Oltre al contenuto e alla struttura dei link, ci sono altre sorgenti informative che possono concorrere a determinare l'ordinamento finale delle pagine. Una di queste è costituita dai cataloghi web, vale a dire un insieme di categorie strutturate gerarchicamente che coprono un vasto spettro di argomenti, con pagine web significative associate alle singole categorie mediante un processo manuale. Il catalogo sviluppato da Yahoo! e la Open Directory, con decine di migliaia di categorie aggiornate da una vasta comunità globale di redattori volontari, sono probabilmente gli esempi più noti. Quando le parole dell'interrogazione coincidono con quelle di una categoria, il risultato puo essere fornito direttamente dal catalogo, che contiene pagine più controllate e verosimilmente pertinenti di quelle recuperate con una ricerca automatica. In effetti, sono molti i motori di ricerca che utilizzano in modo diretto o indiretto i cataloghi.
Un'altra tecnica per migliorare l'ordinamento consiste nell'utilizzare le ancore del codice HTML per estrarre termini e concetti con i quali arricchire la descrizione delle pagine puntate. In questo caso l'osservazione è che possono esserci siti anche molto importanti in cui l'oggetto primario del sito non viene menzionato nella *home* (per esempio automobile per Toyota e computer per IBM), mentre è probabile che esso compaia nella descrizione associata ai link entranti.
I differenti metodi di ordinamento possono essere integrati in vari modi. Si può usare per esempio la popolarità per elaborare a posteriori i risultati ottenuti utilizzando il contenuto o, dualmente, come criterio di filtraggio a monte della ricerca per contenuto. Un'altra possibilità è quella di calcolare in modo separato l'ordinamento prodotto da ciascun metodo e poi combinare direttamente i risultati. Google dice che utilizza più di 100 ingredienti, ma la ricetta è segreta, un pò come quella della Coca Cola. E come per la Coca Cola, questo ha probabilmente una importanza secondaria, se non altro perché nessuno oggi potrebbe essere seriamente intenzionato a riprodurre esattamente i risultati di Google con un altro marchio. Per uno studio sistematico di natura accademica sulla combinazione dei risultati prodotti da differenti metodi di ordinamento si può consultare l'ottima indagine di Yang [8].

3. LIMITAZIONI INTRINSECHE E INGANNI DELIBERATI

In alcuni casi i motori di ricerca sono di scarsa utilità. Anche con una scelta oculata delle parole dell'interrogazione o con una utilizzazione corretta delle opzioni della ricerca avanzata i risultati non soddisfano le aspettative, e possono indurre negli utilizzatori inesperti un senso di frustrazione. In questa sezione cercheremo di capire meglio che cosa un motore di ricerca non può fare o riesce a fare solo in parte.

3.1. Web nascosto

Il primo limite oggettivo dei motori di ricerca è che essi indicizzano solo una parte del web. L'esistenza del **web nascosto** è dovuta ad una serie di ragioni che qui cercheremo di riassumere brevemente. Innanzitutto, certe pagine non sono collegate alle altre e quindi il crawler non può fisicamente accedere ad esse. Nell'insieme delle pagine collegate invisibili si possono distinguere poi due categorie, quelle tecnicamente accessibili e quelle tecnicamente inaccessibili (o accessibili con difficoltà). Nella categoria delle pagine invisibili che sono tecnicamente accessibili rientrano le pagine che il crawler decide autonomamente di non scaricare, per esempio utilizzando un limite di profondità intra-sito per limitare i costi e i tempi di visita del web oppure una soglia sulla dimensione minima per scartare pagine poco significative, e le pagine che non è autorizzato a scaricare (specificate con il file *robots.text*).
La categoria delle pagine tecnicamente inaccessibili comprende quelle che contengono formati che i crawler normalmente non gestiscono (immagini, audio, Flash, Shockwave, file compressi ecc.) e le pagine generate dinamicamente previa interazione con una *form* HTML, che il crawler non sa come riempire. Queste ultime tipicamente vengono generate da basi di dati specializzate in risposta ad una interrogazione e costituiscono la parte più pregiata del web nascosto. Si pensi a previsioni del tempo, orari dei voli, quotazioni di borsa, ricerche bibliografiche e molte altre sconosciute ai più, che sono state raccolte in un elenco ragionato in [7]. Anche in quest'ultimo caso, l'inaccessibilità è un concetto relativo che riflette scelte commerciali insieme a obiettive difficoltà tecniche, perché un crawler può essere programmato anche per estrarre informazioni da una base di dati.

