

# Motori di Ricerca

presente e futuro prossimo

## Il quadro presente

Paolo Ferragina, Università di Pisa

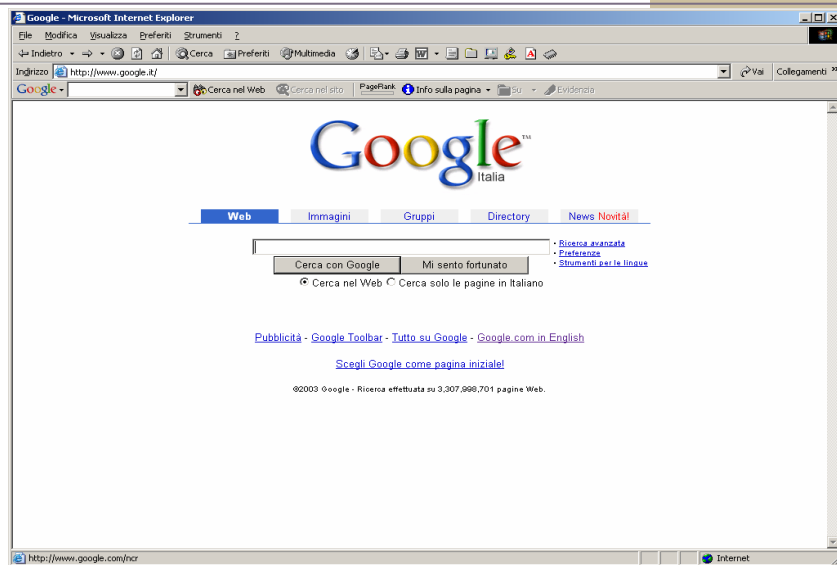
## Fino a pochi anni fa...

- Yahoo (migliore del 1995)
- Inktomi (migliore del 1997)
- Altavista (migliore del 1999)
- Lycos, Excite, Northern Light,...

✓ Oggi Google (60%), Yahoo (28%), Msn (12%), Ask (6%). Ogni utente visita più motori di ricerca per le sue query.

Paolo Ferragina, Università di Pisa

## Il motore più famoso ...



Paolo Ferragina, Università di Pisa

## Cosa non è Google

- **Indice su tutti i documenti disponibili sul Web**
  - Nessun motore lo è
- **Credibile in ogni cosa che ci segnala**
  - Non esiste controllo sulla pubblicazione delle pagine
- **Perfettamente aggiornato**
  - Non riesce a seguire le modifiche giornaliere (milioni di pagine)
- **Protetto da contenuto offensivo**
  - Dispone di un meccanismo di *filtering*, ma non sicuro al 100%

Paolo Ferragina, Università di Pisa

## Cosa è oggi Google

- Alcuni dati interessanti (NY Times, Aprile 2003):
  - Più di 1000 persone
  - 54,000 server - 100,000 processor - 261,000 dischi
  - xMld pagine, 200 milioni query/giorno (60% del totale)
  - 300 milioni di dollari di fatturato 2002 (750 nel 2003 ?)
  - "google" è la parola più utile del 2002 [American Dialect Society]
- Un nuovo scenario di:
  - **Business:** tra i pochissimi a fare molti profitti !
  - **Gestione ed estrazione della conoscenza:** non solo Web
  - **Problemi matematici interessanti:**
    - Qualità risposte, Efficienza, Copertura del web
    - Nuove applicazioni (news,prodotti), Nuovi domini (audio,video)

Paolo Ferragina, Università di Pisa

## Google: Il modello di business in 2 iniziative

- **Search services** via la **Google search appliance**
  - Soluzione hardware+software per un motore di ricerca in ambito intranet o singolo website
  - Hardware fissato e quindi limitati problemi di sviluppo e mantenimento del software
  - Per ora disponibile soltanto in USA e Canada (??)
- **Advertising programs** (100.000 sottoscrittori)
  - **AdSense:** Un sito può fornire spazio sulla sua pagina; le pubblicità da visualizzare vengono scelte da AdSense in funzione dei contenuti della pagina così da rivolgersi a probabili clienti. Il sito riceve un pagamento in funzione del numero di click sul banner.
  - **AdWords:** Una società può scegliere quanto pagare al giorno/mese e indicare le *parole chiave* che descrivono il suo business. Un *banner* viene visualizzato da Google all'atto di ricerche per quelle parole chiave, e la società paga in funzione del numero di click ricevuti.

Paolo Ferragina, Università di Pisa

## Google: altre notizie...

- Il nome deriva dalla parola *GOOGOL*, coniata da un bambino americano di 9 anni per riferirsi al numero  $10^{100}$
- Un po' di storia:
  - [1996-97] Esce il primo prototipo (BackRub).
  - [1998-99] Nasce *Google*, risponde a 10,000 Qpg → 3MI Qpg
  - [2000] 1MI di pagine e 60MI Qpg
  - [2001] 2MI di pagine e 100MI Qpg, ricerche limitabili a 26 linguaggi.  
Introduce Image e File type search, Usenet dal 1981, Google Catalog.
  - [2002] 2,5MI di pagine, ricerche limitabili a 40 linguaggi.  
Introduce AdWords, Google news, Web API, Froogle, Google Labs.
  - [2003] 3MI di pagine, più linguaggi supportati.  
Il programma di business raggiunge i 100,000 sottoscrittori e viene promosso in Italia. Introduce Google AdSense, Local Search.

