

# Metodologie di analisi e caratterizzazione della classe di utenza del servizio di Emeroteca Virtuale al CASPUR\*

Marco Scarnò (mscarno@caspur.it)  
Donatella Sforzini (d.sforzini@caspur.it)

Bibliocom 2004  
Roma, 27 Ottobre

\*Consorzio interuniversitario CASPUR, Università di Roma *La Sapienza*

# Ambito della Ricerca

The screenshot shows the homepage of the Emeroteca Virtuale website. The browser window title is "Emeroteca Virtuale - Periodici Elettronici Full Text - Home Page - Microsoft Internet Explorer". The address bar shows "http://periodici.caspur.it/". The website header includes the logo "EV Emeroteca Virtuale" and "CASPUR" with the tagline "Consorzio interuniversitario per le Applicazioni di Supercalcolo per l'Università e Ricerca". A navigation menu contains links for "Helpdesk", "Contatti", "Area Riservata", "L'Emeroteca Virtuale", "Il Ciber", and "Il Caspur". A statistics box reports: "Dati statistici al Mon Oct 25 00:34:57 MEST 2004 Journals: 3,593 Issues: 190,149 Articles: 3,239,704". A yellow banner indicates "31 Marzo 2004: Attivato il servizio di link da PUBMED." and "Archivio News".

