

RICERCHE SU INTERNET

versione 1

di

Gianluca Sironi

(pagina inserita per una corretta gestione delle stampe in fronte retro)

Sommario

1	Obiettivi del documento.....	3
2	FDL (Free Documentation License).....	4
3	Motori di ricerca.....	5
3.1	Analisi, catalogazione, risposta.....	6
4	Google.....	8
4.1	Ricerche semplici.....	10
4.2	Tabella riassuntiva degli operatori.....	12
4.3	Ricerche avanzate.....	13
4.4	Interpretazione dei risultati.....	15
4.5	Problemi legati ai risultati delle ricerche.....	17
4.6	Google Directory.....	18
5	Riferimenti.....	19

(pagina inserita per una corretta gestione delle stampe in fronte retro)

1 Obiettivi del documento

Scopo del presente documento è fornire indicazioni su come effettuare in modo semplice ma efficace ricerche di informazioni, testi, documenti, immagini, ... su Internet.

Si assume che le ricerche di informazioni vengano effettuate utilizzando un browser, quali Microsoft Internet Explorer, Mozilla Firefox, Opera, ecc

Nel documento verranno dati alcuni cenni sui motori di ricerca e sulle tecniche utilizzabili per ricerche avanzate.

Nel documento non viene affrontato il tema della ricerca di informazioni nei servizi di Directory (come ad esempio Yahoo!) che sono strutturati con uno scopo diverso: l'organizzazione delle informazioni è fatta attraverso una gerarchia strutturata, una struttura ad albero, suddivisa per aree tematiche, dove scegliendo un tema specifico si scende ad un livello più profondo poter raffinare il livello della ricerca; accade spesso che i due sistemi di ricerca (motore di ricerca e servizio di directory) si trovino associati in un unico portale.

Nel documento verrà a fatto riferimento esplicitamente al noto motore di ricerca **Google**, che è diventato lo standard de facto nell'ambito dei motori di ricerca.

Obiettivo principale del documento è quindi illustrare le tecniche di semplice utilizzo di Google, come ad esempio l'utilizzo di operatori booleani, per raffinare in modo semplice ma efficace le ricerche di testi, documenti, immagini, ecc. ; i principi e le tecniche illustrati possono comunque essere facilmente trasportati su altri motori di ricerca.

Obiettivi secondari sono dare alcune informazioni sui motori di ricerca, sul posizionamento e sulle ricerche complesse.

2 FDL (Free Documentation License)

Questo documento è rilasciato sotto licenza FDL (Free Documentation License)

Copyright © 2006 Gianluca Sironi

Via Stradella, 7 – 20129 Milano MI

gianluca.sironi @ gmail.com

E' garantito il permesso di copiare, distribuire e/o modificare questo documento seguendo i termini della Licenza per Documentazione Libera GNU, Versione 1.2 oppure ogni versione successiva pubblicata dalla Free Software Foundation;

senza Sezioni Non Modificabili

senza Testi Copertina

senza Testi di Retro Copertina

mantenendo intatte le indicazioni di Copyright ©

la versione originale della GNU FDL è disponibile su: <http://www.gnu.org/copyleft/fdl.html>



3 Motori di ricerca

La quantità di informazioni presenti su Internet è enorme ed in costante crescita. Internet è diventato sia la più grande base dati dove inserire e reperire le informazioni sia il punto di accesso, lo strumento privilegiato, per la raccolta delle informazioni stesse.

Insieme alla ricchezza e complessità delle informazioni sono sorte anche le difficoltà nel cercare contenuti particolari ed informazioni specifiche. Le informazioni oggetto di ricerca sono tipicamente documenti, immagini, notizie, gruppi di discussione, ecc ...

In ausilio a chi ricerca informazioni su Internet sono disponibili diversi strumenti, tra cui i motori di ricerca. Un motore di ricerca è in sostanza un sistema automatico complesso la cui funzione principale è analizzare un insieme di dati e di informazioni, sovente raccolti dal motore di ricerca stesso, e restituire un "SERP" (Search Engine Result Pages, pagine dei risultati dei motori di ricerca), un indice dei contenuti disponibili classificati in base a formule matematiche che riportano il "grado di rilevanza" in relazione ad una determinata chiave di ricerca (keyword).