Paolo Ferragina, Università di Pisa

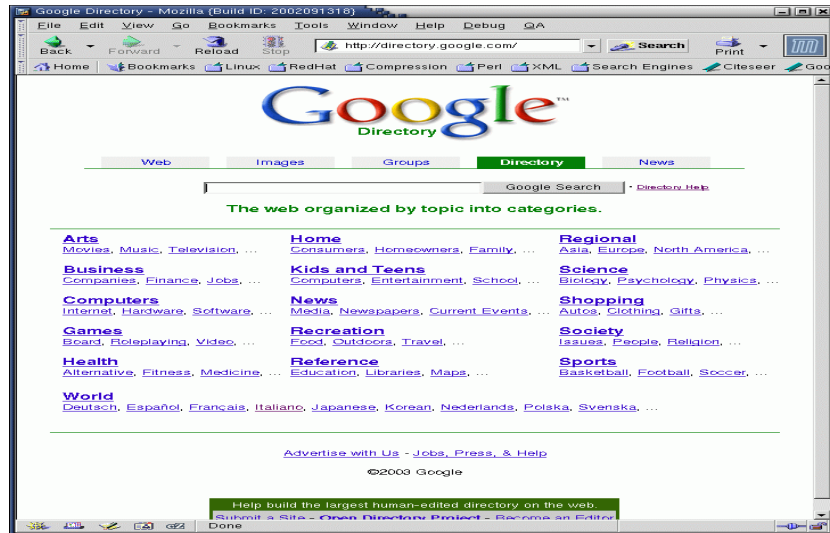
## Motori di Ricerca

presente e futuro prossimo

Altre funzionalità di Google

Paolo Ferragina, Università di Pisa

## L'archivio *dmoz*: open directory project



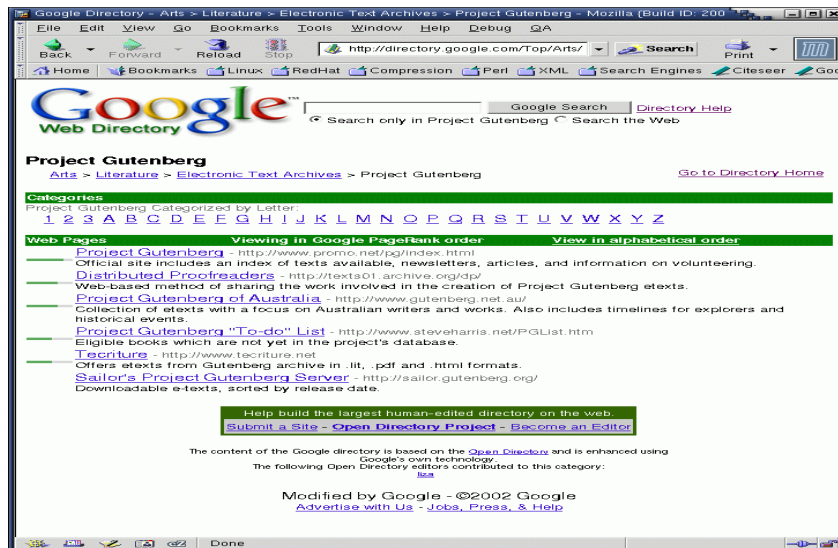
Paolo Ferragina, Università di Pisa

## La directory di Google : *dmoz*

- Raccolto e mantenuto da un gruppo di volontari
  - Siti, non pagine, attentamente selezionati e classificati
- Trade-off
  - Ridotta estensione rispetto all'archivio delle pagine di Google
  - Maggiore qualità delle risposte
- Tipo di ricerche:
  - Per parole chiave, ammette anche *inurl* e *intitle*
  - Per navigazione basata sul soggetto
- Ordine delle risposte:
  - Per *Pagerank* oppure per *ordine alfabetico*

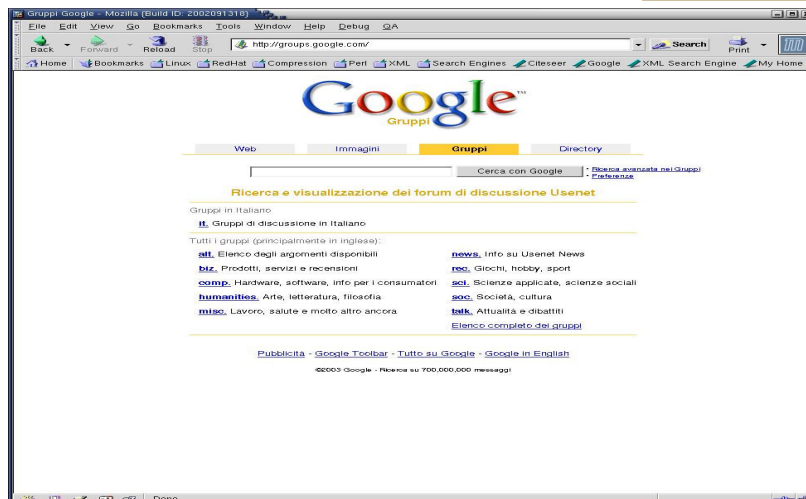
Paolo Ferragina, Università di Pisa

## Un esempio di interrogazione



Paolo Ferragina, Università di Pisa

## L'archivio dei *newsgroup* (Usenet)



- Centinaia di migliaia di topics, secondo una gerarchia
- Oltre 800 milioni di messaggi, raccolti dal 1981