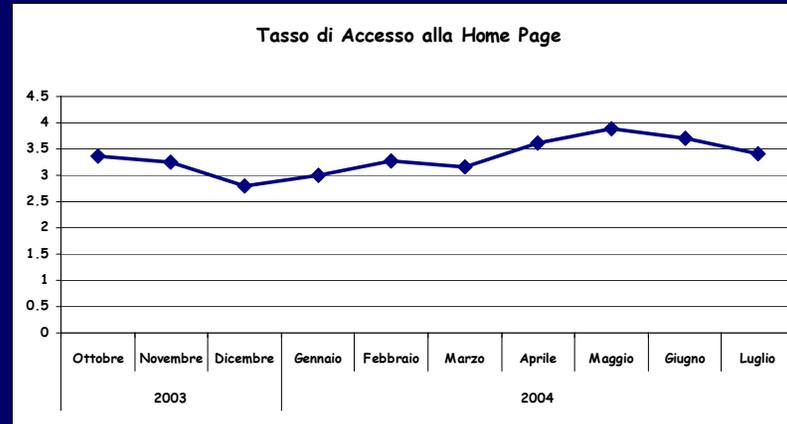
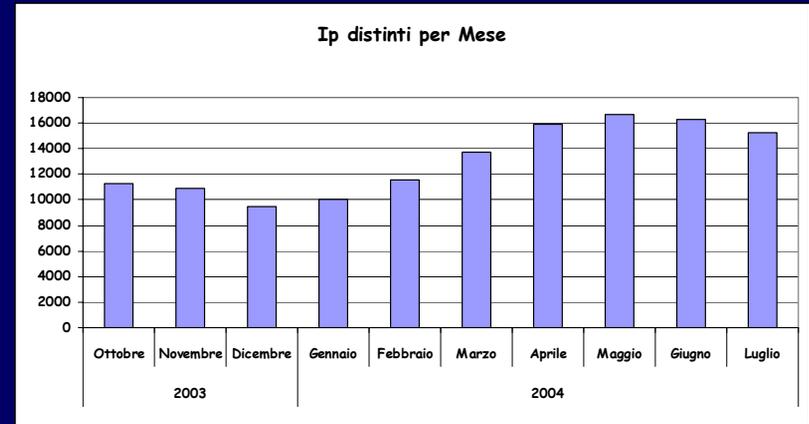
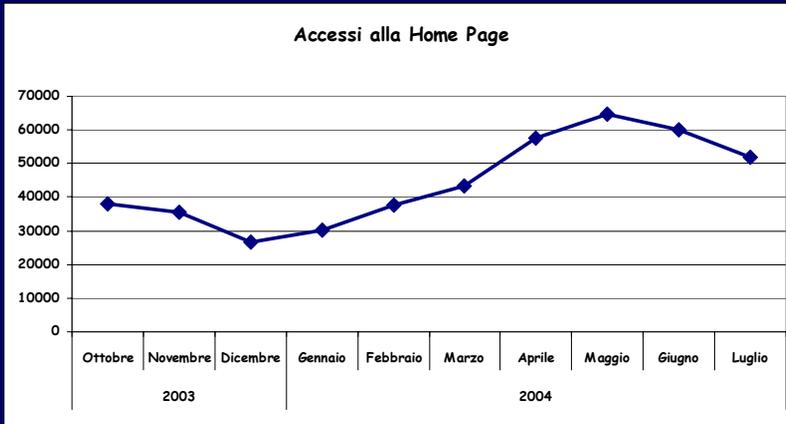
Modalità di Ricerca	Riviste nell'Emeroteca Virtuale	Riviste sul sito dell'Editore
<b>Simple Search</b> ricerca articoli per mezzo di parole chiave nel titolo, abstract ecc.	Elenco delle riviste full-text presenti in Emeroteca per ciascun Editore. :: <a href="#">American Chemical Society</a> 32 riviste link al full-text [...More Info]	Accesso alle riviste direttamente sul sito dell'Editore. :: <a href="#">American Chemical Society</a> 32 riviste [...More Info]
<b>Advanced Search</b> ricerca articoli per mezzo di parole chiave nel titolo, abstract e molto di più		:: <a href="#">Blackwell Publishing</a> 672 riviste [...More Info]
<b>Categorie</b> elenco per categorie dei periodici disponibili		:: <a href="#">Institute of Physics</a> 41 riviste [...More Info]
<b>Elenco Alfabetico</b> elenco alfabetico dei periodici disponibili	:: <a href="#">Elsevier Science</a> 1921 riviste full-text [...More Info]	:: <a href="#">Kluwer Academic Publisher</a> 700 riviste [...More Info]
<b>Servizi Avanzati</b>		:: <a href="#">Wiley InterScience</a> ... riviste [...More Info]
<b>Servizio di Accesso Remoto</b> consulta l'Emeroteca da qualsiasi computer del mondo		:: <a href="#">JSTOR (journal storage)</a> 400 riviste [...More Info]
<b>Servizio di Ricerca Personalizzata</b> memorizza le tue ricerche e ricevi i risultati tutti i giorni via email		:: <a href="#">EIO - Editoria Italiana Online</a> Trial 600 pubblicazioni [...More Info]
<b>Modulo di registrazione</b> registrati ai servizi di Accesso Remoto & Ricerca Personalizzata		

**Area Ente**  
Cataloghi disponibili presso il Caspur:  
● [Web of Science](#)  
● [UlrichsWeb](#)  
● [Catalogo d'Ateneo RomaTRE](#)

Two blue callout boxes are overlaid on the page:  
1. "Più di 3,500 Riviste" (More than 3,500 Journals)  
2. "Più di 3,200,000 Articoli" (More than 3,200,000 Articles)

