[Pocherighe #7] Scrivere per farsi trovare nei motori di ricerca

Paolo Ferragina, professore associato presso il Dipartimento di Informatica dell'Università di Pisa e coordinatore per l'area informatica del centro di ricerca SIGNUM della Scuola Normale Superiore di Pisa, ci svela quali sono i segreti per creare siti web che siano raggiungibili dai motori di ricerca e quindi dai lettori.

di Mariella Minna

Perchè la ricerca sul web è difficile?

Molte delle difficoltà relative alla ricerca sul web sono legate alla struttura del web stesso.
Il web è una rete "sconfinata" di documenti variamente interconnessi tra loro, il cui numero è cresciuto in modo esponenziale: se ne contavano infatti circa 110 mila nel 1994, e oggi se ne contano più di 8 miliardi, stando alla dimensione di Google. Questi documenti sono anche fortemente eterogenei, per lo stile variegato e a volte malizioso con il quale gli utenti compongono le loro pagine (commerciali, soprattutto) per risultare più rilevanti nei motori di ricerca. Le lingue utilizzate sono più di un centinaio: tra queste, quelle asiatiche diventano sempre più preponderanti, rendendo ancora più sofisticata l'analisi dei documenti e l'estrazione delle informazioni da essi. A tutto ciò si aggiunge anche il fatto che il web è dinamico: ciascuno scrive le proprie pagine e le modifica di continuo. Alcuni studi recenti hanno dimostrato, sia pure su un campione di pagine molto ridotto, che tale dinamicità è significativa: in un anno sopravvivono solo il 40% delle pagine e solo il 20% dei link.

Puoi raccontarci a grandi linee la storia dei motori di ricerca?

Quali sono le caratteristiche che differenziano le tre generazioni fondamentali dei motori di ricerca? Altavista era un motore di ricerca incentrato sul contenuto testuale della pagina. Questo approccio produsse risultati eccellenti, finché i documenti disponibili sul web erano pochi e di elevata qualità. Nel 1998 nacque Google, la cui fortuna è essenzialmente legata a un meccanismo di rilevanza che teneva conto non soltanto del contenuto testuale dei documenti, ma anche della loro interconnessione mediante hyper-link e dei commenti collegati a essi (*anchor text*). La rilevanza di una pagina dipendeva dunque dal suo contenuto, da ciò che altre pagine "dicevano" di lei e, soprattutto, da quanto queste erano rilevanti. La generazione corrente di motori di ricerca annovera molti attori tra i quali spiccano Google, Yahoo, AskJeeves e MSN. Questi motori di ricerca sono oggi un'evoluzione molto sofisticata del Google del 1998. Consentono non soltanto di recuperare documenti tramite le parole chiave inserite dagli utenti, ma anche di offrire suggerimenti, di affinare le ricerche e personalizzarne i risultati, di cercare *topics* in collezioni selezionate (*directory*), e in genere, di scoprire nuove informazioni, le più personalizzate possibili.

Come progettare pagine web autorevoli?

Quando si progetta una pagina web occorre considerare: le parole che un visitatore interessato alla pagina scriverà per formulare la sua interrogazione di ricerca, e le tecniche (note) di cui un motore di ricerca si avvale per il *ranking* delle pagine in risposta a una interrogazione.
Il titolo della pagina deve essere breve e deve contenere tutte le parole chiave che la descrivono e che potenziali clienti utilizzeranno per cercarla. Esistono poi i cosiddetti *metatag*, che non vengono visualizzati, ma che permettono di introdurre una descrizione della pagina o delle parole chiave. È importante indicare qual è il linguaggio usato all'interno della pagina con un altro metatag, così che il motore sappia quale analizzatore lessicale adottare per estrarre le parole da indicizzare.
Il primo paragrafo gioca un ruolo determinante. Sebbene i motori di ricerca indicizzino l'intera pagina web, la posizione in testa alla pagina ha per alcuni motori un peso maggiore. È opportuno che le parole chiave nella pagina figurino con una certa frequenza. È anche utile che altre pagine web puntino alla nostra pagina, e che questi link contengano un testo che riporti le parole chiave che a noi interessano (*anchor text*). Tanto più questa pagina puntante è importante, tanto maggiore rilevanza acquisirà la nostra pagina, secondo il criterio adottato da Google. Pertanto, facciamoci puntare da pagine autorevoli: università, siti governativi, portali...

e facciamo in modo che questi link riportino una descrizione adeguata della nostra pagina web.