Paolo Ferragina, Università di Pisa

## I gruppi di *Usenet*

- Raccolta di messaggi dal 1981, circa 845 milioni
- Tipo di ricerche:
  - Per parole chiave
  - Per navigazione basata sul soggetto
    - **comp.sys** computer system
    - **humanities** letteratura
  - Con sintassi speciale:
    - **intitle:** o **author:**
    - **group:comp\***
- La ricerca per data qui è molto precisa e utile !!
- **E' una sorgente di informazioni validissime su i più disparati soggetti, ottenibili dalle discussioni tra utenti**

Esiste una  
interfaccia per  
ricerche avanzate

Paolo Ferragina, Università di Pisa

## L'archivio delle immagini



Paolo Ferragina, Università di Pisa

- Circa 800 milioni di immagini dal Web

## Un esempio di interrogazione



Paolo Ferragina, Università di Pisa

## L'archivio delle immagini (contd)

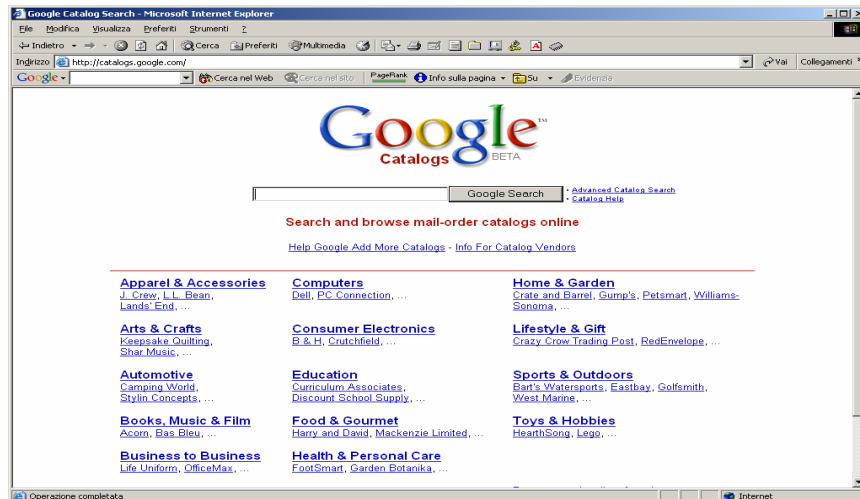
- 800 milioni di immagini prese dal Web e indicizzate per parole chiave, solitamente molto specifiche
- I risultati includono un preview dell'immagine, dimensione, URL
- **Problemi:** 1 parola → molti risultati, alcune parole → nessuno !!!
- **Ricerche:** esiste una sintassi speciale
  - intitle: o inurl: o site:
  - filetype:jpg
- Può essere utile per le vostre presentazioni o articoli

Esiste una interfaccia per ricerche avanzate

Paolo Ferragina, Università di Pisa



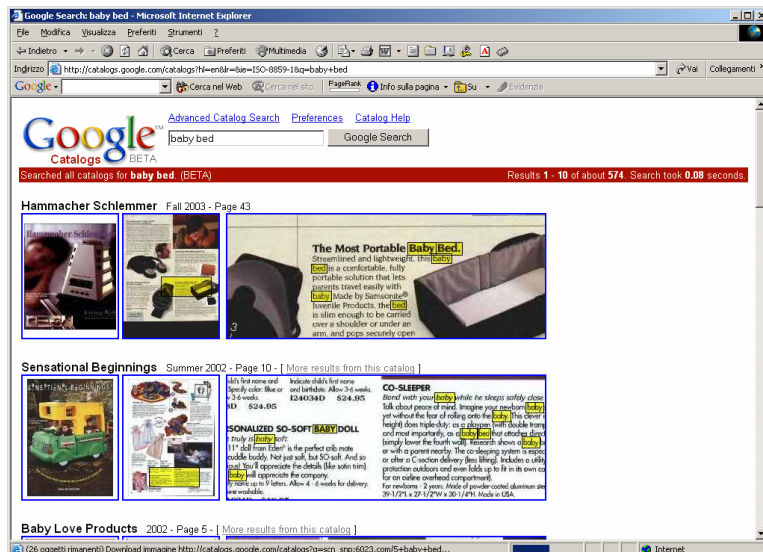
# I cataloghi



- Oltre 6000 cataloghi digitalizzati
- Ricerca per parole chiave o navigazione
- Catalogo (anche vecchio) visualizzabile, o sito del venditore

