

Recupero di Documenti

(Motori di Ricerca: presente e futuro prossimo)

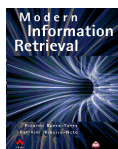
Paolo Ferragina
Dipartimento di Informatica
Università di Pisa

Paolo Ferragina, Università di Pisa

Libri di testo



Google, The pocket guide
T. Calishain *et al.*, O'Reilly 2003.
"Una guida semplice all'uso di Google"



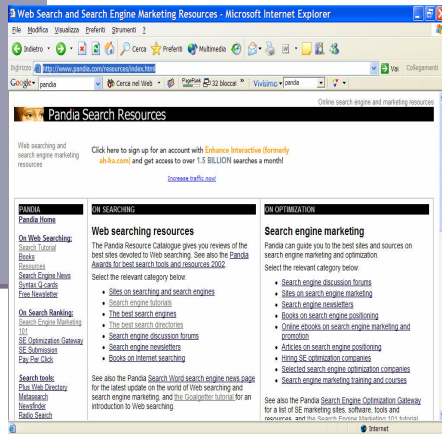
Modern Information Retrieval
R. Baeza-Yates e B. Ribeiro-Neto, Addison-Wesley, 1999.
"Libro di livello universitario sul progetto di un motore di ricerca"

Consultare la pagina web del corso per avere una
indicazione precisa sulle parti da studiare !!

Paolo Ferragina, Università di Pisa

Tre riferimenti interessanti

- <http://www.pandia.com>
- <http://websearch.about.com/>
- <http://searchenginewatch.com/>



Paolo Ferragina, Università di Pisa

Motori di Ricerca presente e futuro prossimo

Prologo

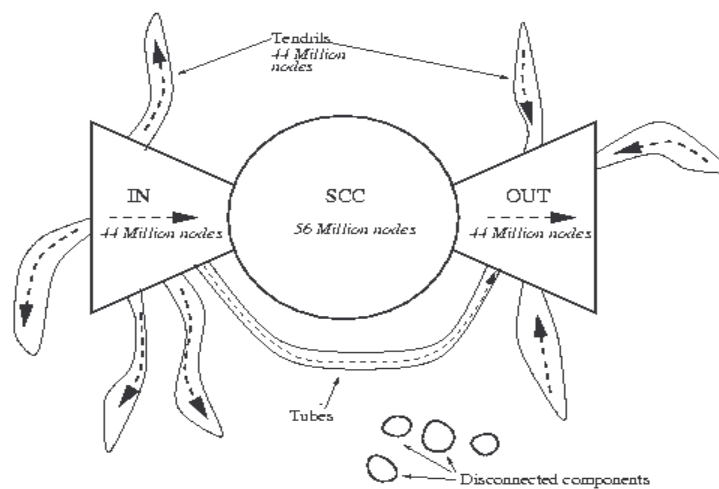
Paolo Ferragina, Università di Pisa

Il Web

- “Surface Web”: 25 ÷ 75 Terabytes (1Tb = 1000 Gb)
 - 11.5 miliardi di pagine (cambiano circa 10 milioni al giorno)
 - Pagina in media 5 ÷ 40Kb, #links ~ 10
 - Circa il 23% delle pagine è duplicato, altro 20% è spam
- “Hidden Web”: circa 500 volte più grande
 - Siti intranet, database, pagine dinamiche,...
 - Circa 4,200 Tb di dati testuali interessanti

Paolo Ferragina, Università di Pisa

Una immagine pittorica del Web



Paolo Ferragina, Università di Pisa

Velocità di cambiamento

[snapshot settimanale nel 2004: 154 web sites, 3÷5 mil pg, 65Gb]

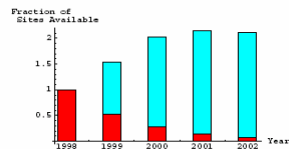


Figure 4: Percent of IP addresses identifying a Web site in Year A also identifying a Web site in Year B. For example, 56% of IP addresses identifying a Web site in the 1998 sample also identified one in 1999 sample. Taken from <http://wcp.oclc.org/>.

