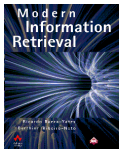
Recupero di Documenti

(Motori di Ricerca: presente e futuro prossimo)

Paolo Ferragina
Dipartimento di Informatica
Università di Pisa

Libri di testo

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

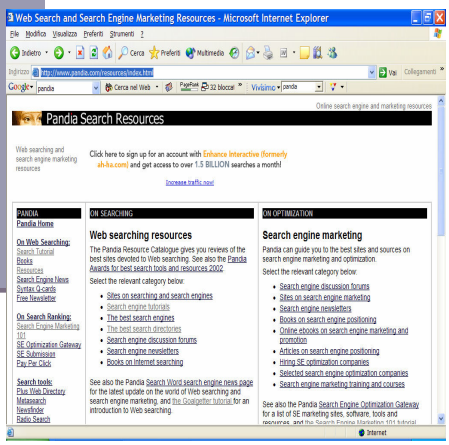


Modern Information Retrieval
R. Baeza-Yates e B. Ribeiro-Neto, Addison-Wesley, 1999.
"Libro di livello universitario sul progetto di un motore di ricerca"

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

Tre riferimenti interessanti

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



Motori di Ricerca

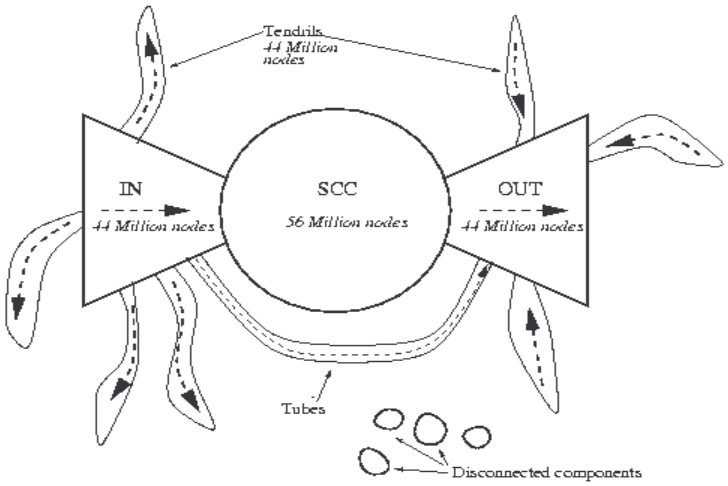
presente e futuro prossimo

Prologo

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

Una immagine pittorica del Web



Velocità di cambiamento

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

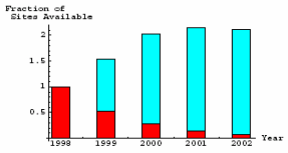


Figure 4: Percent of IP addresses identifying a Web site in Year A also identifying a Web site in Year B. For example, 56% of IP addresses identifying a Web site in the 1998 sample also identified one in 1999 sample. Taken from `http://wcp.oclc.org/`.

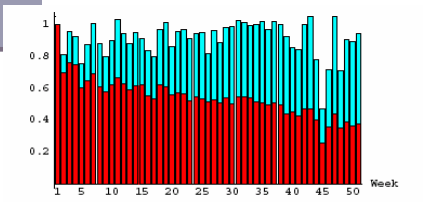


Figure 2: Fraction of pages from the first crawl still existing after n weeks (dark bars) and new pages (light bars).

Normalizzata rispetto prima settimana

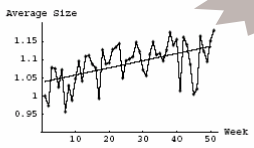


Figure 5: Average page sizes in our snapshots over time.

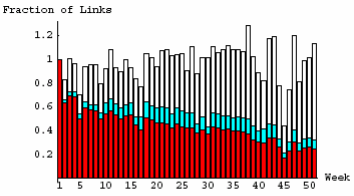


Figure 8: Fraction of links from the first weekly snapshot still existing after n weeks (dark/bottom portion of the bars), new links from existing pages (grey/middle) and new links from new pages (white/top).