Anche la descrizione di immagini è importante?

Sì, un altro aspetto da curare con attenzione è infatti quello legato alla descrizione delle figure. Nei dati che permettono di caricare le immagini, il tag ALT contiene la descrizione della figura. I motori non riescono a fare un'analisi delle immagini e si basano su ciò che ha scritto l'utente che le ha caricate. Oppure su quanto scritto da altri, che puntano alle stesse immagini. Se la descrizione è adeguata, l'immagine verrà trovata. E poiché oggi tutti i motori di ricerca offrono strumenti per cercare in collezioni di immagini, la nostra immagine può essere ricercata da altri.

Come stabilire se abbiamo progettato una pagina web autorevole?

Alcuni strumenti ci consentono di valutare la nostra bravura nel comporre una pagina web autorevole. Utile adottarli, dunque, alla fine della nostra progettazione. Per esempio, Google ha recentemente messo a disposizione uno strumento, *Sitemaps*, grazie al quale è possibile valutare la rilevanza della propria pagina web. Con *Seekbot,* invece, basta specificare l'indirizzo di una pagina web per sapere se la sua struttura è adeguata per l'indicizzazione sui motori di ricerca. Infine, *Word Tracker* e *Inventory* di Overture sono utili "suggeritori" per le parole chiave da adottare nelle nostre pagine web. Infatti, questi ci consentono di scoprire quali interrogazioni sono state frequentemente eseguite negli ultimi mesi dagli utenti su un certoargomento. In tal senso, è chi scrive a doversi adeguare al linguaggio utilizzato dagli utenti e non viceversa.

Qual è il futuro dei motori di ricerca?

La personalizzazione e la localizzazione geografica delle informazioni risulteranno sempre più cruciali, alla luce soprattutto dell'esplosione dell'informazione in formato digitale, e dell'uso di tecnologie *mobile*, quali palmari e *smart-phone*. Informazioni fresche, rilevanti e personalizzate sulla base della nostra collocazione geografica (GPS) e dei "nostri bisogni". Un sogno? Forse. E forse, presto, una realtà.
Ci si muove inoltre verso un'integrazione totale dei media. Non più solo testo. Negli ultimi anni si è parlato di immagini e foto, ora si parla di video e di audio.
Nell'ultima conferenza CES (Consumer Electronics Show) 2006 di Las Vegas, sono intervenuti Google e Yahoo mostrando i nuovi prodotti relativi al video e al *mobile*. Google ha presentato un progetto in collaborazione con la CBS: la creazione di un immenso archivio di video amatoriali e commerciali, sui quali l'utente può effettuare le ricerche usando il solito motore, ma con alcune caratteristiche specifiche per i video, e poi acquistare i video che gli interessano. Yahoo si sta muovendo invece verso la *TV on-demand*. Intende infatti realizzare un portale di trasmissioni televisive, usufruibili da ogni angolo della terra attraverso PC, laptop e telefonini di ultima generazione. Ovviamente, le ricerche sui video sono guidate da informazioni aggiuntive, introdotte ad-hoc sui filmati da esperti o produttori (i cosiddetti *meta-dati*). Quello che però si vorrebbe fare in un prossimo futuro è progettare un *multimedia search-engine*, ossia un motore di ricerca in grado di cercare i video servendosi di informazioni estratte automaticamente dal loro audio o dal filmato, estendendo così enormemente le capacità di recupero delle informazioni in essi contenute.

****

Pocherighe è la newsletter della Palestra della scrittura, fondata da Alessandro Lucchini e Paolo Carmassi.