3.2. Andare oltre le parole chiave

Una volta che le pagine sono state scaricate e il file indice costruito, ciascuna interrogazione viene elaborata recuperando soltanto le pagine che contengono esattamente le parole specificate nell'interrogazione. Questa è una limitazione molto forte, alla luce della ricchezza e dell'ambiguita del linguaggio naturale. In particolare, se una pagina contiene lo stesso

Il **web nascosto** (o profondo, o invisibile) è la parte che non viene indicizzata dai motori di ricerca e che si stima essere centinaia di volte più grande di quella visibile. La sua esistenza è dovuta ad un insieme di ragioni, riconducibili essenzialmente alle politiche di visita selettiva del web adottate dai crawler e alla presenza di contenuti non testuali che sono difficilmente indicizzabili dato lo stato della tecnologia e i vincoli operativi e commerciali cui i motori di ricerca sono soggetti. Le principali tipologie d'invisibilità sono riassunte nella seguente tabella.

Tipologia del contenuto web	Perché è nascosto
Pagine disconnesse	Non ci si può arrivare
Pagine periferiche	Il crawler si ferma prima
Immagini, audio, o video	Non c'è testo
Flash, Shockwave, .zip,.tar ecc.	Indicizzazione costosa
Informazioni fornite da basi dati	Il crawler non sa fare le interrogazioni
Dati che cambiano in tempo reale	Informazioni effimere e voluminose

concetto espresso con parole differenti essa non viene recuperata (*problema del vocabolario*). L'utente alle prese col problema di scegliere le parole giuste per descrivere le pagine che non conosce fa venire in mente la situazione di quel proverbio cinese che dice: "Sei sai quello che cerchi perchè lo cerchi? E se non lo sai come speri di trovarlo?".
La situazione è ulteriormente complicata dal fatto che le interrogazioni sono brevi (di solito non più di due o tre parole) e il web è estremamente ricco ed eterogeneo in contenuti. In queste condizioni, a causa dei problemi di sinonimia (parole differenti con lo stesso significato) e polisemia (una stessa parola con significati differenti), è ancora più probabile che il sistema non riesca a recuperare pagine pertinenti che non contengono gli stessi termini dell'interrogazione oppure, simmetricamente, che recuperi molte pagine non pertinenti.
Per alleviare questo problema si possono utilizzare tecniche in grado di estrarre ed utilizzare informazioni che *non* sono contenute esplicitamente nella rappresentazione di pagine e interrogazioni. Una di tali tecniche, detta retroazione di pseudo pertinenza (pseudo-relevance feedback), sfrutta la ridondanza del linguaggio naturale per arricchire la formulazione dell'interrogazione. Essa consiste in un doppio ciclo di ordinamento, nel secondo dei quali è utilizzata una interrogazione espansa con parole estratte dalle pagine recuperate dal sistema nel primo ciclo. La retrazione di pseudo pertinenza è interessante per la sua semplicità e per le sue superiori prestazioni documentate sperimentalmente, specialmente se paragonata ad altri approcci basati sulla similarita inter-pagina, però fatica ad essere adottata dai sistemi commerciali, sia per l'aggravio dei tempi di risposta sia per la sua relativa robustezza.
Indipendentemente dalla possibilità di migliorare la formulazione delle interrogazioni, rimane il fatto che un meccanismo basato su parole chiave non riesce a cogliere la semantica di pagine e interrogazioni, e può fraintendere la loro apparente somiglianza. Per esempio, la frase "Giovanni ama Maria" non sarebbe discriminata da "Giovanni non ama Maria" o "Giovanni ama la migliore amica di Maria", perchè ciò richiederebbe la capacità di identificare concetti e relazioni fra concetti (chi ama e chi è amato). Sembrerebbe quindi che l'uso di indici concettuali, legati alla elaborazione del linguaggio naturale, sia la strada maestra per andare oltre il paradigma delle parole chiave, ed effettivamente ricerche in questo senso sono in corso da vari anni. C'è da dire però che questi tentativi hanno finora avuto poca fortuna, perché in genere i sistemi risultano più lenti e meno accurati di quelli statistici, specialmente se vengono applicati a compiti tradizionali di ricerca delle informazioni su collezioni omogenee. Più recentemente, però, sono emersi compiti di ricerca delle informazioni più specializzati, di cui vedremo un esempio più avanti, in cui tali metodi giocano in effetti un ruolo essenziale. Tecniche linguistiche "superficiali" possono rivelarsi utili anche per ricerche generiche sul web: alcuni motori di ricerca commerciali stanno cercando di sfruttare certi "pattern" linguistici (per esempio "such as....") per trovare entità con nome proprio o altri concetti rilevanti all'interno delle pagine e scartare il resto già in fase di indicizzazione.
Esiste poi una vasta classe di richieste che non possono essere esaudite senza assumere l'esistenza di qualche forma di strutturazione dei dati. È il caso di quei compiti di ricerca delle informazioni in cui vengono espressi vincoli temporali, spaziali, o di altro tipo, la cui risoluzione richiede il possesso di dati espressi mediante rappresentazioni strutturate. In effetti, poiché stanno diffondendosi nuovi linguaggi per la descrizione semantica dei dati sul web, i motori di ricerca si stanno attrezzando per riuscire a cogliere queste opportunità. Questo tema verrà ripreso nel paragrafo 4.
Un'ulteriore difficoltà pratica è rappresentata dalla modalità di visualizzazione dell'output, che costringe l'utente ad una scansione seriale dei risultati con ispezione diretta dei riassunti associati a ciascuno di essi. L'utente inoltre non ha la possibiltà di riordinare i risultati o selezionare "viste" d'interesse. La conseguenza è che per motivi di tempo e di costo tipicamente ci si sofferma solo sui dieci risultati offerti nella prima pagina di risposta, a fronte di migliaia di risultati reperiti, con una vistosa sottoutilizzazione delle capacità del sistema. Per superare questo problema possono essere utilizzati schemi interattivi di visualizzazione grafica dei risultati dei quali parleremo piu avanti.