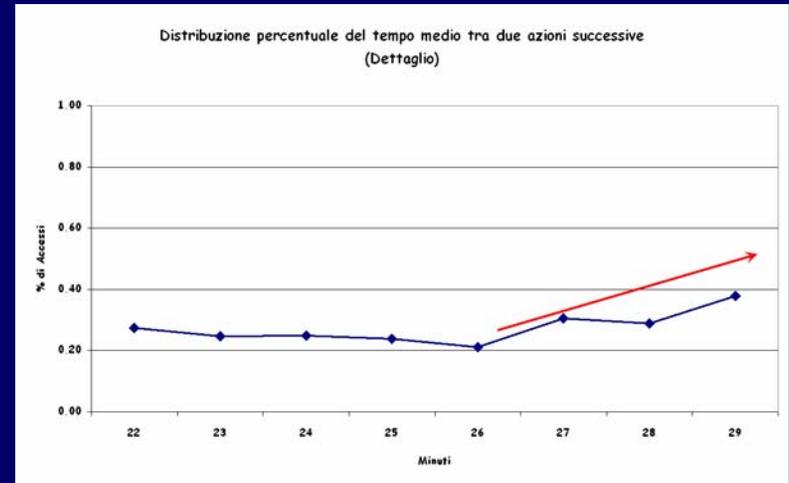
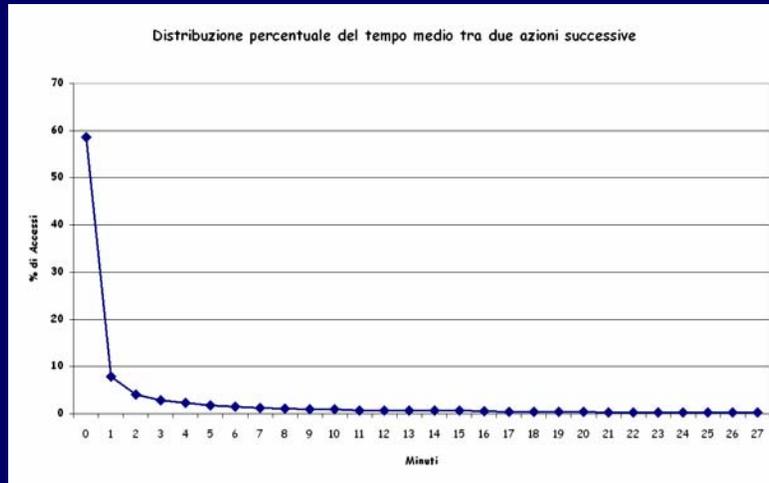
# Gli accessi al sito...

..da Ottobre 2003 a Luglio 2004:



Ma.. quando si è entrati nel sito?

# ..accade che l'IP (il computer) compia più azioni successive...



Lasciando questa *traccia...*

19x.16y.10z.23t - - [01/Nov/2003:07:30:28 +0100] "GET /cgi-

bin/search.pl?Database=journals&search\_field=oncogenes+and+bladder+cancer&GetSearchResults=Search&fields=An  
y&canned\_search= HTTP/1.1" 200 42145

# Definizioni:

- *Contenuti*: insieme delle pagine che il server restituisce all'utente; possono essere o l'indice di una rivista, o l'indice dell'issue, o un abstract o un full text (articolo di una rivista);
- *Attività di Browsing*: insieme delle pagine web di ricerca degli articoli basate sulla visualizzazione degli elenchi delle riviste per lettera alfabetica, per argomento o per editore;
- *Attività di Ricerca*: insieme delle pagine web di ricerca che consentono all'utente di trovare un articolo tramite delle parole chiave;
- *Sessione di Ricerca*: la sessione durante la quale un utente cerca un articolo (si è identificata una nuova sessione ogni qual volta l'utente rimane inattivo per 27 minuti);