Ma non solo il Web

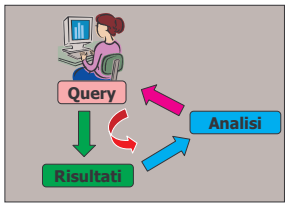
- Email:
  - 610Mld di messaggi al giorno
  - 150,000 Mailing List (circa 675Tb all'anno)
- Ogni anno: Libri (8 Tb), Quotidiani (25 Tb), Periodici (12 Tb), documenti di ufficio (210 Tb).

… 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"

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,

[...] 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... "

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

Corriere della Sera, 12 Aprile 2001

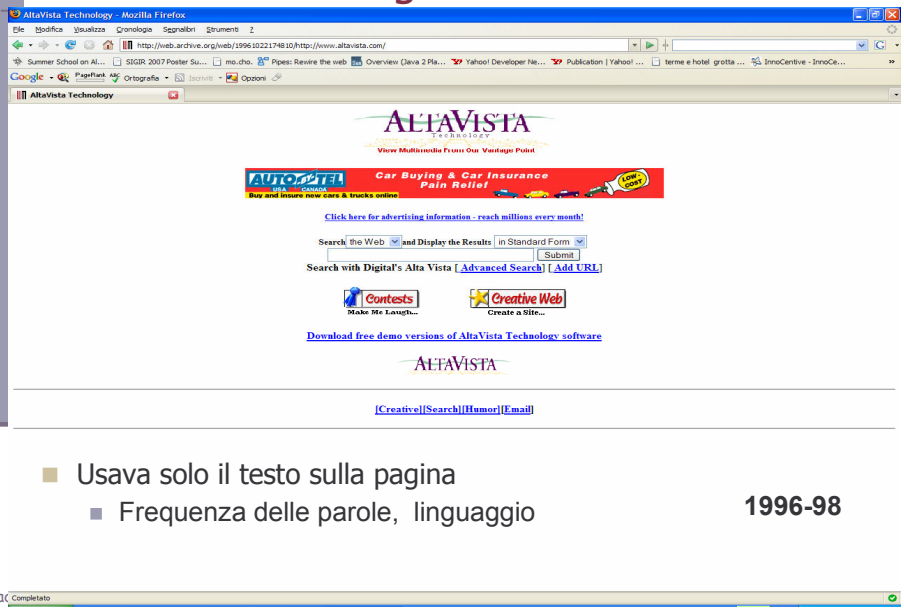
Interrogazioni sul Web: vari tipi

- 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"

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

La storia: Prima generazione



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

1996-98

La storia: Seconda generazione

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

1998-0?

La storia: Terza generazione

- focus sull'utente
  - Analisi semantica, Contesto
  - Aiuti, click-through
  - Integrazione ricerca e browsing **oggi**

Quarta generazione ???

Quarta generazione → Information Supply

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

Ieri.....

Oggi...

Ieri...