Esistono numerosi motori di ricerca attivi su Internet. Il motore di ricerca attualmente più utilizzato su scala mondiale è Google; altri motori di ricerca molto utilizzati sono MSN (il motore di ricerca di Microsoft) e Yahoo!. In Italia, oltre a Google, MSN e Yahoo!, i motori di ricerca italiani maggiormente utilizzati sono Libero e Virgilio.

La maggior parte dei motori di ricerca che opera sul web è gestito da compagnie private che utilizzano algoritmi proprietari e i database di SERP sono tenuti segreti.

Oltre ai risultati delle ricerche "neutre" vi sono risultati "sponsorizzati": spesso i motori di ricerca sul web forniscono anche risultati sponsorizzati: mostrano in maggiore evidenza nelle SERP (Search Engine Result Pages) i riferimenti a siti web di aziende che pagano per risultare tra i primi risultati quando si cercano termini (keyword, parole chiave) che sono in relazione all'ambito di competenza dell'azienda sponsor stessa.

Le ricerche avvengono semplicemente inserendo delle keyword (parole chiave della ricerca) nelle caselle di testo dei motori di ricerca. Le ricerche possono essere facilmente raffinate, e la possibilità di raffinare una ricerca varia da motore a motore, ma quasi tutti permettono di utilizzare operatori booleani (AND, NOT, OR, ..): ad esempio è possibile cercare "*Ganimede AND satellite NOT figlio*" per cercare informazioni su Ganimede inteso come satellite (del pianeta Giove) e non come figura mitologica (il minore dei figli del re di Troia)

Su Google (e su altri motori di ricerca) è possibile raffinare la ricerca a seconda della localizzazione (a seconda della lingua del documento), delle parole o delle frasi presenti o assenti, del formato dei file (file testo, file di Word, PDF, PostScript ecc.), della data di ultimo aggiornamento, e diversi altri criteri ancora.

La quantità di informazioni disponibili è enorme, ma è noto che i motori di ricerca tengono conto solo di una piccola parte di informazioni realmente presenti su internet, che spesso sono presenti in archivi dati aziendali o di altre basi di dati a cui si accede solo attraverso query specifiche e non attraverso il lavoro dei crawler o degli spider.

Va osservato che il lavoro dello spider si limita alla cattura del testo presente nelle pagine, ma in caso di query non è in grado di “scegliere” quali parole inserire né quali siano particolarmente rilevanti; e non può neanche inserire parole a caso nei campi di query. In questo modo molte pagine di informazioni in profondità (“deep web”) rimangono escluse dai SERP dei motori di ricerca.

Con il passare del tempo questo “deep web” diventa sempre più ricco e più profondo, rimanendo sempre poco (o per nulla) accessibile ai motori di ricerca; sono in corso modifiche al funzionamento dei motori ed è lecito prevedere che nel giro di pochi anni i motori di ricerca baseranno le proprie tecnologie oltre che sull'analisi quantitativa dei contenuti (come attualmente viene effettuato con le parole in sé) sempre più sull'analisi qualitativa (il “senso” delle parole).

I motori di ricerca saranno, ad esempio, sempre più in grado di distinguere il senso della parola “pesca” a seconda di quale sia il contesto in cui la parola è contenuta (capire se sia il frutto, la disciplina sportiva, o altro). Per muoversi in questa direzione Google ha acquisito Oingo (un tempo noto come “il motore dei concetti”) e la tecnologia dell'azienda che lo creò, la Applied Semantics.

3.1 Analisi, catalogazione, risposta

Le attività di un motore di ricerca possono essere sintetizzate in tre aree: analisi, catalogazione e risposta.