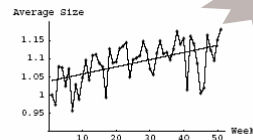


Figure 5: Average page sizes in our snapshots over time.

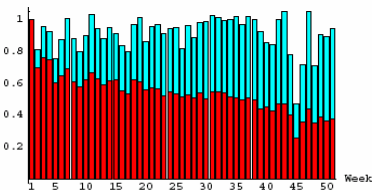


Figure 2: Fraction of pages from the first crawl still existing after n weeks (dark bars) and new pages (light bars).

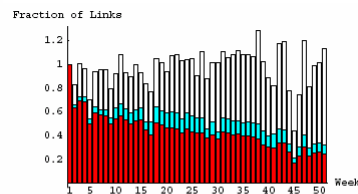


Figure 8: Fraction of links from the first weekly snapshot still existing after n weeks (dark/bottom portion of the bars), new links from existing pages (grey/middle) and new links from new pages (white/top).

Ma non solo il Web

- Email:
 - 610Mld di messaggi al giorno
 - 150,000 Mailing List (circa 675Tb all'anno)
- Ogni anno: Libri (8 Tb), Quotidiani (25 Tb), Periodici (12 Tb), documenti di ufficio (210 Tb).

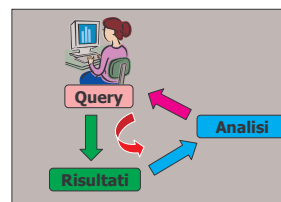
... e non solo testo ma anche audio, video, immagini,

Motore di ricerca vs. Utente

Obiettivo: Recuperare i documenti che sono "rilevanti" per l'interrogazione formulata dall'utente.

- Documento: file word o pdf, pagina web, email, ...
- Interrogazione: lista parole chiave
- Rilevante ?
 - Concetto soggettivo e mutevole
 - Dimensione archivi elettronici in continua espansione
 - Interrogazioni "selettive" sono difficili da formulare
 - Pagine eterogenee

☹ La ricerca è un processo difficile
... e quindi necessariamente "ciclico"



Paolo Ferragina, Università di Pisa

Dimensione vs. precisione

6 navigatori italiani su 10 non sanno cercare nel Web:

"... Su un campione di 856 navigatori italiani, fra i 25 e i 55 anni, che utilizzano Internet regolarmente,

- il 28% trova difficoltà solo «alcune volte»,
- il 33% incontra sempre serie difficoltà,

Il 75% degli utenti ritiene che i motori di ricerca siano il servizio più importante del Web

[...] Tutto ciò genera stress, frustrazione e senso di smarrimento nel mare del Web. [...]

... Quasi un italiano su tre sogna un motore di ricerca automatico e intelligente che non agisca solo per parole chiave... "

Corriere della Sera, 12 Aprile 2001

Paolo Ferragina, Università di Pisa

Interrogazioni sul Web: vari tipi

- Lo “user need”
 - Informational – vogliamo apprendere qualcosa (~40%)
 - Influenza cinese
 - Navigational – vogliamo andare su una pagina (~25%)
 - United Airlines
 - Transactional – vogliamo fare qualcosa con il web (~35%)
 - Accesso a un servizio
 - Tempo a Roma
 - Download
 - Immagini di Marte
 - Shop
 - Nikon CoolPix
 - Altre possibilità
 - Trovare un buon “hub”
 - Affitto macchina a Roma
 - Esplorare il web “see what’s there”

Paolo Ferragina, Università di Pisa

Interrogazioni sul Web: gli utenti

- Query mal definite
 - Brevi
 - AV 2001: 2.54 termini
 - 80% < 3 parole
 - Termini imprecisi
- Grande diversità
 - Bisogni
 - Conoscenza
 - Pazienza
- Comportamento
 - 85% guardano soltanto alla prima pagina
 - 78% delle query non sono modificate