Posizionamento

Per posizionamento s'intende l'adozione di un insieme di tecniche che hanno l'obiettivo di migliorare la posizione di un sito web nei risultati prodotti dai motori di ricerca. La letteratura su questo argomento (designato anche come ottimizzazione o marketing) è fiorente, con molti libri pubblicati. Il posizionamento non va confuso con lo *spam*, perché nelle tecniche di posizionamento si pone l'accento sui contenuti e sull'organizzazione del sito, e non sull'utilizzazione di trucchi per ingannare i motori di ricerca. A titolo di esempio, si riportano alcuni consigli che è facile trovare per siti commerciali:

✓ Inserire la parola chiave più importante nel marcatore "title" e all'inizio del *Body Text*.

✓ Includere nel *Body Text* un testo descrittivo dei prodotti o servizi offerti di alcune centinaia di parole, con una ripetizione non consecutiva di 4/5 volte della parola chiave.

✓ Inserire link alle altre pagine del sito e a siti esterni pertinenti.

✓ Non utilizzare testo nascosto, testo dello stesso colore dello sfondo, link nascosti, tecniche di *cloaking*, o reindirizzamenti automatici, perché possono essere facilmente scoperti, con conseguente penalizzazione del sito.

✓ Indicizzare il sito anche nella Open Directory.

3.3. Spam

Oltre ai problemi di calcolo e di presentazione, bisogna tenere conto di un importante fenomeno esogeno che ostacola il buon funzionamento dei motori di ricerca. Parliamo dello "spam", cioé del tentativo di influenzare con metodi scorretti il posizionamento delle pagine nei risultati. Questa è una novità sostanziale rispetto ai sistemi tradizionali di reperimento delle informazioni, come OPAC (*Online Public Access Catalogue*) e basi di dati bibliografiche, i cui dati di partenza hanno invece un elevato grado di veridicità e affidabilità.

Molte tecniche possono essere adoperate per confondere i motori di ricerca. Il fatto è che i principali meccanismi di ordinamento presentano un alto grado di vulnerabilità alle manipolazioni. Le più semplici sono l'utilizzazione di parole civetta nascoste nelle pagine e la creazione di fabbriche di link con contenuti artificiali, per alterare il contenuto delle pagine nei file indice. L'aneddotica su questo tipo di inganni è fiorente e ricorrente, come testimoniano i casi recenti in cui le interrogazioni "evil" e "miserable failure" producevano rispettivamente le pagine su Microsoft e su George Bush.