Analisi

Per analizzare il web i motori di ricerca utilizzano dei programmi detti crawler (o spider), che si occupano di visitare in modo automatico gli URI (Uniform Resource Indicator) contenuti nel database (a partire quindi da una lista iniziale di siti web) e seguire i successivi URI che trovano all'interno dei documenti analizzati, inserendo di volta in volta nel database tutte le informazioni “sensibili” della pagina (il contenuto testuale, varie informazioni su di essa come la data di ultimo aggiornamento, ed altro).

Catalogazione

Dopo l'analisi delle pagine, a seconda di criteri che variano da motore a motore, alcune di esse vengono inserite ed indicizzate nel database e nell'indice del motore di ricerca.

La parte testuale archiviata durante la fase di analisi viene successivamente analizzata per fornire le risposte alle ricerche degli utenti. Molti motori di ricerca sul web rendono anche disponibile una copia dei dati testuali di ogni pagina archiviata per quando la risorsa originale sia irraggiungibile. (questa tipo di funzione è detta copia cache) .

Risposta

I motori di ricerca rispondono alle richieste degli utenti (keyword) elencando i siti in ordine di rilevanza rispetto alla richiesta ricevuta.

Per stabilire la rilevanza di un sito vengono cercate nel database quei documenti che contengono la parola chiave inserita dall'utente, dopodiché ogni motore di ricerca sfrutta propri algoritmi per classificare le pagine, controllando, per esempio, quante volte le parole chiave vengono ripetute, quanti link riceve quel documento, in quali punti della pagine sono poste le parole chiave, quanti siti del database contengono link verso quella pagina, o quante volte un utente ha visitato quel sito dopo una ricerca.

4 Google

Tra i motori di ricerca Google occupa una posizione di particolare rilievo.

La missione di Google è "organize the world's information and make it universally accessible and useful".

Google come motore di ricerca non si limita ad indicizzare e catalogare le pagine presenti sul web, ma si occupa anche di immagini, newsgroup, notizie, ecc .. oltre a mantenere una copia cache di tutte le pagine che conosce; attualmente l'indice di Google comprende più di otto miliardi di pagine web.

Google è riconosciuto come il più grande ed affidabile tra i motori di ricerca, occupandosi di circa l'80% di tutte le ricerche effettuate su internet attraverso il suo sito.

La popolarità di Google è tale che in inglese è nato il verbo "*to google*" col significato di "fare una ricerca sul web".

Relativamente alle fase di catalogazione, Google si basa principalmente su un algoritmo chiamato "PageRank" per giudicare l'importanza delle pagine che corrispondono ad una ricerca. (il nome Page non deriva da "pagina" ma da Larry Page, uno dei due fondatori di Google).

La formula del PageRank viene applicata attraverso un algoritmo che calcola in modo iterativo un "indice di merito" per ogni pagina, basandosi sulla media ponderata dei PageRank delle pagine dove compare un link a quella pagina; l'algoritmo si avvicina al concetto umano di "importanza", cercando di fornire sia una analisi quantitativa sia un inizio di analisi qualitativa.

Oltre a PageRank, Google usa anche molti altri criteri segreti per determinare l'importanza di una pagina e la sua rilevanza con la ricerca effettuata.

Il motore di ricerca Google è stato creato nella seconda metà degli anni '90 da Larry Page e Sergey Brin, dell'Università di Stanford in California; i due, dopo aver sviluppato una teoria secondo la quale un motore di ricerca basato sull'analisi matematica delle relazioni tra siti web avrebbe prodotto risultati migliori rispetto alle tecniche empiriche usate sino ad allora, decisero di approfondire la loro teoria e posero le basi per il loro motore di ricerca

Il successo di Google come motore di ricerca ha provocato anche delle modifiche nelle impostazioni dei siti web: da quando infatti Google è diventato uno dei motori di ricerca più popolari, molti webmaster hanno cominciato a seguire e cercare di spiegare i cambiamenti nel posizionamento (*rank*) del loro sito, di conseguenza cercando di comprendere il sistema di catalogazione hanno iniziato a modificare il contenuto dei siti in modo da avanzare nel rank..