# Problema: esiste un metodo *ufficiale* per trattare i dati?

Obiettivo: eliminare i dati non validi

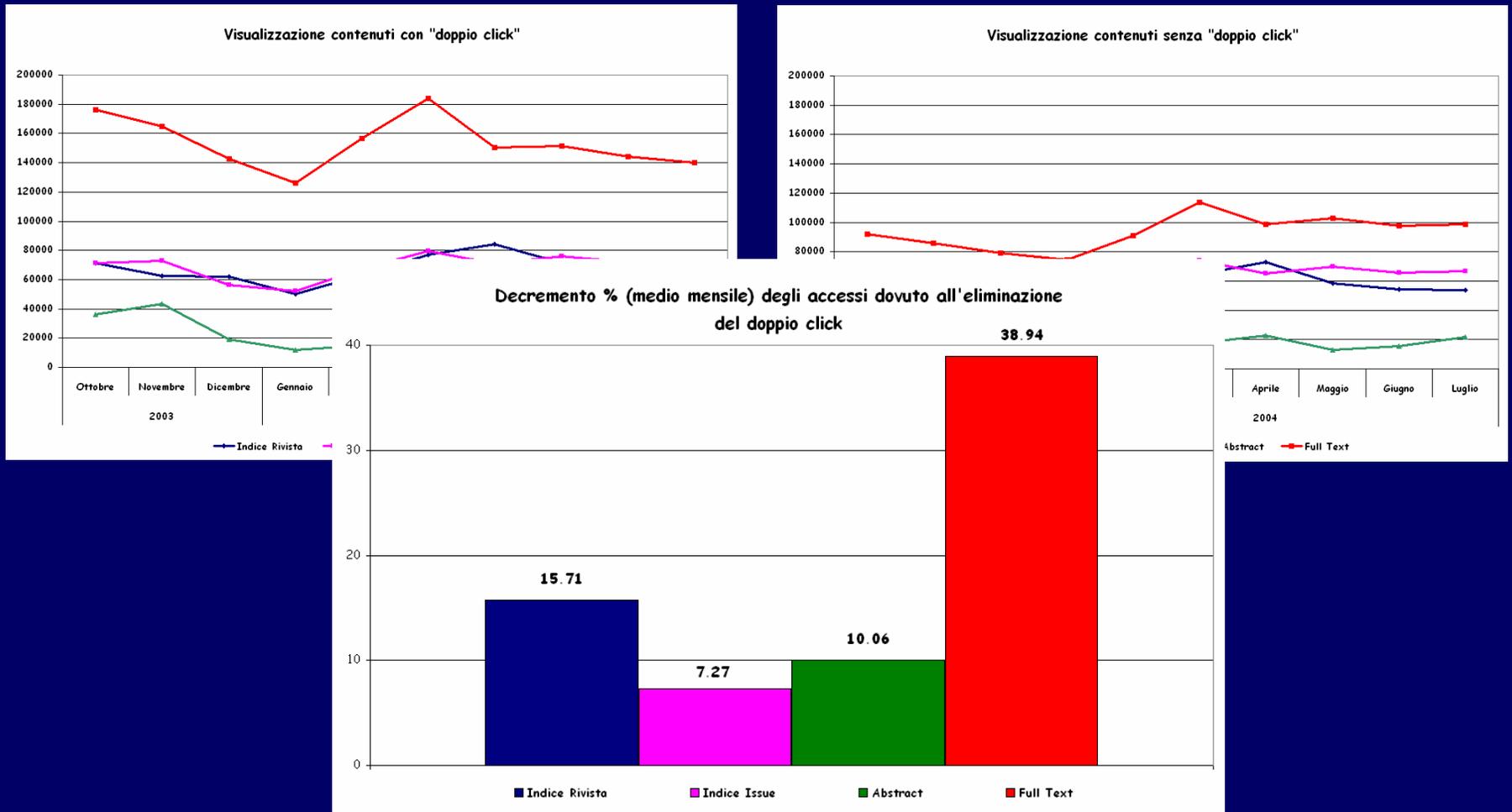


- Eliminazione dei doppi *click*;
- Eliminazione dei *comportamenti anomali*;