Paolo Ferragina, Università di Pisa

La storia: Prima generazione



The screenshot shows the AltaVista homepage in a Mozilla Firefox browser window. The page features the AltaVista logo at the top, followed by a navigation bar with links like 'HOME', 'ABOUT', 'CONTACT', 'SUPPORT', 'ADVANCED SEARCH', and 'SERVICES'. Below the navigation bar is a large advertisement for 'AUTO-TEL' with the text 'Car Buying & Car Insurance Pain Relief'. Underneath the ad is a search bar with the text 'Search the Web and Display the Results in Standard Form' and a 'Submit' button. Below the search bar are links for 'Advanced Search' and 'Add URL'. Further down are links for 'Contests' and 'Creative Web'. At the bottom of the page, there is a footer with the text '[Creative|Search|Humor|Email]'.

- Usava solo il testo sulla pagina
 - Frequenza delle parole, linguaggio

1996-98

La storia: Seconda generazione



The screenshot shows the Google search engine homepage. At the top is the 'Google! BETA' logo. Below the logo is a search bar with the text 'Search the web using Google!' and a 'Google Search' button. Below the search bar are three columns of links: 'Special Searches' (Stanford Search, Linux Search), 'Help!' (About Google!, Company Info, Google Links), and 'Get Google! updates monthly: your e-mail' (Subscribe, Archive). At the bottom of the page is the copyright notice 'Copyright ©1998 Google Inc.'.

- Sfruttare la struttura del web
 - Link (or connectivity) analysis
 - Anchor-text

1998-0?

La storia: Terza generazione

The screenshot shows the Yahoo! Italia homepage. At the top, there's a search bar with 'Ricerca:' and a 'Siti Web' button. Below it, a navigation menu includes 'Mio Yahoo!', 'Mail', and 'Opzioni'. The main content area is divided into several sections: 'In evidenza' with a featured article on shopping trends; 'Shopping, nuove tendenze' with sub-articles on online shopping and car insurance; 'Ciao Paolo' with a Nissan advertisement; 'Yahoo! Giochi' with a link to games; and 'Yahoo! Pulse' with a link to entertainment news. A left sidebar contains various service links like 'Auto', 'Chat', 'Cinema', etc. At the bottom, there's a 'focus sull'utente' section with bullet points.

■ focus sull'utente

- Analisi semantica, Contesto
- Aiuti, click-through
- Integrazione ricerca e browsing

oggi

Paolo Ferragina, Università di Pisa

Quarta generazione ???

Quarta generazione → Information Supply

[Andrei Broder, VP emerging search tech, Yahoo! Research]

Paolo Ferragina, Università di Pisa

Ieri....



Oggi...



Paolo Ferragina, Università di Pisa

Ieri...

```
> finger karger@CSAIL.MIT.EDU
[CSAIL.MIT.EDU]
KARGER David Karger Theory of Computation Faculty
<karger@theory.lcs.mit.edu>
Project: Analysis of Algorithms and/or Information Retrieval
Work: NE43-321; 258-6167
Home: 1600 Mass. Ave., Apt. 407, Cambridge, MA 02138; 4919592
>
```

Tutti questi
Tool usano un
Search Engine

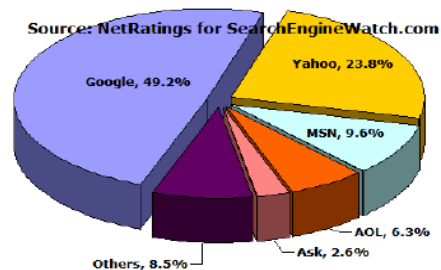
...oggi



Paolo Ferragina, Università di Pisa

Perché tanto interesse sui "motori" ?