Uno dei probabili motivi di successo di Google, oltre alla efficacia nei risultati delle ricerche, è stata la sua interfaccia molto spartana. Nella figura di seguito è riportata la pagina iniziale che si presenta digitando l'indirizzo <http://www.google.it> nel browser



Si possono semplicemente inserire parole chiave (keyword) e far partire la ricerca utilizzando il tasto “Cerca con Google”; il tasto “Mi sento fortunato” semplicemente reindirizza direttamente al sito web del primo risultato in ordine della ricerca.

Prima di affrontare l'utilizzo dei primi semplici operatori di ricerca, facciamo un breve cenno agli altri elementi presenti nella pagina di di Google.

Di default la ricerca è su tutto “il Web”, ma si può limitare alle sole “pagine in italiano” o per provenienza dei siti (in questo caso “pagine provenienti da: Italia”. Sopra la casella dove inserire i termini della ricerca vi sono i link per le ricerche di “Immagini”, nei “Gruppi”, nelle “News”, ecc ... Ognuno di questi link fa riferimento ad aree di ricerca più specifiche.

4.1 Ricerche semplici

Google visualizza solo le pagine di risultati che contengono **tutte** le keyword (parole chiave) ricercate, per ottimizzare o restringere le ricerche è sufficiente aggiungere altre parole a quelle già inserite. Se si aggiungono altre parole alla ricerca precedente, viene visualizzato solo un sottoinsieme delle pagine visualizzate dopo la prima ricerca contenente anche i termini di ricerca aggiunti.

E' possibile per raffinare velocemente i risultati di una ricerca in modo semplice attraverso l'utilizzo di operatori.

Va premesso che Google non fa distinzione tra lettere minuscole e maiuscole poiché considera tutte le lettere come minuscole. Ad esempio, ricercando "google", "GOOGLE" e "GoOgLE" si ottengono sempre gli stessi risultati.

Operatore AND

L'operatore AND è automatico in Google. In sostanza Google cerca tutte le keyword (parole chiave) inserite. In effetti Google fa di più che l'operazione di AND: cerca infatti tutte le parole chiave inserite dando maggiore importanza a quelle pagine che contengono tutte le parole nell'ordine in cui sono state inserite.

Una query può avere una lunghezza massima di 10 parole. Keyword inserite oltre il limite di dieci non verranno prese in considerazione nella ricerca.

Esclusione di parole da una ricerca (-)

Per escludere una parola dalla ricerca, è sufficiente inserire il segno meno ("-") davanti al termine da escludere (ricordarsi di inserire anche uno spazio prima del segno meno).

Ricerca di frasi (" ")

Per ricercare le frasi con Google, è sufficiente racchiuderle tra virgolette. In tutti i documenti trovati, la frase ricercata risulterà racchiusa tra virgolette ("come in questo esempio") (a meno che tali parole non siano parole comuni o parole comuni speciali che richiedano l'uso del segno "+"; per informazioni più dettagliate fare clic qui). L'uso delle virgolette è particolarmente utile se si ricercano detti famosi o nomi specifici.

Ricerca di parole chiave in sequenza (+)

Per cercare testi dove compaiono in sequenza (una subito dopo l'altra) due o più parole chiave occorre separarle col segno "+": ad esempio l'istruzione "wikipedia+google" fa comparire le pagine in cui compare la parola wikipedia subito seguita da google.

Nelle posizioni più basse (quelle con minore ranking) si trovano le pagine che risultano anche da una ricerca senza l'operatore "+", che non è esattamente l'operatore booleano "AND", in quanto oltre a cercare entrambi i termini nel testo, li cerca nell'ordine indicato nella ricerca.

Ricerche per sinonimi (~)

E' possibile utilizzare l'operatore "~" ed effettuare la ricerca per la keyword (parola chiave) o un suo sinonimo

Ricerche di definizioni (define:)

Per cercare la definizione di un termine o il significato di un acronimo basta digitare nel campo di ricerca di Google "define: termine" dove "termine" indica la parola da cercare.