Fornire rapporti e analisi non falsati da valori anomali

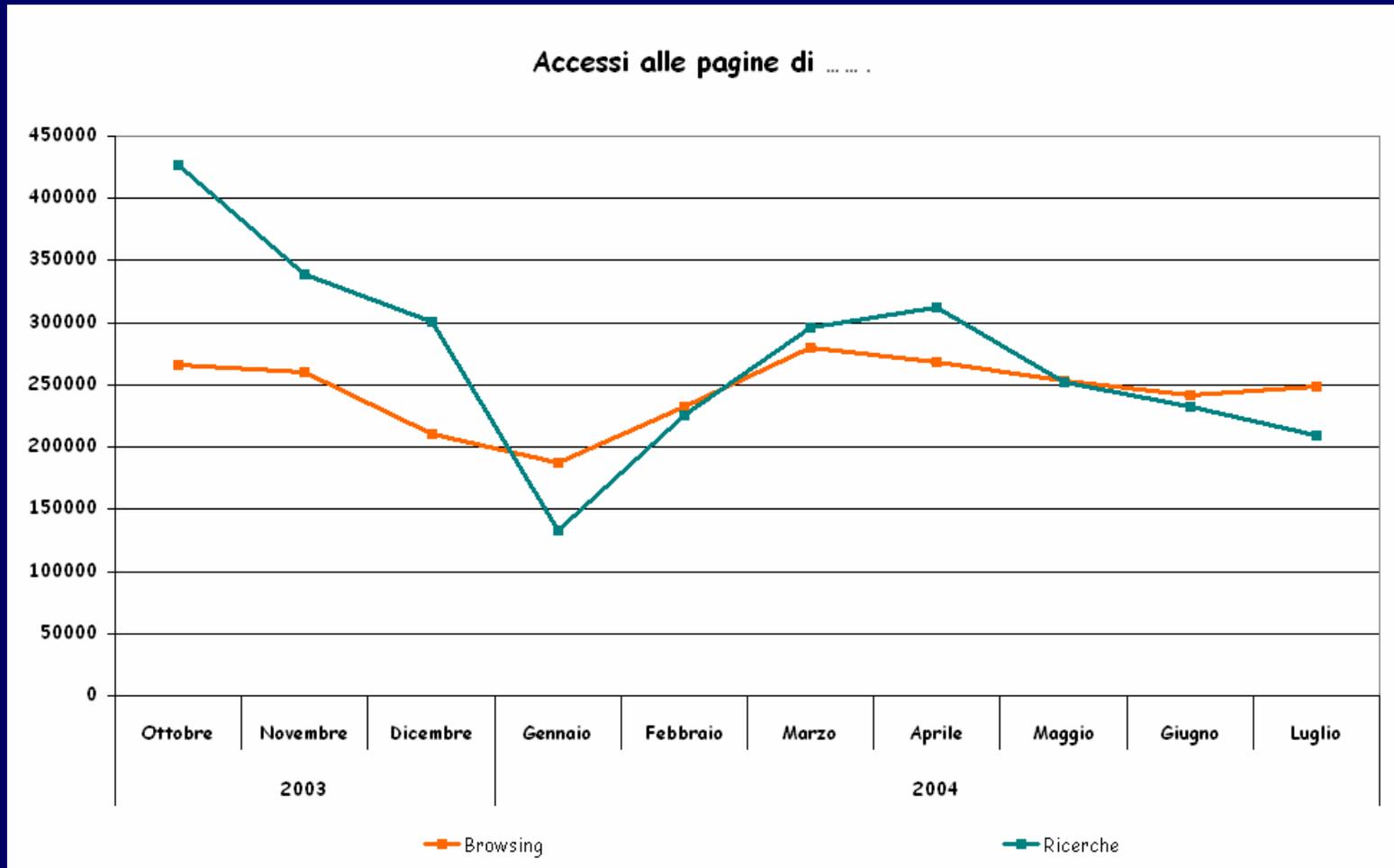
# Pre-trattamento dei dati: il problema del doppio click