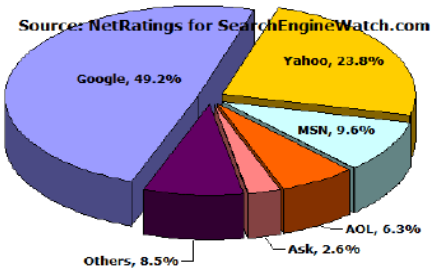
```
> finger karger@CSAIL.MIT.EDU
[CSAIL.MIT.EDU]
KARGER   David Karger         Theory of Computation          Faculty
         <karger@theory.lcs.mit.edu>
         Project: Analysis of Algorithms and/or Information Retrieval
         Work: NE43-321; 258-6167
         Home: 1600 Mass. Ave., Apt. 407,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

Tutti questi Tool usano un Search Engine

...oggi

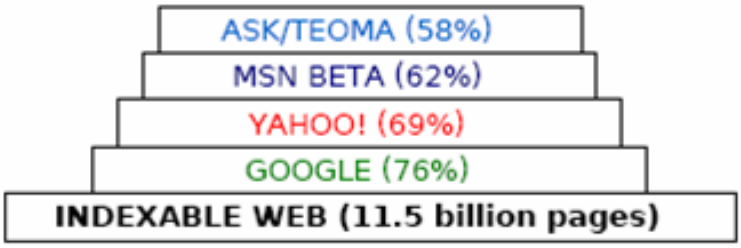