I punteggi di popolarità sono piu robusti, ma anche essi possono essere falsati manipolando la struttura dei link; un'analisi approfondita della vulnerabilità di PageRank è stata recentemente presentata in (Bianchini *et al.*, in pubblicazione). Contenuti civetta possono essere anche serviti direttamente nella fase di crawling, utilizzando in modo scorretto i normali meccanismi previsti per consentire di fornire al crawler versioni parallele che sono piu appropriate o informative delle pagine visualizzate dai browser (cloaking).

Naturalmente i motori di ricerca adottano una serie di contromisure, per esempio analizzando la distribuzione delle parole nei testi per verificare scostamenti sospetti, o verificando l'esistenza di concentrazioni anomale di link. Va da sé che nel mondo segreto dei motori di ricerca, le tecniche di anti-spam costituiscono, per definizione, la parte piu segreta, e la comunità accademica non se ne interessa più di tanto perché lo ritiene un problema essenzialmente commerciale. Il fenomeno dello spam è molto esteso perché estremamente redditizio, con un giro d'affari che Singhal, un ricercatore di Google, ha recentemente stimato in una decina di milioni di euro al giorno.

L'utilizzazione della parola spam ha una storia divertente, che evoca l'ottenimento di cose non richieste. Essa si riferisce ad una scenetta di una serie televisiva americana del 1972, Monty Python's Flying Circus, in cui i due protagonisti entrano in un bar e tentano invano di ordinare dei piatti senza SPAM (una nota carne di maiale in scatola), con la cameriera che contropropone piatti contenenti carne in scatola e un gruppo di avventori vestiti da vichinghi che urlano "spam, spam,.." ad ogni sua proposta.

4. LA PROSSIMA GENERAZIONE: DAI MOTORI DI RICERCA AI MOTORI DI RISPOSTA

Cosa ci riserva il futuro? Anche se l'argomento nel mondo industriale è ovviamente riservato, si possono interpretare i segnali provenienti dai laboratori di ricerca e osservare i risultati delle conferenze di settore, alimentati principalmente dagli studi compiuti nel mondo accademico. L'obiettivo di gran parte delle ricerche in corso, perseguito in vari modi, può essere ricondotto al tentativo di costrui-

re sistemi in grado di fornire direttamente la risposta desiderata, invece di limitarsi ad indicare un insieme di pagine che probabilmente la contengono.
Anticipando alcune delle conclusioni di questo articolo, la sensazione è che non vedremo un motore di risposta generalista. Già oggi sono disponibili servizi di ricerca specializzati che offrono risposte mirate in domini ristretti, come ad esempio gli acquisti in rete, le ricerche bibliografiche, e le notizie. Questa tendenza probabilmente si estenderà e si generalizzerà. Avremo motori di risposta abili nell'eseguire certi compiti di ricerca, oppure ricerche su alcuni tipi di dati, o ancora nel rispondere alle richieste di determinate tipologie di utenti. Cercheremo ora di precisare meglio le principali direzioni di sviluppo, indicando anche alcuni sistemi prototipali suscettibili di tramutarsi in innovazioni tecnologiche stabili nei motori di ricerca di prossima generazione.