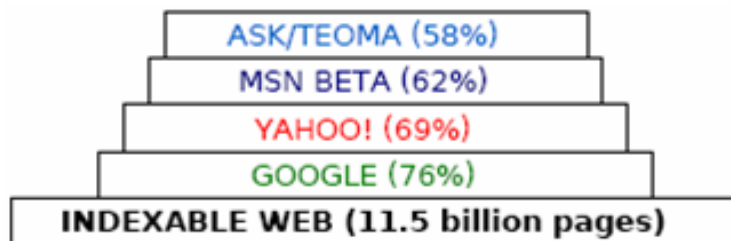
- Più dell' 85% utenti arrivati a un sito attraverso una ricerca
- Il 33% degli utenti crede che i primi risultati di una ricerca sono il posto migliore dove spendere i soldi
- Distribuzione delle ricerche negli USA



- <http://searchenginewatch.com> -

Paolo Ferragina, Università di Pisa

Dimensione reale [2005]

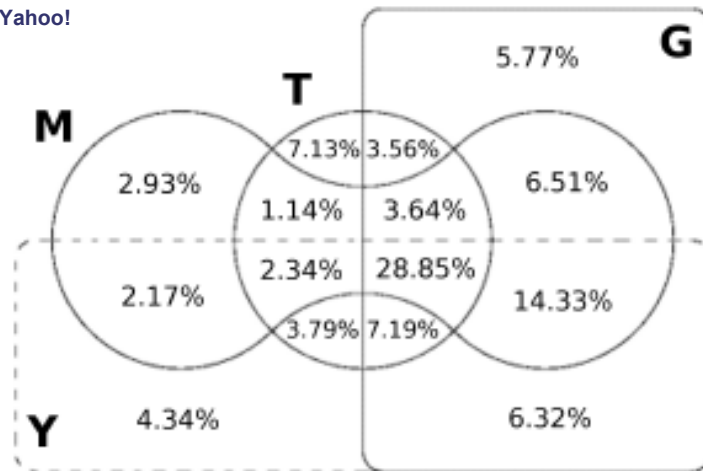


Google vs Yahoo:
20-30% risultati identici

Paolo Ferragina, Università di Pisa

La visione globale

G = Google
M = Msn
T = ASK/Teoma
Y = Yahoo!



Paolo

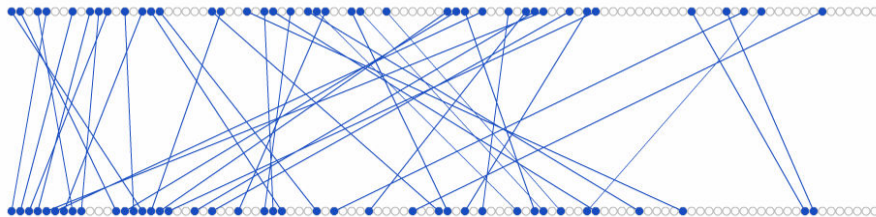
Ranking: Google vs Yahoo!

Comparing Google and Yahoo! Search results 1 - 100 for "mousetrap".

mousetrap

Google

touch the dots!



YAHOO!

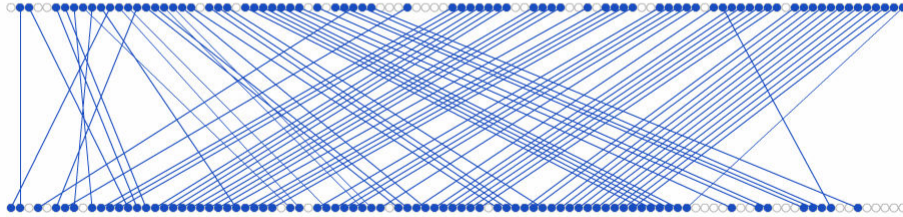
Ranking: Google vs Google.cn

Comparing Google and Google.cn results 1 - 100 for "comunism":

comunism

Google

http://www.marxists.org/romana/dictionar/c/Comunism.htm



Google.cn

Più confronti

Engine Comparison - Microsoft Internet Explorer

Indirizzo: <http://roquefort.di.unipi.it/clus-bin/c?q=trappola+per+topi+&google=1&yahoo=1&altavista=1&alltheweb=1>