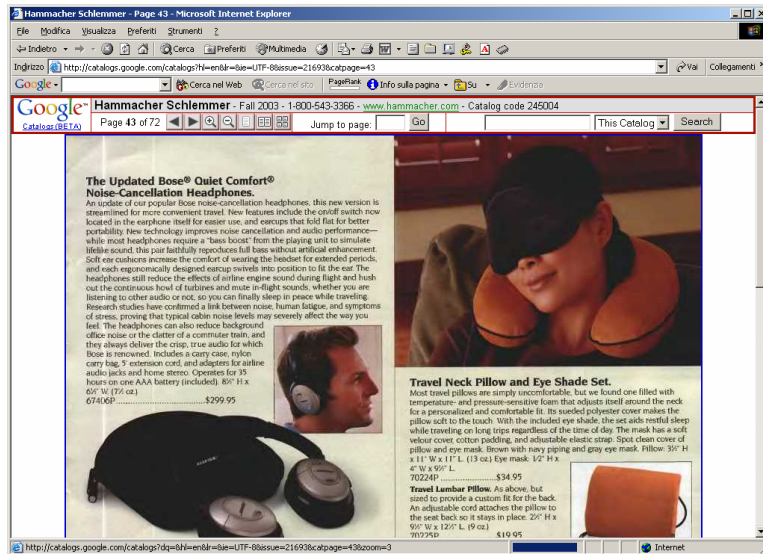
Paolo Ferragina, Università di Pisa

# Cercando "baby bed"



Paolo Ferragina, Università di Pisa

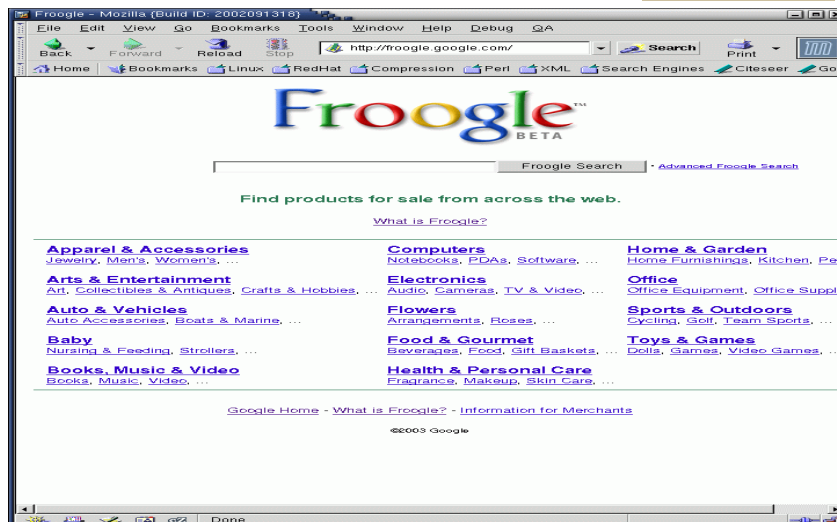
## Cliccando su una immagine....



Paolo Ferragina, Università di Pisa

- ... basta scorrere l'immagine!!

## Acquisti on-line...



Paolo Ferragina, Università di Pisa

- Froogle = Google + frugal
- Altri approcci noti: shopping robots (**Jango**)

## Un esempio di interrogazione

View  
List view  
> Grid view  
Sort By  
> Best match  
Price: low to high  
Price: high to low  
Price Range  
\$ to \$  
Go  
Group By  
> Store  
> Show All Products  
Search within  
> All Categories  
Electronics  
Cameras  
Digital Cameras

Sponsored Links  
Canon G5  
Find the Lowest Price on the Latest Canon Digital Cameras! BizRate.com  
Canon g5  
Decine di nuove occasioni su eBay  
Compra e vendi in sicurezza  
www.ebay.it  
Compare prices  
and find the lowest price on the net for the powershot g-5  
www.pricerarrow.com  
See your message here...

Paolo Ferragina, Università di Pisa

## Notizie USA e internazionali

World  
U.S.  
Business  
Sci/Tech  
Sports  
Entertainment  
Health  
Text Version  
About Google News

Top Stories  
Dawn raids target Baghdad  
BBC - 30 minutes ago  
Military targets and Iraq's main TV station were hit. The station couldn't broadcast anything for three hours.  
US warplanes try to weaken Iraqi Republican guard units  
Deutsche Welle  
Resistance Endures Amid the Rubble  
Washington Post  
Ananova - ABC News - CNN Asia - Gulf Daily News - and 1302 related »  
Blair Warns Of Resistance 'To the End'  
Washington Post - 2 hours ago  
Prime Minister and Bush To Meet at Camp David  
Washington Post Foreign Service  
Wednesday, March 25, 2003  
Bush to Meet Military Planners, Hold Talks with Blair  
Voice of America  
Can Blair convince Bush to share his belief in the international institutions?  
Guardian  
Arab News - San Jose Mercury News - The Scotman - Washington Times - and 270 related »

World »  
U.S. »  
Tanker blown up in Kashmir  
BBC - 1 hour ago  
One person has been killed and six others injured in a bomb explosion in Indian-administered Kashmir.  
'Another leaf in the bloody history.'  
Guardian  
US-based groups condemn J&K massacre  
Rediff  
Daily Times - The Hindu - Mid-Day Mumbai - PTI News - and 290 related »  
Second US Soldier Dies After Grenade Attack  
Washington Post - 1 hour ago  
A second US soldier died in Kuwait on Tuesday from wounds suffered when a fellow soldier allegedly lobbed grenades into three tents on March 23, the Idaho Air National Guard said early Wednesday.  
Idaho Guardsman From Boise Remains in Serious Condition After Kuwait Grenade Attack  
KBSI

Paolo Ferragina, Università di Pisa

# Google news

- Oltre 4500 sorgenti di informazione
- I risultati includono:
  - Notizie vecchie o recenti (fino a 1 minuto fa...)
  - ❖ Indicazione della provenienza
  - ❖ Raggruppate per soggetto o storia
  - ❖ Presentazione stile Rivista o Quotidiano
- Tutto eseguito in maniera automatica
  - Loro dichiarano "senza influenze politiche o personali"  
.... noi osserviamo che operano "senza un oggettivo filtro" !!!
- E' sorprendente nella sua efficienza, efficacia e ampiezza !!
- Esistono altre proposte, come quella di Yahoo (offre un free alert !!)