Perché tanto interesse sui "motori" ?

- 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 -

Dimensione reale [2005]



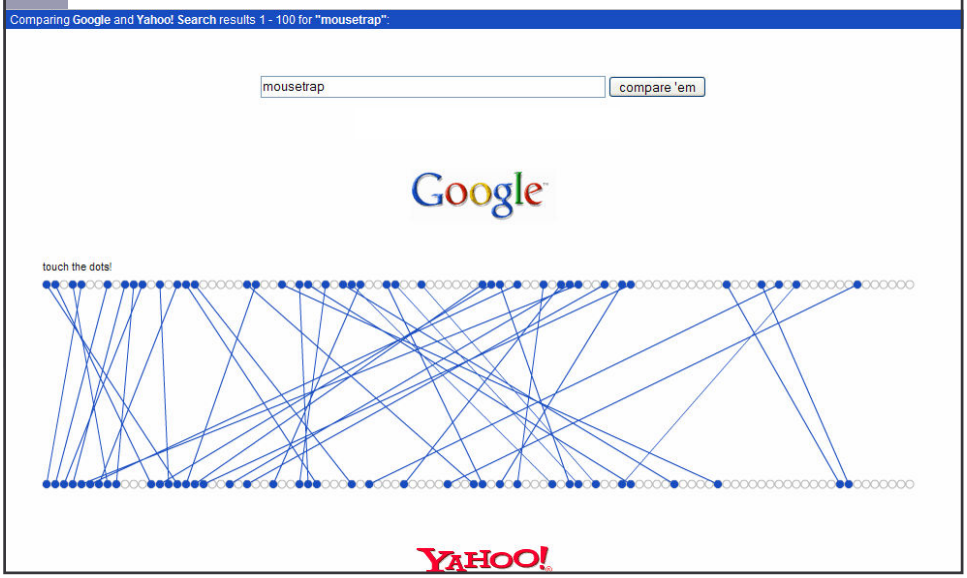
Google vs Yahoo:
20-30% risultati identici

La visione globale

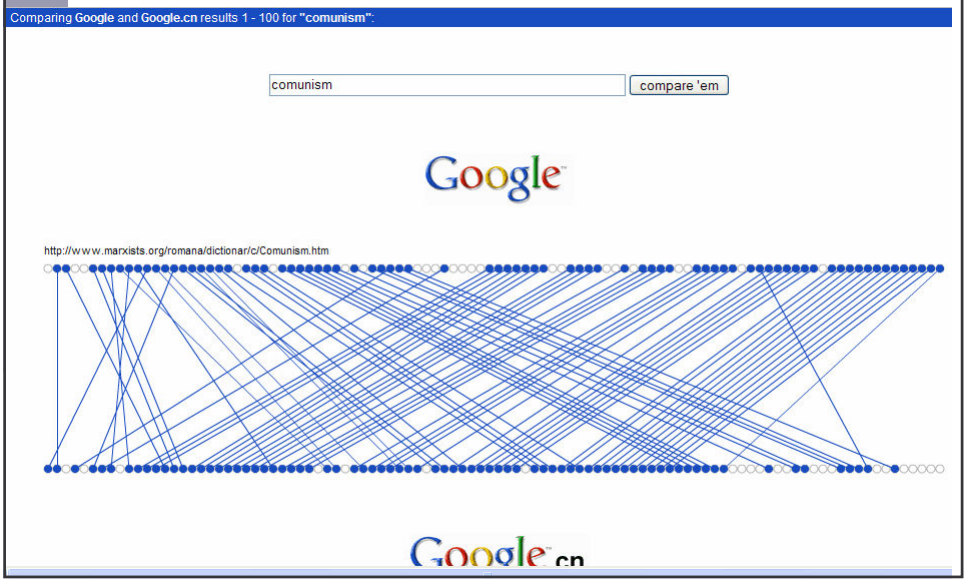
G = Google
M = Msn
T = ASK/Teoma
Y = Yahoo!

Paolo

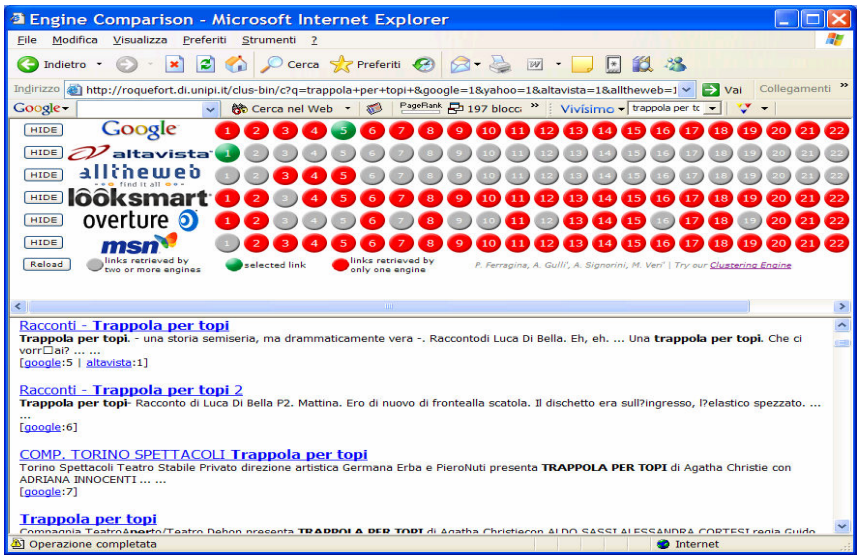
Ranking: Google *vs* Yahoo!



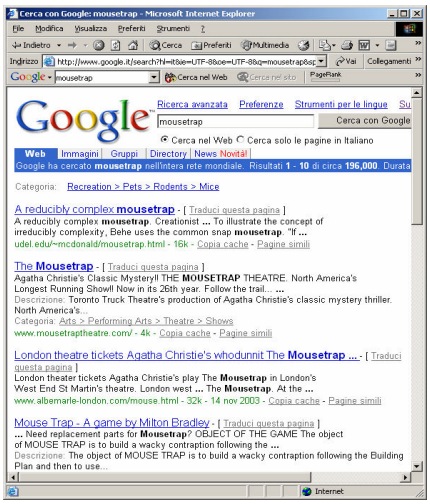
Ranking: Google *vs* Google.cn

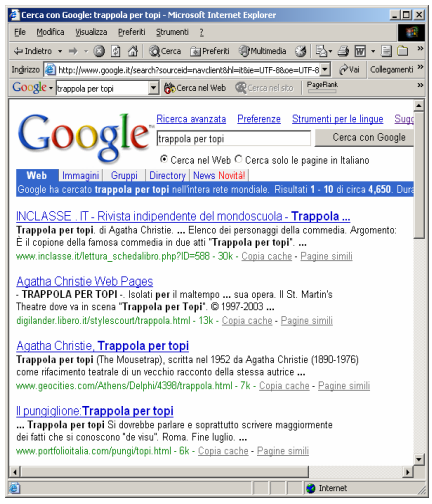


Più confronti