Ricerche all'interno di un sito specifico (site:)

Per effettuare una ricerca di parole chiave all'interno di un sito occorre digitare l'istruzione: "site:nomesito.com" uno spazio bianco e le parole chiave (o viceversa la parola chiave uno spazio e l'istruzione site:nomesito.com).

Ad esempio È anche possibile cercare contenuti presenti in un determinato sito web, indicandolo esplicitamente ("site:") nelle chiavi di ricerca: ad esempio "*Ganimede site:nasa.gov*" cercherà le informazioni su Ganimede presenti sul sito della NASA.

Un altro modo di effettuare lo stesso tipo di ricerca è scrivere nella barra degli indirizzi: www.google.it/custom?domain=www.nomesito. Compare una barra di Google con un'opzione selezionabile per la ricerca all'interno del sito richiesto

Conversione di valuta

Google è in grado di fare anche la conversione di valuta. Digitando ad esempio "1€ to USD" Google ritornerà il valore di un euro in dollari americani. Funziona con tutte le valute e con le unità di misura di tutto il mondo.

Stop Words

Google ignora le parole ed i caratteri comuni, detti "stop words", scartando automaticamente termini come "http" e ".com" nonché alcune singole cifre e lettere, dal momento che questi termini non aiutano a restringere il campo di ricerca ma anzi contribuiscono a rallentare i tempi di ricerca.

4.2 Tabella riassuntiva degli operatori

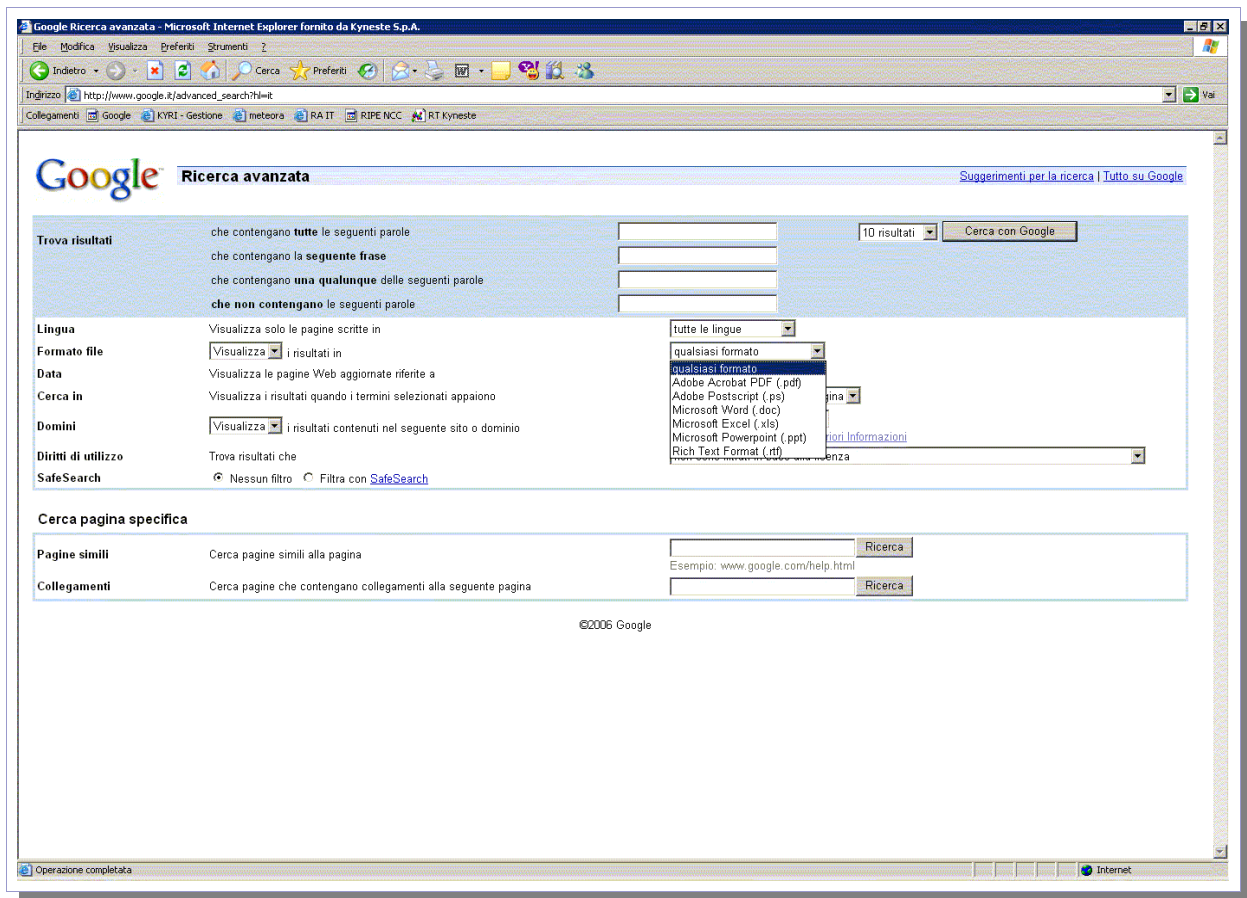
Nella tabella seguente sono riassunti i principali operatori di ricerca semplice