# Le attività possibili (e realmente effettuate) sui dati:

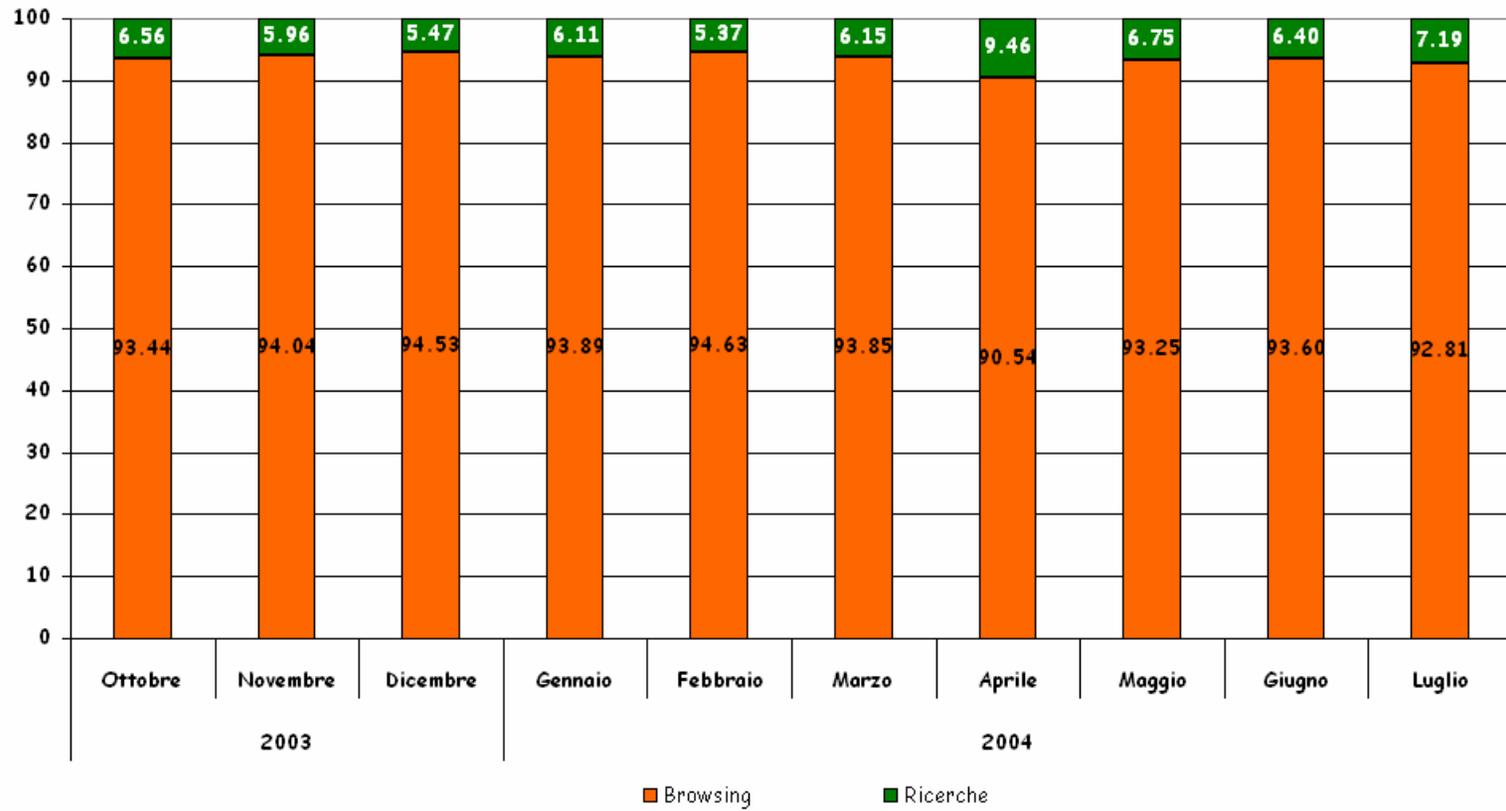
- Istituzionale (rapporti sulle consultazioni ad uso degli *utenti-responsabili*)
- Descrittiva:
  - Attività di browsing
  - Attività di ricerca tramite parole chiave
  - Visualizzazione di Full Text