4.1. Visualizzazione e raffinamento dei risultati

Per superare i limiti connessi alla visualizzazione dei risultati mediante lista di riassunti testuali, da vario tempo sono allo studio metodi di presentazione grafica che consentano di mostrare le informazioni associate a più pagine risultato in modo simultaneo, dando al contempo all'utente la possibilità di focalizzare l'attenzione su determinate parti dell'insieme dei risultati recuperati specificandone interattivamente proprietà e vincoli.
Varie rappresentazioni 2D e 3D sono state proposte, specialmente nelle conferenze CHI (*Computer-Human Interaction*) e SIGIR (*Special Interest Group on Information Retrieval*) dell'ACM (*Association for Computing Machinery*), ma la loro adozione da parte dei sistemi commerciali è ostacolata da una serie di problemi tecnici e da una relativa difficolta d'uso. Un buon compromesso fra la modalità di presentazione attuale e quelle grafiche più dirompenti è costituito dal "clustering" dei risultati, in cui i risultati vengono partizionati in un insieme di categorie organizzate gerarchicamente che riflettono i contenuti principali delle pagine recuperate. Questo metodo combina interrogazione diretta e "browsing" di una gerarchia: l'effetto per l'utente è quello di navigare attraverso un catalogo web costruito automaticamente sull'insieme dei risultati. I vantaggi sono molteplici: ci si può fare rapidamente un'idea dei contenuti delle pagine web che referenziano l'oggetto dell'interrogazione, si hanno a disposizione scorciatoie per trovare le pagine con le accezioni desiderate nel caso di interrogazioni ambigue, e si possono scoprire facimente informazioni non note.
Il sistema più famoso è Vivisimo (http://vivisimo.com), che fra l'altro non produce direttamente i risultati dei quali fa il clustering, ma li attinge da altri motori di ricerca (secondo una metafora zoologica diffusa, questi ultimi sarebbero gli erbivori delle informazioni e Vivisimo, così come altri motori specializzati nella post-elaborazione, i carnivori). Gli autori del presente articolo hanno sviluppato una variante che consente una navigazione più flessibile (genera un reticolo invece di un albero) ed è basata su un formalismo algebrico – i reticoli concettuali (concept lattices) – che produce cluster più giustificabili e comprensibili di quelli prodotti utilizzando metodi statistici. Questo approccio, descritto in [3], è stato implementato in un sistema prototipale denominato CREDO (*Conceptual REorganization of Documents*), che è disponibile in rete (http://credo.fub.it). CREDO ha una piccola comunità di utenti affezionati che noi chiamiamo scherzosamente credenti. È da notare che il clustering dei risultati è presente anche nell'agenda di Google, ed è probabile che lo vedremo presto in linea.
I metodi illustrati finora servono sostanzialmente ad analizzare più velocemente i risultati recuperati in seguito ad una singola interrogazione. La ricerca delle informazioni sul web è però un processo iterativo, oltre che interattivo, in cui l'esame delle pagine recuperate può spingere l'utente a formulare una nuova interrogazione i cui risultati soddisfino meglio il suo bisogno informativo. Come abbiamo già visto la scelta delle parole giuste non è sempre facile, specialmente nel caso di argomenti generici o ambigui che possono essere esaminati da varie angolazioni, potenzialmente sconosciute all'utente.

Poiché è ben noto che in alcuni casi è più facile riconoscere piuttosto che descrivere, alcuni sistemi, per esempio Altavista e Teoma, oltre ai risultati suggeriscono parole o concetti afferenti all'interrogazione che possono essere direttamente adoperati dall'utente per formulare una nuova richiesta, senza perdere troppo tempo ad esaminare i risultati della vecchia. I metodi impiegati spaziano dall'analisi del testo dei primi documenti recuperati o delle pagine che puntano ad essi, alla mappatura dei termini dell'interrogazione su cataloghi per il web o su reti semantiche, o ancora alla similarità con interrogazioni precedenti. Questo tipo di servizio è stato adoperato con alterna fortuna e convinzione da alcuni motori di ricerca nel recente passato; ora sembra che ci sia un interesse maggiore, giustificato anche dal miglioramento della qualità dei suggerimenti forniti e da una probabile maggiore propensione degli utenti alla interazione assistita.