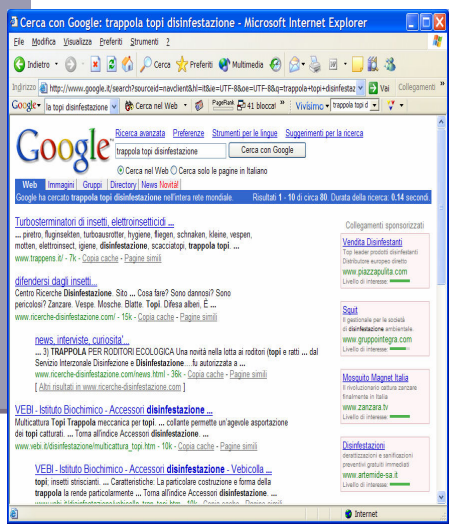
Formulare una interrogazione (quante difficoltà)

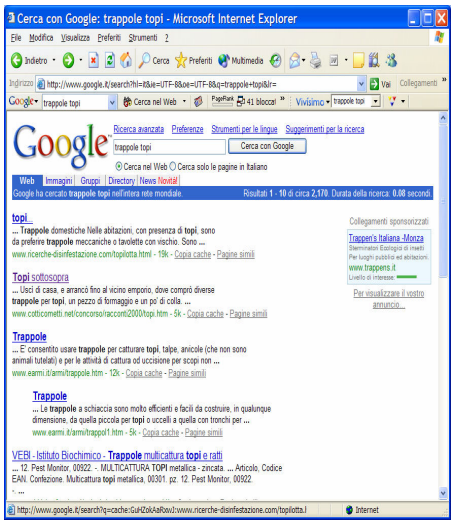




- Varie difficoltà: problemi di astrazione, sinonimia, polisemia, 10 risultati

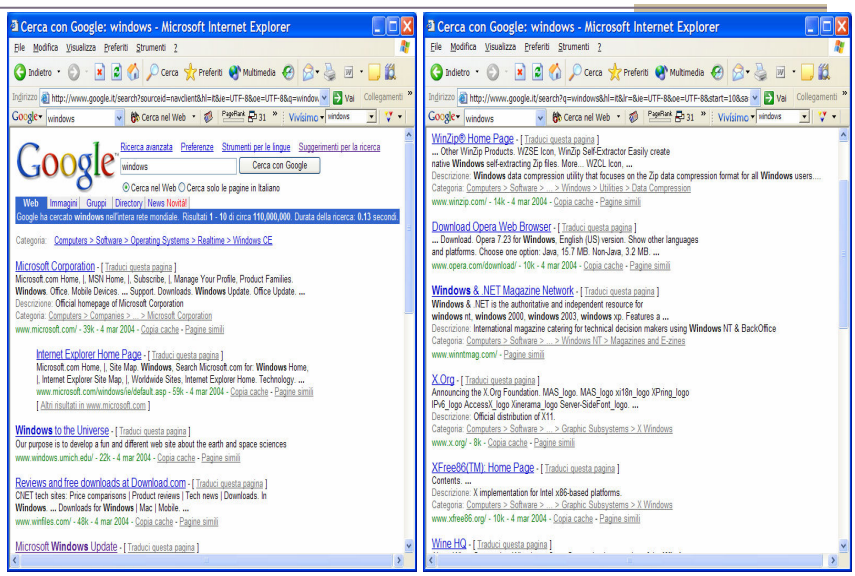
Formulare una interrogazione (quante difficoltà)



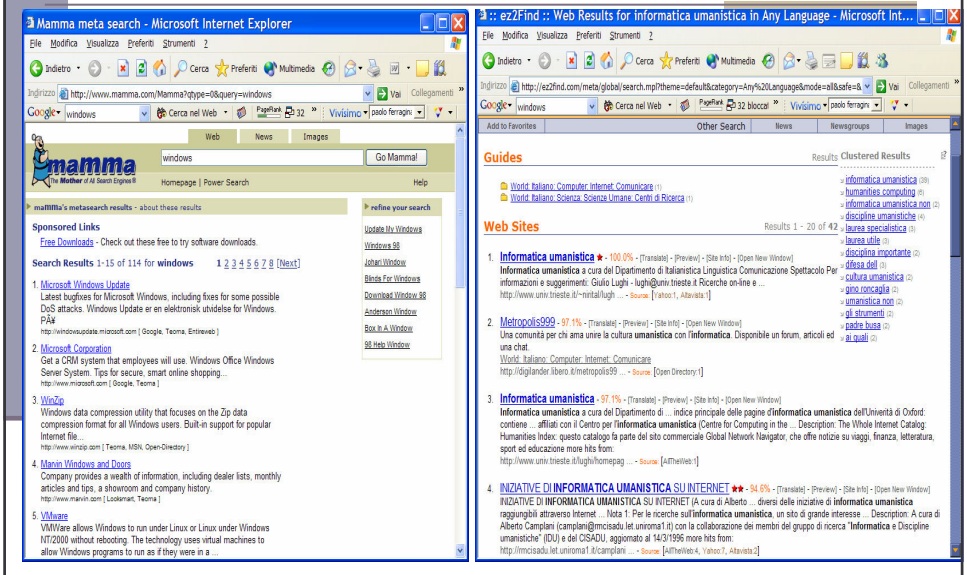


- Varie difficoltà: problemi di astrazione, sinonimia, polisemia, 10 risultati

Formulare una interrogazione (quante difficoltà)