Cerca nel Web PageRank 197 blocchi Vivisimo trappola per topi

| | | | | | | | | | | | | | | | | | | | | | | | |
|------|-----------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|
| HIDE | Google | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| HIDE | altavista | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| HIDE | alltheweb | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| HIDE | looksmart | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| HIDE | overture | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| HIDE | msn | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |

Links retrieved by two or more engines selected link links retrieved by only one engine

F. Ferragina, A. Gulli, A. Signorini, M. Venti | Try our Clustering Engine

Racconti - Trappola per topi
Trappola per topi - una storia semiseria, ma drammaticamente vera -. Raccontodi Luca Di Bella. Eh, eh. ... Una trappola per topi. Che ci vorrà mai? ...
[google:5 | altavista:1]

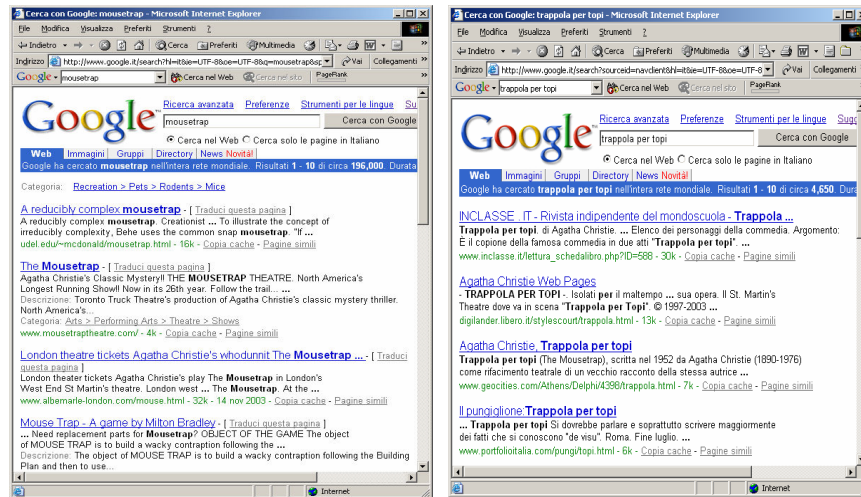
Racconti - Trappola per topi 2
Trappola per topi - Racconto di Luca Di Bella P2. Mattina. Ero di nuovo di fronte alla scatola. Il dischetto era sull'ingresso, l'elastico spezzato. ...
[google:6]

COMP. TORINO SPETTACOLI Trappola per topi
Torino Spettacoli Teatro Stabile Privato direzione artistica Germana Erba e Piero Nuti presenta TRAPPOLA PER TOPI di Agatha Christie con ADRIANA INNOCENTI ...
[google:7]

Trappola per topi
Compagnia Teatro Anzani/Teatro Dehon presenta TRAPPOLA PER TOPI di Agatha Christie con ALDO SASSI ALESSANDRA CORTESE regia Guido ...
Operazione completata