Esistono versioni per vari paesi

Paolo Ferragina, Università di Pisa

# La versione italiana

Google News Italia - Microsoft Internet Explorer

Indirizzo: <http://news.google.it/>

Google News Italia BETA

Ricerca e leggi tra 250 fonti di notizie costantemente aggiornate.

Ultimo aggiornamento automatico: 9 minuti fa

**Prima pagina**

**Dal mondo**

- Italia
- Economia
- Scienze
- Sport
- Spettacolo
- Salute

**Prima pagina**

**Governo: scorie nucleari da stoccare in Basilicata**  
Corriere della Sera - 1 ora fa  
ROMA - Una decisione che ha già scatenato un mare di polemiche. Il Consiglio dei ministri ha stabilito che l'impianto nazionale di stoccaggio delle scorie nucleari venga realizzato nel comune di Scanzano Jonico in Basilicata, un paese a circa 5 km dal ...  
[Nucleare, presidente Regione, si oppongono in ogni sede](#) Trentino  
[Jean, scelta sito fondata su basi tecniche](#) La Gazzetta del Mezzogiorno  
[Il Nuovo - Libero.it - L'Espresso - Villaggio Globale - e altri 24 articoli simili >](#)

**Lutto nazionale per le vittime in Iraq**  
In Italia - 35 minuti fa  
Martedì i funerali dei 19 caduti nell'attentato di Nassiriya. Il ministro della Difesa Martino: "La missione non cambia". Partiti per l'Iraq altri 50 carabinieri ...  
"Pisgno come a Ground Zero": Panorama  
[Martino a Nassiriya: "Sembra Ground Zero"](#) Yahoo! Italia Notizie  
[La Repubblica - La Gazzetta del Mezzogiorno - L'Espresso - AudioNews - e altri 67 articoli simili >](#)

**Alitalia, da Cdm primo sì a privatizzazione**  
Yahoo! Italia Notizie - [e altri 6 articoli simili >](#)

**Un Celeron per i portatili**  
PC Open - [e altri 4 articoli simili >](#)

**Diritti tv, scontro in Lega. Le società di Gioco Calcio insistono: noi non giochiamo**  
L'Unità - [e altri 24 articoli simili >](#)

**Ingerman: ho abortito dopo "Scherzi a parte"**  
Corriere della Sera - [e altri 17 articoli simili >](#)

**Mafia, indagati 2 medici: curarono Provenzano**  
L'Espresso - [e altri 18 articoli simili >](#)

**Ultime notizie**

Scanzano Jonico	Consiglio Comunale
Roberto Merandini	Alessandra Canale
Reggio Calabria	Guido Bellini
Ground Zero	Stato Maggiore
Santa Maria	Marco Biagi