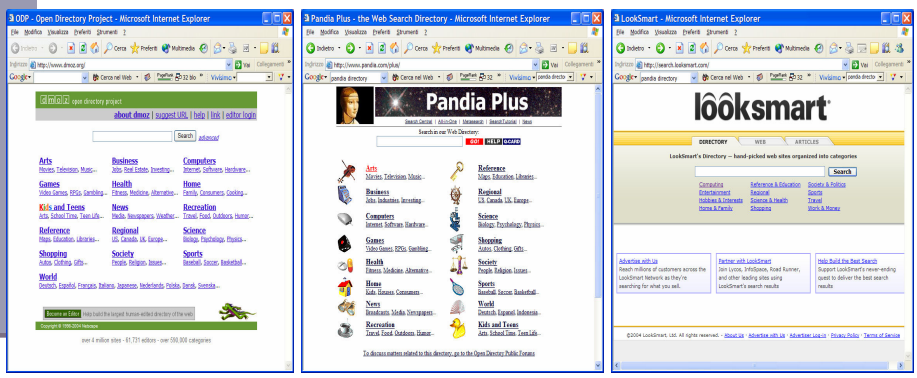


Tipi di ricerche: Meta-Motori



Tipi di ricerche: Directory

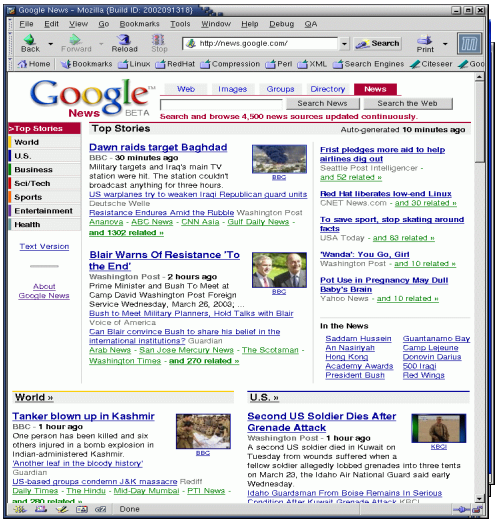
- DMOZ, Yahoo, Pandia, Looksmart, MSN



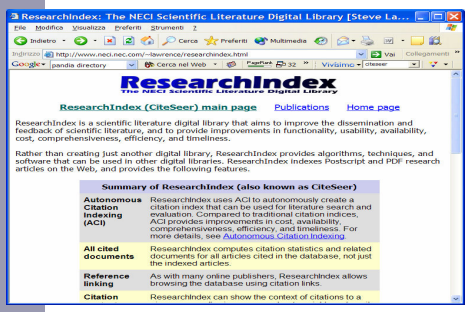
- Deep web: Invisible-web.net, Completeplanet, ResoruceDiscovery Network

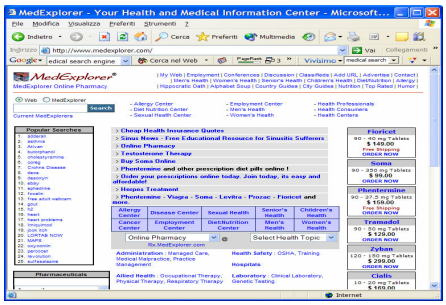
Interesting Tools

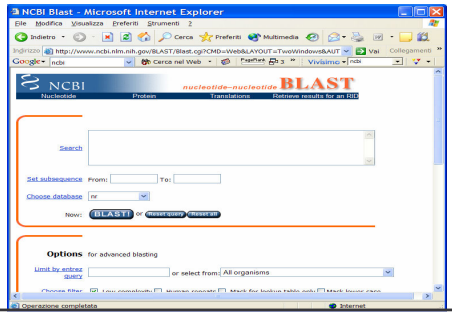
- Vivisimo
  - Clustering engine
- AskJeeves
  - Question answering
  - Suggerimento termini
- Froogle
  - Cerca prodotti