Paolo Ferragina, Università di Pisa

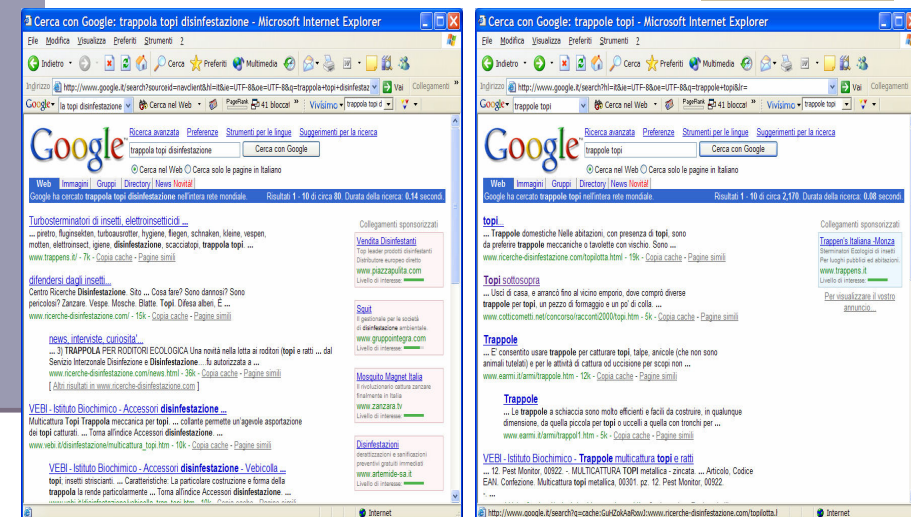
Formulare una interrogazione (quante difficoltà)



🔍 **Varie difficoltà:** problemi di astrazione, sinonimia, polisemia, 10 risultati

Paolo Ferragina, Università di Pisa

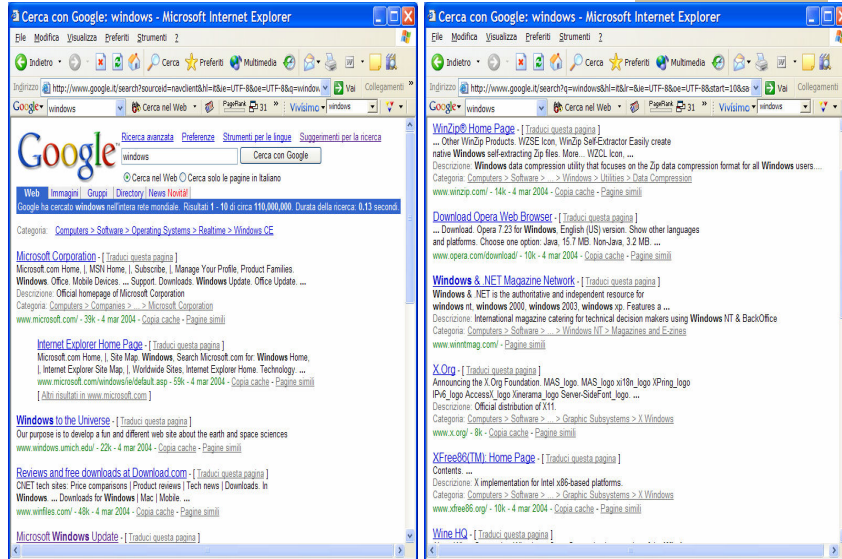
Formulare una interrogazione (quante difficoltà)



🔍 **Varie difficoltà:** problemi di astrazione, sinonimia, polisemia, 10 risultati

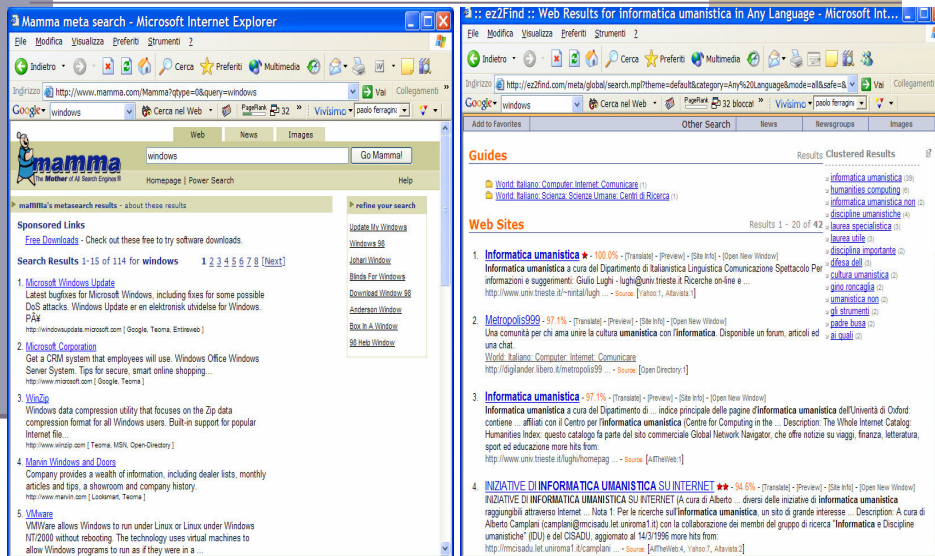
Paolo Ferragina, Università di Pisa

Formulare una interrogazione (quante difficoltà)