4.2. Personalizzazione

La personalizzazione dei risultati è un vecchio cavallo di battaglia dell'accesso intelligente alle informazioni, per la quale è stata utilizzata anche la metafora degli agenti software. L'osservazione è che il risultato ottimale di una ricerca di informazioni dipende non solo dall'interrogazione, ma anche da chi la fa e per quale motivo. Direct Hit, uno dei primi motori a fornire risultati personalizzati, ha scoperto ad esempio che con un'interrogazione "flower" gli uomini vogliono spedire fiori, le donne ordinare semi e piante da giardino. Queste informazioni non vengono esplicitate dall'utente, ma possono essere ricostruite da varie sorgenti esterne.
Una possibilità, investigata presso i Google Labs, è quella di utilizzare un profilo d'utente e di graduare la sua influenza sui risultati ottenuti senza il profilo, lungo uno spettro di combinazioni che vanno dalla personalizzazione totale all'assenza di personalizzazione. Un approccio più ambizioso ed invasivo, studiato soprattutto in ambiente Microsoft, si basa sull'analisi dei comportamenti dell'utente desunti dalle informazioni presenti sul suo computer: le ricerche antecedenti che ha fatto, i file memorizzati, con chi scambia messaggi di posta e su quale argomento. Un terzo paradigma di calcolo è basato sulle reti sociali. Ci si affilia ad un gruppo, per esempio un gruppo di amici o di colleghi o persone con le quali si hanno interessi comuni e le preferenze espresse da ciascuno influenzano i risultati di ricerche simili condotte dagli altri. Un esempio di ricerca sociale delle informazioni è il sistema Eurekster (http://eurekster.com).
Il concetto di personalizzazione puo evolvere ulteriormente per abbracciare una serie di variabili ambientali legate al momento, al luogo e al tipo di dispositivo utilizzato. I recenti sviluppi dei servizi di reperimento delle informazioni basati sulla localizzazione geografica e le ricerche sull'accesso alle informazioni da terminale mobile convergono, insieme alla personalizzazione, verso la definizione di una ricerca *contestuale* delle informazioni che potrebbe diventare sempre piu pervasiva ed efficace.

4.3. Risposte fattuali

Uno dei filoni di ricerca più promettenti è il "question answering", cioè fornire risposte molto sintetiche a domande precise. Un esempio è "chi sono stati gli ultimi tre Presidenti del Consiglio?", oppure "quanto è alto il Monte Everest?". La risposta contiene di solito un'entità prestabilita (data, nome di persona, misura, luogo, quantità ecc.) e viene calcolata combinando metodi statistici ed elaborazione del linguaggio naturale, con la seconda che gioca un ruolo essenziale prima nel riconoscere il tipo d'interrogazione, e poi nell'estrarre il risultato esatto dalle pagine che contengono le parole dell'interrogazione.
Vari risultati sperimentali indicano che con i prototipi sviluppati già è possibile raggiungere percentuali elevate di risposte fattuali corrette. Uno di tali prototipi è stato messo in rete dall'università del Michigan (http://tangra.si.umich.edu/clair/NSIR/html/nsir.cgi).
Realizzazioni in corso presso vari laboratori di ricerca industriali, inclusi IBM e Microsoft, mirano sia a rendere più efficiente, robusto e preciso il processo di generazione delle risposte fattuali, sia ad estendere il paradigma "question anwering" a domande non fattuali.

Fra i motori di ricerca commerciali, Ask Jeeves (http://www.ask.com) è stato uno dei primi ad utilizzare tecniche di elaborazione del linguaggio naturale per aumentare la qualità dei risultati o, in taluni casi, per fornire direttamente la risposta. Per esempio, per l'interrogazione "Who won the best actor Oscar in 2003?", Ask Jeeves restituisce la frase "The 2003 Best Actor award was given to Sean Penn in Mystic River".

4.4. Reperimento di informazioni multimediali

Sul web sono presenti milioni di file non testuali (immagini, audio, video) che teoricamente si prestano ad essere indicizzati e quindi reperiti utilizzando varie caratteristiche multimediali di base, combinate a differenti livelli di astrazione [5], per esempio riconoscendo un oggetto visuale dagli elementi visuali atomici (pixel, linee e regioni) e dai loro attributi (dimensione, posizione, colore). Le applicazioni sono molteplici: informazione, intrattenimento, istruzione, turismo. Strumenti per il riconoscimento automatico dei contenuti sarebbero preziosi anche per il filtraggio dei siti, la prevenzione dei crimini e la tutela della proprietà intellettuale sui beni multimediali.