Notation	Find result	Example
<u><i>terms1 terms2</i></u>	with both <i>term1</i> and <i>term2</i>	[carry-on luggage]
<u><i>term1 OR term2</i></u> <u><i>term1 term2</i></u>	with either <i>term1</i> or <i>term2</i> or both	[Tahiti OR Hawaii] [Tahiti Hawaii]
<u><i>+term</i></u>	with <i>term</i> (The + operator is typically used in front of stop words that Google would otherwise ignore or when you want Google to return only pages that match your search terms exactly. However, the + operator can be used on any terms.)	[+i spy]
<u><i>-term</i></u>	without <i>term</i>	[twins minnesota -baseball]
<u><i>~term</i></u>	with <i>term</i> or one of its synonyms (currently supported on Web and Directory search)	[google ~guide]
<u><i>number1..number2</i></u>	with a number in the specified range	[annual report 2000..2003]
<u><i>"phrase"</i></u>	with the exact <i>phrase</i> , a proper name, or a set of words in a specific order	["I have a dream"] ["Rio de Janeiro"]

4.3 Ricerche avanzate

Oltre a poter impostare delle preferenze, a destra della casella di inserimento dei termini della ricerca vi è il link alla pagina di ricerca avanzata.

Nella figura di seguito la pagina di ricerca avanzata



Nella pagina di ricerca avanzata è possibile inserire diversi criteri per affinare la ricerca:

Lingua

E' possibile utilizzando il menu a tendina selezionare la localizzazione (ad esempio "italiano") o lasciare il parametro di ricerca di default "tutte le lingue"

Formato file

Utilizzando il menu a tendina è possibile selezionare un tipo di documento: Google è in grado di effettuare ricerche oltre che per pagine HTML anche per PDF, documenti Word .DOC, file Excel .XLS, presentazioni PowerPoint .PPT, ecc ..

Data aggiornamento

E' possibile, sempre utilizzando l'apposito menu a tendina, indicare il periodo temporale dove ricercare l'aggiornamento del risultato di una ricerca, ad esempio negli ultimi 3 mesi

Ricerca in una parte specifica della pagina

E' possibile selezionare attraverso il menu a tendina se i risultati della ricerca devono appartenere al titolo della pagina, al corpo, nei link, ecc ...

Tipologia di licenza dei risultati

E' possibile selezionare, sempre attraverso un menu a tendina, se i risultati devono essere filtrati in base alla licenza, se possono essere utilizzati gratuitamente, condivisi, se hanno scopo commerciale, ecc.

SafeSearch

E' possibile filtrare i risultati delle ricerche in modo che i siti con contenuto per adulti non vengano inclusi nei risultati delle ricerche. La funzione SafeSearch di Google cerca siti con contenuti esplicitamente sessuali e li elimina dai risultati della ricerca.