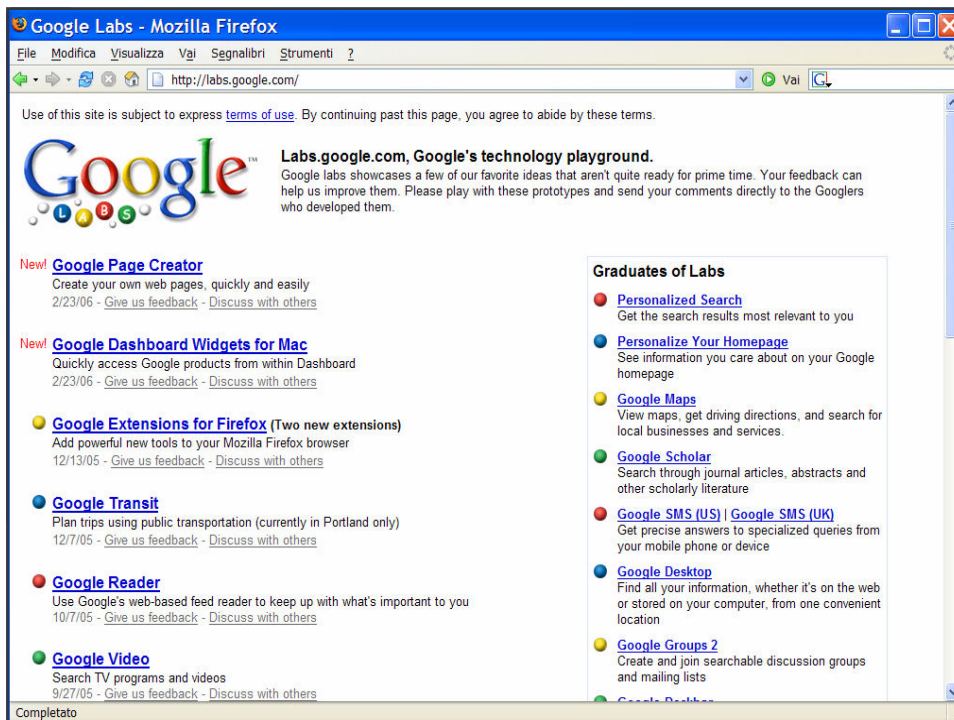
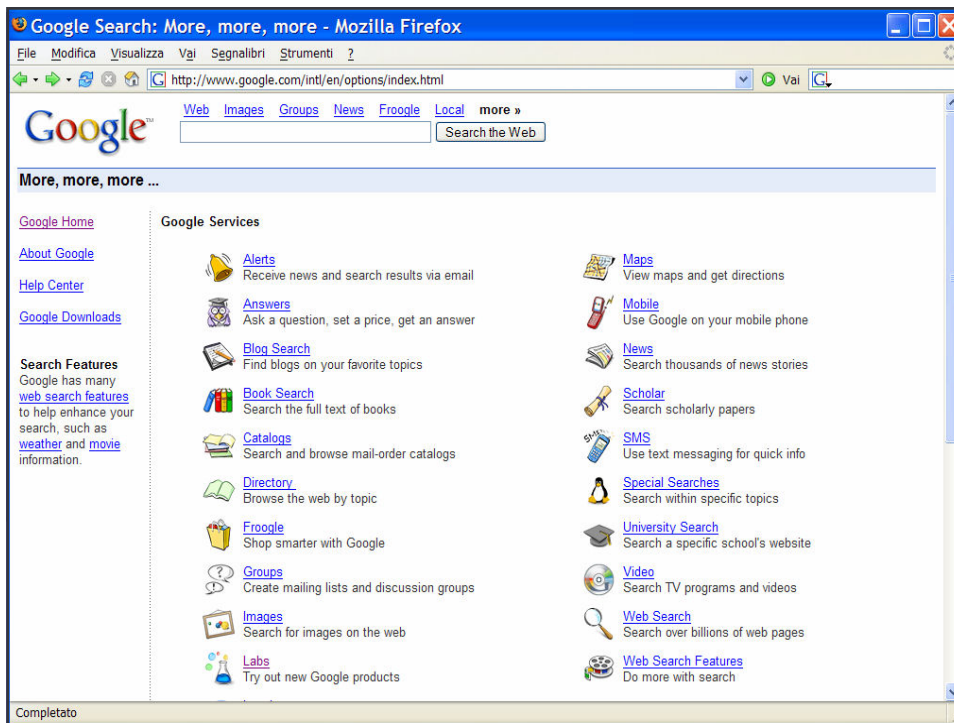
**Dal mondo >**

**Italia >**

**Attacco Nassiriya, colpiti militari italiani giunti per la pace**  
[La Repubblica - 5 ore fa](#)

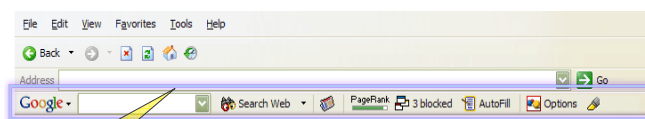
**La Rai si scusa: "Solo un errore"**  
[La Repubblica - 5 ore fa](#)

Paolo Ferragina, Università di Pisa



## Google è tanto altro ancora

- Dizionari, anche per dialetti: `~parola define:parola`
- Elenchi telefonici, pagine gialle,... [anche `reverse search`]
- `Stocks:identificatore_società` [Yahoo finance]
- **Weblogs:** “an online site that keeps running commentary and associated links, updated daily.”
  - [radio.weblogs.com](http://radio.weblogs.com), [www.blogspot.com](http://www.blogspot.com), [www.blogger.com](http://www.blogger.com)



Track user behavior !!

Occhio alla pagina di Google !!

Paolo Ferragina, Università di Pisa

## Motori di Ricerca

presente e futuro prossimo

Altre interessanti proposte

Paolo Ferragina, Università di Pisa

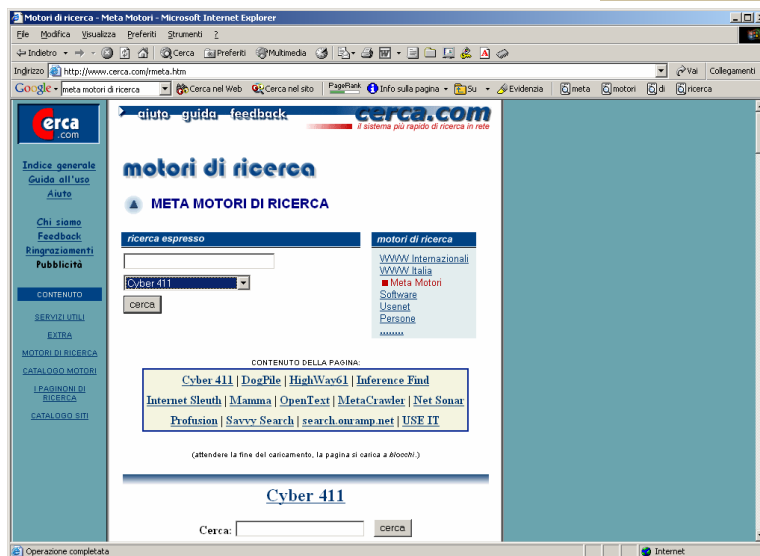
## Funzionalità: Il quadro corrente

- Migliorare il “ranking” delle pagine: **Teoma** e molti altri....
- Migliorare la copertura del Web: “**Meta**” motori di ricerca
- “Supporto all’utente” nella ricerca dei risultati che meglio soddisfano le sue interrogazioni !
  - Suggerimenti sulle parole da cercare (**AskJeeves**)
  - Suggerimenti su *oggetti* visti da “utenti simili” (**Amazon, Epinions**)
  - Categorizzazione delle risposte (il fu **NorthernLight**)
    - Testuale (**Vivisimo, Copernic**) o grafico (**Kartoo**)