Paolo Ferragina, Università di Pisa

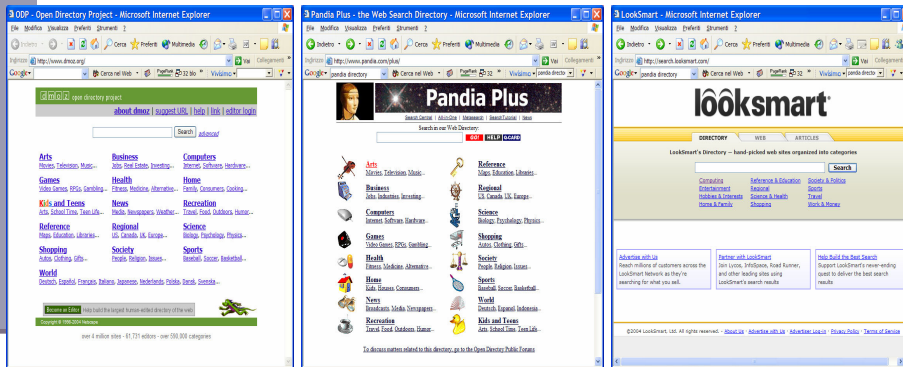
Tipi di ricerche: Meta-Motori



Paolo Ferragina, Università di Pisa

Tipi di ricerche: Directory

- [DMOZ](#), [Yahoo](#), [Pandia](#), [Looksmart](#), [MSN](#)



- Deep web: [Invisible-web.net](#), [Completeplanet](#), [ResorceDiscovery Network](#)

Paolo Ferragina, Università di Pisa

Interesting Tools

- [Vivisimo](#)
 - Clustering engine
- [AskJeeves](#)
 - Question answering
 - Suggerimento termini
- [Froogle](#)
 - Cerca prodotti
- [Google Catalogs](#)
 - More than 6,000 catalogs
- [GoogleNews](#)
 - News service on-the-fly
 - More than 4,200 sources



Paolo Ferragina, Università di Pisa

Ricerche "verticali": <http://vlib.org/>

The collage displays four distinct web search interfaces:

- ResearchIndex (top-left):** The main page of the NECI Scientific Literature Digital Library. It features a navigation menu with 'main page', 'Publications', and 'Home page'. The text describes the library's goal to improve the dissemination and feedback of scientific literature. A 'Summary of ResearchIndex (also known as CiteSeer)' section highlights features like 'Autonomous Citation Indexing (ACI)', 'All cited documents', and 'Reference linking'.
- MedExplorer (top-right):** A Microsoft Internet Explorer window showing the MedExplorer website. It offers various medical information services, including 'Drug Health Insurance Quotes', 'Online Pharmacy', and 'Pharmaceuticals and other prescription drug info online!'. A 'Pharmaceuticals' table lists various drugs and their prices.
- LexisNexis (bottom-left):** The LexisNexis website, which provides legal, public records, company data, and government information. It includes a search bar and several filters to refine results, such as 'View by occupation', 'View by industry', and 'View by task'.
- NCBI Blast (bottom-right):** The NCBI Blast (Basic Local Alignment Search Tool) interface. It features a search input field, a 'Search' button, and options for selecting a database and filtering results.

Motori di Ricerca presente e futuro prossimo

Cosa è un motore di ricerca ?

Paolo Ferragina, Università di Pisa