La funzione di SafeSearch è disponibile in tre diverse opzioni:

- **Livello medio di filtraggio** esclude le immagini più esplicite dai risultati della ricerca di immagini di Google, ma non filtra i risultati di una normale ricerca Web. Questa è l'impostazione predefinita: se non viene modificata dall'utente, SafeSearch viene impostata con l'opzione Livello medio di filtraggio.
- **Livello massimo di filtraggio** applica il filtro di SafeSearch a tutti i risultati di ricerca (sia alla ricerca di immagini che alla normale ricerca Web).
- **Nessun filtro**, come facile immaginare, disattiva completamente il filtro SafeSearch.

4.4 Interpretazione dei risultati

Sul sito di Google è presente una breve guida per l'interpretazione dei risultati di una ricerca in base agli elementi presenti.

Nella figura di seguito sono indicati con le lettere gli elementi di SERP (Search Engine Result Pages).



Nella tabella alla pagina seguente sono indicati elementi e descrizioni.

- A. Collegamenti superiori**
Visualizza la pagina che permette di impostare parametri che limitino l'ambito della ricerca.
- B. Pulsante Cerca con Google**
Selezionare questo pulsante per effettuare un'altra ricerca. In alternativa, è possibile anche utilizzare il tasto Invio.
- C. Ricerca avanzata**
Visualizza la pagina che permette di restringere i campi di ricerca per effettuare delle ricerche avanzate.
- D. Campo di ricerca**
Per effettuare una ricerca, è sufficiente inserire alcune parole chiave descrittive, e premere Invio oppure fare clic sul pulsante Cerca con Google per visualizzare i risultati più rilevanti.
- E. Preferenze**
Visualizza la pagina che permette di impostare le preferenze per le ricerche quali il numero predefinito di risultati da visualizzare su ciascuna pagina, la lingua dell'interfaccia e l'uso del filtro SafeSearch.
- F. Riga delle statistiche**
Questa riga fornisce indicazioni sulla ricerca effettuata ed indica il numero di risultati trovati nonché i tempi di ricerca.
- G. Suggestimenti**
Informazioni per rendere più efficaci le ricerche basate sui termini appena ricercati. Offrono maggiori dettagli sulle funzionalità speciali di Google e indicano gli strumenti per risparmiare tempo.
- H. Risultati OneBox**
Google include molte risorse di informazioni specializzate e quelle più attinenti alle ricerche effettuate, sono riportate in cima ai risultati di ricerca. Queste informazioni sono chiamate “risultati onebox” perché accessibili da qualsiasi casella di ricerca Google. Tipici risultati onebox includono notizie, quotazioni di borsa, previsioni del tempo e siti Web locali correlati alla ricerca.
- I. Titolo della pagina**
La prima riga di un risultato è costituita dal titolo della pagina Web trovata. Qualche volta, al posto del titolo viene visualizzato un URL ad indicare che la pagina non ha un titolo oppure che Google non ha incluso nell'indice tutto il contenuto della pagina stessa. La pagina viene proposta tra i risultati perché altre pagine Web già indicizzate contengono collegamenti che rimandano ad essa. Se il testo associato a questi collegamenti contiene la parola ricercata, è possibile che la pagina venga inclusa nell'elenco dei risultati anche se il testo completo della pagina Web non è stato incluso nell'indice.
- J. Testo sotto il titolo**
Questo testo rappresenta la sintesi della pagina Web trovata e contiene i termini ricercati evidenziati in grassetto. Questa sintesi permette di vedere in quale contesto vengono utilizzati i termini e di decidere se aprire o meno i risultati.
- K. URL della pagina Web**
Indica l'indirizzo Internet della pagina Web trovata.
- L. Dimensioni**
Questo numero indica le dimensioni della sezione di testo della pagina Web trovata. Non vengono visualizzate dimensioni per i siti non ancora indicizzati.
- M. Versione cache**
La selezione del collegamento della versione cache permette di visualizzare la copia archiviata della pagina Web così come si presentava al momento della sua indicizzazione. Questa funzione permette di visualizzare una copia del sito per ricercare le informazioni desiderate nel caso in cui non si possa collegarsi direttamente alla pagina Web. I termini ricercati sono evidenziati anche nella versione cache.
- N. Pagine simili**
Selezionare il collegamento Pagine simili per un risultato specifico per ricercare automaticamente sul Web altre pagine correlate a questo risultato.
- O. Risultato con rientro a sinistra**
Nel caso in cui Google trovi più risultati all'interno di uno stesso sito Web, visualizza per primo nell'elenco il risultato più rilevante. Gli altri risultati vengono visualizzati con un rientro a sinistra.