Knowledge Management Systems

Paolo Ferragina, Università di Pisa

## Un tipo di Meta-motore



The screenshot shows a web browser window displaying the Cerca.com website. The page is titled "Motori di ricerca - Meta Motori" and features a search interface. The main content area is titled "MOTORI DI RICERCA" and includes a search box with the text "Cyber 411" entered. Below the search box, there is a list of search engines: Cyber 411, DogPile, HighWay61, Inference Find, Internet Sleuth, Mamna, OpenText, MetaCrawler, Net Sonar, Profusion, Savvy Search, search.ouamp.net, and USE IT. The page also includes a sidebar with navigation links and a footer with the text "Operazione completata".

Paolo Ferragina, Università di Pisa

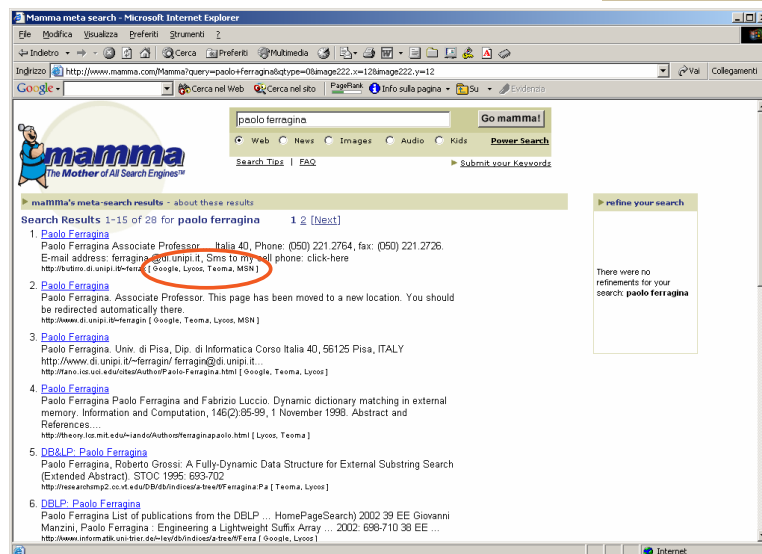


## Qualcosa di più sofisticato



Paolo Ferragina, Università di Pisa

## La struttura delle risposte



Paolo Ferragina, Università di Pisa



## Suggerimenti sui termini: AskJeeves

AskJeeves Results - Mozilla (Build ID: 2002091316)

http://web.ask.com/web?q=computational+biology&x=0&y=0&qsrc=

computational biology

WEB RESULTS NEWS RESULTS SHOPPING RESULTS Help Edit Guidelines

You may find these sponsored results helpful:

**Computational Biology**  
Solve your software, database integration and LIMS problems.  
From: www.cantient.com

You may find my search results helpful:

**EBI, the European Bioinformatics Institute**  
The one-stop portal to all your bioinformatics needs: the European Bioinformatics Institute (EMBL-EBI) is a centre for research and services in...  
From: http://www.ebi.ac.uk/

**Computational Biology at UC Santa Cruz**  
The Computational Biology group at UCSC is dedicated to the discovery and implementation of algorithms that facilitate the understanding of...  
From: http://www.cse.ucsc.edu/research/compbio/

**DMERC: Computational Biology, structure prediction, functional**  
Computational Biology Tools: Who we are: Completed genomes search and analysis. Protein structure prediction. Protein Domain Profile Analysis...  
From: http://bmerc-www.bu.edu/

**EXPASY Molecular Biology Server**  
EXPASY Molecular Biology Server...  
From: http://www.expasy.ch/

**NCBI Home Page**  
The National Center for Biotechnology Information (NCBI) provides an integrated approach to the use of gene and protein sequence information, the sci...  
From: http://www.ncbi.nlm.nih.gov/

**Center for Computational and Experimental Genomics**  
Supports research related to computational biology for the USC Departments of Biological Sciences, Computer Science, and Mathematics.  
From: http://www.hnto.usc.edu/

**Bioinformatics and Computational Biology Graduate Program**  
Bioinformatics and Computational Biology faculty, research, education, and resources at Iowa State University.  
From: http://www.bcb.iastate.edu/

**Computational Molecular biology at NIH**  
Computational Molecular Biology at NIH...  
From: http://ncbio.info.nih.gov/molebio/

Try these search terms to see more answers:

- Bioinformatics
- Computational Biology
- Molecular Biology Problems
- Computational Molecular Biology
- tools
- Journal Of Computational Biology
- Online Journal Bioinformatics
- Computational Biology Companies
- Protein Bioinformatics
- Computational Biology sequence Receptor
- National Bioinformatics Institute

Paolo Ferragina, Università di Pisa

© Si appoggia a Teoma, e risponde a "domande" !!