Anche se la ricerca sulla elaborazione dei contenuti audio e video dovrà fare ancora molta strada prima di arrivare a costruire motori di ricerca multimediale per il web di uso generico, analoghi ai motori di ricerca testuali, la direzione è stata tracciata e lo stato attuale della tecnologia già consente alcune applicazioni. L'esempio più noto è la ricerca d'immagini basata sul testo, che avviene utilizzando il testo circostante l'immagine nella pagina web. I motori di ricerca commerciali offrono questo servizio da tempo con buona efficacia. Il reperimento d'immagini può anche essere eseguito calcolando la similarità visuale fra un esempio fornito dall'utente e le immagini da selezionare, seppure bisogna osservare che queste ultime tecniche hanno una utilizzazione pratica ancora limitata. Probabilmente una integrazione fra tecniche basate sul testo e tecniche visuali consentirebbe in alcuni domini di bilanciare meglio l'aumento di precisione e il contenimento dei tempi di calcolo. Vanno in questa direzione alcune ricerche che mirano a raffinare i risultati di un motore di ricerca di immagini con una post-elaborazione visuale basata su immagini campione.

Anche per l'analisi e il riconoscimento del parlato sono stati fatti grandi progressi. Oggi è possibile estrarre con buona accuratezza le parole da una traccia audio ed utilizzarle per accedere direttamente al segmento che le contiene, in risposta ad una interrogazione specificata in forma testuale (speech retrieval). Il grande pubblico si accorse di quest'applicazione già ai tempi dell'affare Lewinski, con milioni di persone che utilizzarono un servizio disponibile gratuitamente in rete per accedere direttamente ai momenti salienti della deposizione di Clinton, senza doverla ascoltare interamente. Per tornare ai giorni nostri, Speechbot (http://speechbot.research.compaq.com) è un prototipo di speech retrieval sviluppato dagli HP Labs che lavora su una base di trasmissioni radio piuttosto grande (circa 15000 h).

L'idea probabilmente più appassionante e ambiziosa del "multimedia information retrieval" è l'indicizzazione automatica delle sequenze video. Il progetto Informedia (http://www.informedia.cs.cmu.edu), iniziato a metà degli anni 90, è stato il pioniere di queste ricerche. Esso ha cercato di ridurre l'enorme complessità del problema agendo inizialmente sulla trascrizione automatica della traccia audio e il suo allineamento temporale col video, con l'obiettivo di attaccare etichette testuali ai frame candidati ad essere reperiti, ed integrando successivamente questo approccio con varie modalità di sintesi visuale. Lo studio dei metodi per la comprensione e l'indicizzazione automatica del video continuerà a vari livelli, ma mentre la complessità dell'elaborazione diretta del contenuto del segnale multimediale fa ritenere che l'interesse maggiore per questo tipo di applicazioni si concentrerà soprattutto presso le aziende che producono e trasmettono contenuti video, l'attenzione dei motori di ricerca sembra rivolgersi al reperimento di frammenti video basato sulle didascalie di accompagnamento delle trasmissioni o sulla trascrizione della traccia audio, in modo concettualmente analogo a quanto avviene per il reperimento d'immagini basato sul testo associato. Blinkx (http://www.blinkx.tv) e Google Video (http://video.google.com) sono

due motori di ricerca sperimentali per trasmissioni video che dimostrano la fattibilità di questo approccio.

4.5. Motori di ricerca per il web semantico

Nella visione del Web Semantico, le pagine vengono arricchite con annotazioni interpretabili dal computer che catturano il significato del contenuto delle pagine stesse. I vantaggi per l'accesso alle informazioni sono evidenti, in linea di principio. Utilizzando linguaggi come RDF per descrivere il contenuto dei dati e ontologie formali per specificare concetti e regole di derivazione, è possibile trovare risposte che sono collegate *logicamente* alle interrogazioni, superando i limiti sintattici dei motori di ricerca tradizionali.

Attualmente i marcatori semantici non vengono utilizzati dai motori di ricerca nel calcolo dei risultati. Poiché è probabile che vedremo un numero crescente di pagine descritte in modo misto (testo + annotazioni semantiche), il trattamento esplicito di entrambi gli aspetti, con l'integrazione di reperimento testuale ed inferenza logica, rappresenta una sfida ma anche una opportunità per gli attuali motori di ricerca.

Una situazione di complessità intermedia è costituita dalla gestione delle pagine descritte in XML, in cui lo schema, la tipizzazione dei dati e la prossimità strutturale possono essere adoperate a fini semantici. La strutturazione può consentire ad esempio di reperire l'elemento informativo di granularità appropriata, che risponda all'interrogazione in modo esaustivo ma che sia anche sufficientemente specifico. Il vantaggio di considerare pagine XML è anche che si tratta di uno standard di descrizione estremamente diffuso per la pubblicazione e lo scambio dei dati.