4.5 Problemi legati ai risultati delle ricerche

In considerazione del funzionamento di Google, le aziende agiscono per migliorare la propria posizione dei loro siti nei risultati delle ricerche su Google (come su altri motori di ricerca).

L'attività è nota come "ottimizzazione per i motori di ricerca" (SEO: Search Engine Optimization) che si propone di creare pagine e siti che rispettino le regole (i "gusti") dei motori di ricerca riuscendo a migliorare la propria posizione nei risultati delle ricerche.

Con il sistema di catalogazione e rilevanza è possibile agire in modo organizzato e distribuito su molti siti web (come su Newsgroup) per far variare il rank (posizionamento) di un risultato di una ricerca. Questa tecnica è nota come "Google bombing".

Collegato alle tecniche di riposizionamento nel rank vi è un nuovo tipo di problema che sta diventando importante: il problema della presenza di malware (trojan, virus, ...) nelle pagine catalogate come rilevanti nei risultati di una ricerca.

Oltre ai limiti ed ai problemi sopra indicati ve ne sono alcuni legati al sistema di catalogazione: Google indicizza solo i primi 101 KByte di HTML di una pagina e circa i primi 120 KByte di un documento PDF.

Google manifesta inoltre alcuni problemi a cercare caratteri speciali come: &, %, #, @, \$, £, €, +, -, /, *, =, <, {, [, (,),], }, >, :, ;, ! e ?. Solitamente questi caratteri vengono ignorati, ma sta cominciando a introdurre il supporto per alcuni come:

I/O (Input/Output) e le note musicali in lingua inglese: A_, A#, B_, B#, C_, C#, D_, D#, E_, E#, F_, F#, G_, G#.

Relativamente all'uso di caratteri speciali nelle ricerche si possono utilizzare i caratteri speciali "&", "+" e "_" una o più volte nel mezzo (o alla fine) di una parola. Ad esempio: A+, a_, C++, net_, page_count, i++++, a&b&c, i&&, "Johnson +& Johnson". Non si può però usare il "+" di fronte ad una parola o numero.

Stando alla documentazione ufficiale, "per impostazione predefinita, Google non riconosce accenti o altri segni diacritici". In sostanza se si ricerca "Munchen" e "München" si otterranno (dovrebbero ottenere) gli stessi risultati.

In realtà provando a fare ricerche per queste due parole si otterranno risultati molto diversi tra di loro. Altra stranezza Ancora più strano, è però il fatto che la stessa ricerca, limitando i risultati a specifiche lingue, da risultati ancora diversi.

In caso di ricerche importanti è consigliabile provare più modi di scrivere la stessa parola.

4.6 Google Directory

Anche Google ha un servizio di directory dove le informazioni sono suddivise per canali tematici e sezioni; nella figura di seguito la pagina iniziale della directory di Google



Attraverso la scelta di canali tematici o categorie si può raffinare l'ambito di ricerca sino ad arrivare al livello di ricerca desiderato con una struttura organizzata di informazioni, contenuti e collegamenti. La directory Google (così come gli altri servizi di directory) è quindi uno strumento molto utile per effettuare ricerche mirate (e limitate) su un argomento specifico. Ad esempio, per visualizzare solo risultati relativi al pianeta Saturno, è sufficiente selezionare "Saturno" nella categoria Scienze > Astronomia della directory Google.

Il servizio di directory di Google è all'indirizzo: <http://directory.google.com/>

5 Riferimenti

Sito di Google: <http://www.google.com>