- Google Catalogs
  - More than 6,000 catalogs
- GoogleNews
  - News service on-the-fly
  - More than 4,200 sources



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







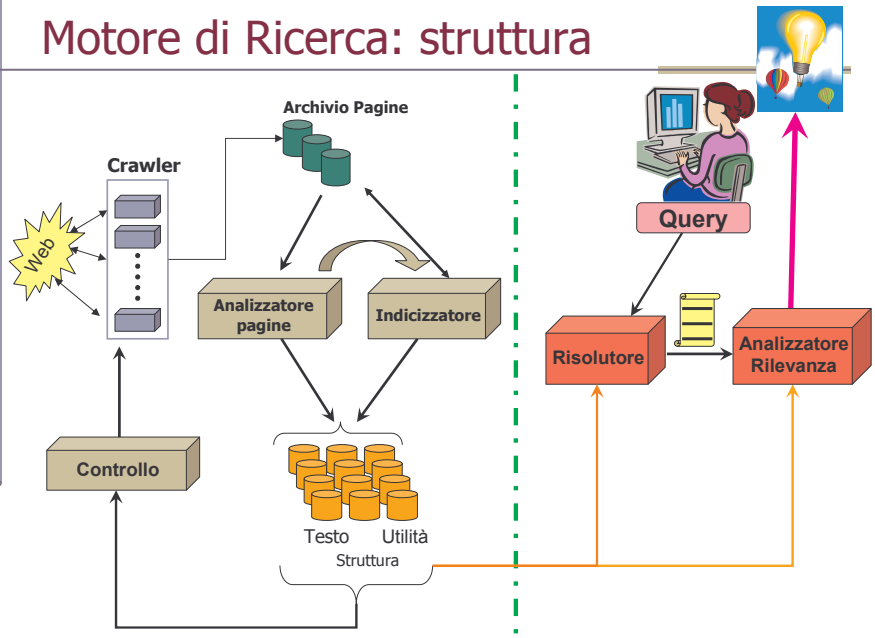
Motori di Ricerca

presente e futuro prossimo

Cosa è un motore di ricerca ?

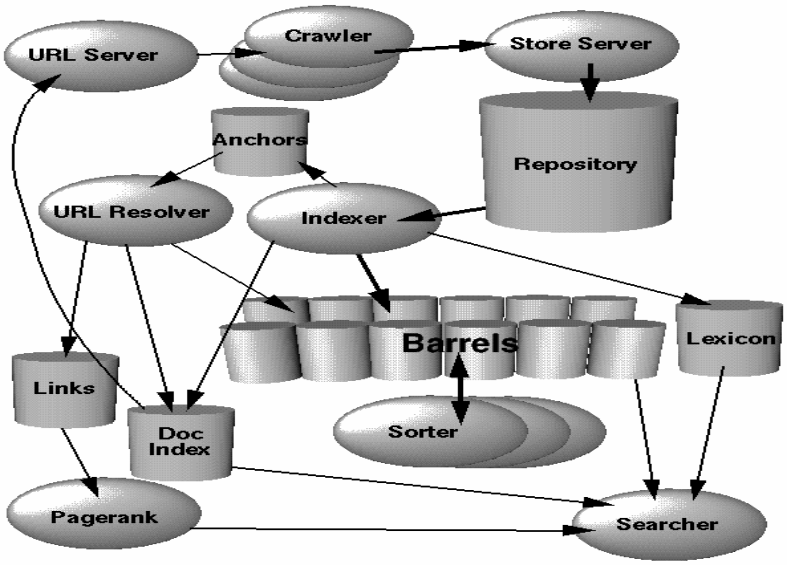
Paolo Ferragina, Università di Pisa

Motore di Ricerca: struttura



Paolo Ferragina, Università di Pisa

Un lavoro storico: Brin & Page [1998]



Paolo Ferragi