Il reperimento di informazioni da dati XML è l'oggetto di INEX (http://inex.is.informatik.uni-duisburg.de:2004), un programma pluriennale di ricerca finanziato in parte dalla Comunità Europea. L'obiettivo di INEX è lo sviluppo di tecniche di reperimento delle informazioni che integrino contenuto e struttura, cercando di fondere i metodi di "information retrieval" coi linguaggi di interrogazione per basi di dati strutturate. Per la sperimentazione e valutazione dei prototipi viene utilizzata una collezione di pubblicazioni scientifiche della IEEE descritte in linguaggio XML.

5. CONCLUSIONE

Anche se in un settore caratterizzato da trasformazioni così rapide è sempre difficile fare previsioni, ci sembra di poter dire che la tecnologia della ricerca delle informazioni sul web sarà sempre più caratterizzata, usando un'espressione del linguaggio psicoanalitico, da un processo di "differenziazione e integrazione". La differenziazione consentirà di ottenere servizi ad elevate prestazioni in settori verticali mirati, per rispondere a bisogni specifici utilizzando particolari tipi di dati. Di questo già vediamo vari segnali e le prime manifestazioni, come abbiamo cercato di spiegare nell'articolo. Alle accresciute capacità di differenziazione dovrà necessariamente accompagnarsi un grosso sforzo di integrazione (di cui gli attuali meta-motori di ricerca costituiscono un aspetto modesto), con l'obiettivo di presentare all'utente in modo sintetico e possibilmente adattativo i vari "canali" di risposta che possono essere esaminati. L'integrazione di insiemi di risposta differenti può essere un aspetto cruciale per il miglioramento dell'esperienza finale dell'utente, specialmente se emergeranno anche nuovi paradigmi di presentazione delle informazioni più diretti e familiari, per esempio con l'adozione di metafore legate all'attività lavorativa e allo stile di vita degli utenti stessi.

Bibliografia

[1] Baeza-Yates R., Ribeiro-Neto B.: *Modern Information Retrieval*. Addison Wesley, 1999.

[2] Bianchini M., Gori M., Scarselli F.: *Inside PageRank*. To appear in ACM Transactions on Internet Technology (TOIT).

[3] Carpineto C., Romano G.: *Concept Data Analysis: Theory and Applications*. John Wiley & Sons 2004.

[4] Chakrabarti S.: *Mining the Web*. Morgan Kaufmann, 2003.

[5] Maybury M.: *Intelligent Multimedia Information Retrieval*. The MIT Press, 1997.

[6] Salton G., McGill M.: *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[7] Sherman C., Price G.: *The Invisible Web*. CyberAge Books, 2001.

[8] Yang K.: *Combining Text-, Link-, and Classification-based Retrieval Methods to Enhance Information Discovery on the Web*. Doctoral Dissertation, University of North Carolina, 2002.

Per approfondire

Un elenco aggiornato di risorse:
http://dmoz.org/Computers/Internet/Searching/Search_Engines/

Un sito di riferimento per il marketing:
http://searchenginewatch.com

Gli strumenti di search:
http://searchtools.com

Google Labs:
http://labs.google.com

Tutto sui Crawlers:
http://www.robotstxt.org

Text REtrieval Conference (TREC):
http://trec.nist.gov

Cross Language Evaluation Forum:
http://www.clef-campaign.org

Claudio Carpineto è ricercatore presso la Fondazione Ugo Bordoni. È autore o co-autore di circa 70 pubblicazioni nei settori dell'intelligenza artificiale, dei sistemi informativi e dell'"information retrieval", incluso il recente libro *Concept Data Analysis: Theory and Applications* (John Wiley & Sons, 2004). Le sue ricerche attuali riguardano lo sviluppo di motori di ricerca intelligenti per internet e intranet.
carpinet@fub.it

Giovanni Romano è ricercatore presso la Fondazione Ugo Bordoni. Ha svolto ricerche nel campo dell'intelligenza artificiale, dell'interazione uomo-calcolatore e dell'"information retrieval". Attualmente lavora allo sviluppo di strumenti software per la costruzione di motori di ricerca. È co-autore del libro *Concept Data Analysis: Theory and Applications* (John Wiley & Sons, 2004).
romano@fub.it