## Suggerimenti su oggetti visti: Amazon

Amazon.com: Books: Computational Molecular Biology: An Algorithmic Approach (Computational Mole - Microsoft Internet Explorer...)

http://www.amazon.com/exec/obidos/tg/detail/-/0262161974/qid=1051020169/sr=1-4/ref=sr\_1\_4/102-8687298-56193C

SEARCH Books

**Computational Molecular Biology: An Algorithmic Approach (Computational Molecular Biology)**  
by Pavel A. Pevzner

List Price: \$47.00  
Price: \$47.00 & This item ships for FREE with Super Saver Shipping. See details

Availability: Usually ships within 24 hours

Used & new from \$42.12

Edition: Hardcover

See more product details

READY TO BUY?  
Add to Shopping Cart  
or  
Add to Wish List  
Add to Wedding Registry

RECENTLY VIEWED ITEMS

- Handbook of Comparative Genomics: Principles and Methodology by Cecilia Saccone (Author), Graziano Pesole (Author)
- Algorithms on Strings, Trees, and Sequences: Computational Biology by Dan Gusfield (Author)

Customers who bought this book also bought:

- Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids by Richard Durbin (Author), et al (Paperback)
- Introduction to Computational Biology: Maps, Sequences and Genomes by Michael S. Waterman (Paperback)
- Bioinformatics: Sequence and Genome Analysis by David W. Mount (Hardcover)
- Statistical Methods in Bioinformatics: An Introduction by Gregory R. Grant, Warren J. Ewens (Hardcover)

Paolo Ferragina, Università di Pisa

© Ora anche ricerche sul testo dei libri! (anche Google)

## Vivísimo: Raggruppare le risposte

The screenshot shows the Vivísimo search engine interface. The search query is "computational biology", which has returned 173 documents. The results are clustered into categories such as:

- computational biology (173)
- Computational Molecular Biology (43)
- Bioinformatics (34)
- Laboratory (15)
- Center for Computational Biology (11)
- Software (8)
- Computational Biology Group (10)
- Structural (10)
- Journal of Computational Biology (4)
- Computational Biology Centers (3)
- Pennsylvania, University (3)

The main content area displays several search results, including sponsored links for "Buy Computational Biology Products" and "Computational Biology At BizRate". It also lists three numbered results from the International Society for Computational Biology (ISCB) and the National Institutes of Health (NIH).

Paolo Ferragina, Università di Pisa

- ⊙ Offre categorizzazione risultati di FirstGov.com
- ⊙ Copernic funziona sul nostro desktop

## Kartoo: Non solo testo

The screenshot shows the Kartoo search engine interface. The search query is "computational biology". The results are displayed as a network diagram where nodes represent websites and lines represent links between them. The nodes include:

- nervana.montana.edu
- academic
- www.research.ibm.com
- contact
- www.iscb.org
- science
- research
- links
- software centers
- www.cbc.med.umn.edu
- interface
- molbio.info.nih.gov
- combio.ornl.gov
- bioinformatics
- www.cse.ucsc.edu
- molecular
- www.bcb.iastate.edu
- www.liebertpub.com
- www.hto.usc.edu

The interface also includes a sidebar with a list of topics and a legend explaining the symbols used in the network diagram.

Paolo Ferragina, Università di Pisa

## Un mix interessante: Ez2find

The screenshot shows the ez2find search engine interface. The search term 'computational biology' is entered in the search box. The results are displayed in a list format. Several results are circled in red, including 'The International Society for Computational Biology', 'Computational Biology at UC Santa Cruz', 'DNA Computing, Computational Biology, and Molecular Computing', 'Bioinformatics and Computational Biology Graduate Program', 'Center for Computational Biology', 'Computational Biology and Informatics Laboratory at UPenn', 'Laboratory of Experimental and Computational Biology', 'Journal of Computational Biology: A Journal of Computational Molecular Cell Biology', 'Computational Biology Software', and 'EXPASY Molecular Biology Server'. A red circle highlights the 'Open Directory' link in the left sidebar. Another red circle highlights the 'amazon.com' logo in the right sidebar.

Paolo Ferragina, Università di Pisa

© Metasearch + Directory + Cluster dei risultati !!

## Un tool interessante

The screenshot shows the WEBSOM zoomed map interface. The map displays a heatmap of search terms, with 'pc.video' and 'pc.chips' being prominent. A legend on the right lists the terms and their corresponding categories. A tip box indicates that clicking arrows on the map allows moving up to the overall view. The bottom of the interface includes a 'Click any area on the map to get a zoomed view!' instruction and a 'Completato' status.

**WEBSOM zoomed map - Million documents - Mozilla Firefox**

File Modifica Visualizza Vai Segnalibri Strumenti ?

http://websom.hut.fi/websom/milliondemo/html/1\_ex3.html

**WEBSOM zoomed map - Million documents**

Click arrows to move to neighboring areas on the map, and to move up to the overall view.

acorn - comp.sys.acorn.hardware  
 blues - rec.music.bluesnote  
 books - rec.arts.books  
 cdrom - comp.publish.cdrom.hardware  
 classical - rec.music.classical  
 humor - rec.humor  
 movies - rec.arts.movies.current-films  
 pc.cdrom - comp.sys.ibm.pc.hardware.cd-rom  
 pc.chips - comp.sys.ibm.pc.hardware.chips  
 pc.storage - comp.sys.ibm.pc.hardware.storage  
 pc.video - comp.sys.ibm.pc.hardware.video  
 ps2 - comp.sys.ibm.ps2.hardware  
 sgi - comp.sys.sgi.hardware

Click any area on the map to get a zoomed view!

Completato