# Come cerco gli articoli che mi interessano? Browsing o Ricerche?



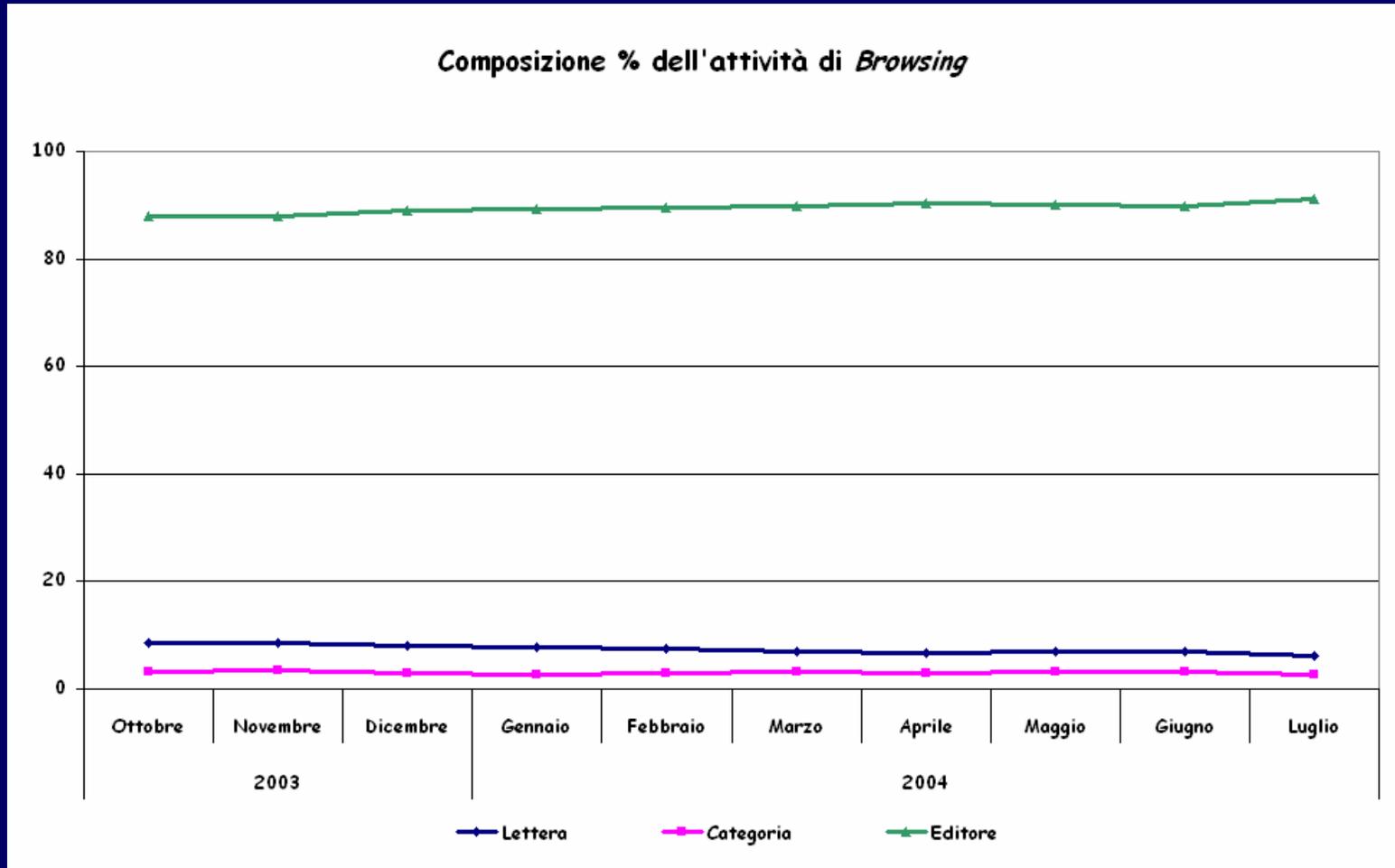
# ...e quale "strategia" mi aiuta di più a visualizzare gli articoli che cerco?

Articoli Full Text visualizzati raggiunti da pagine di ...  
(valori percentuali)



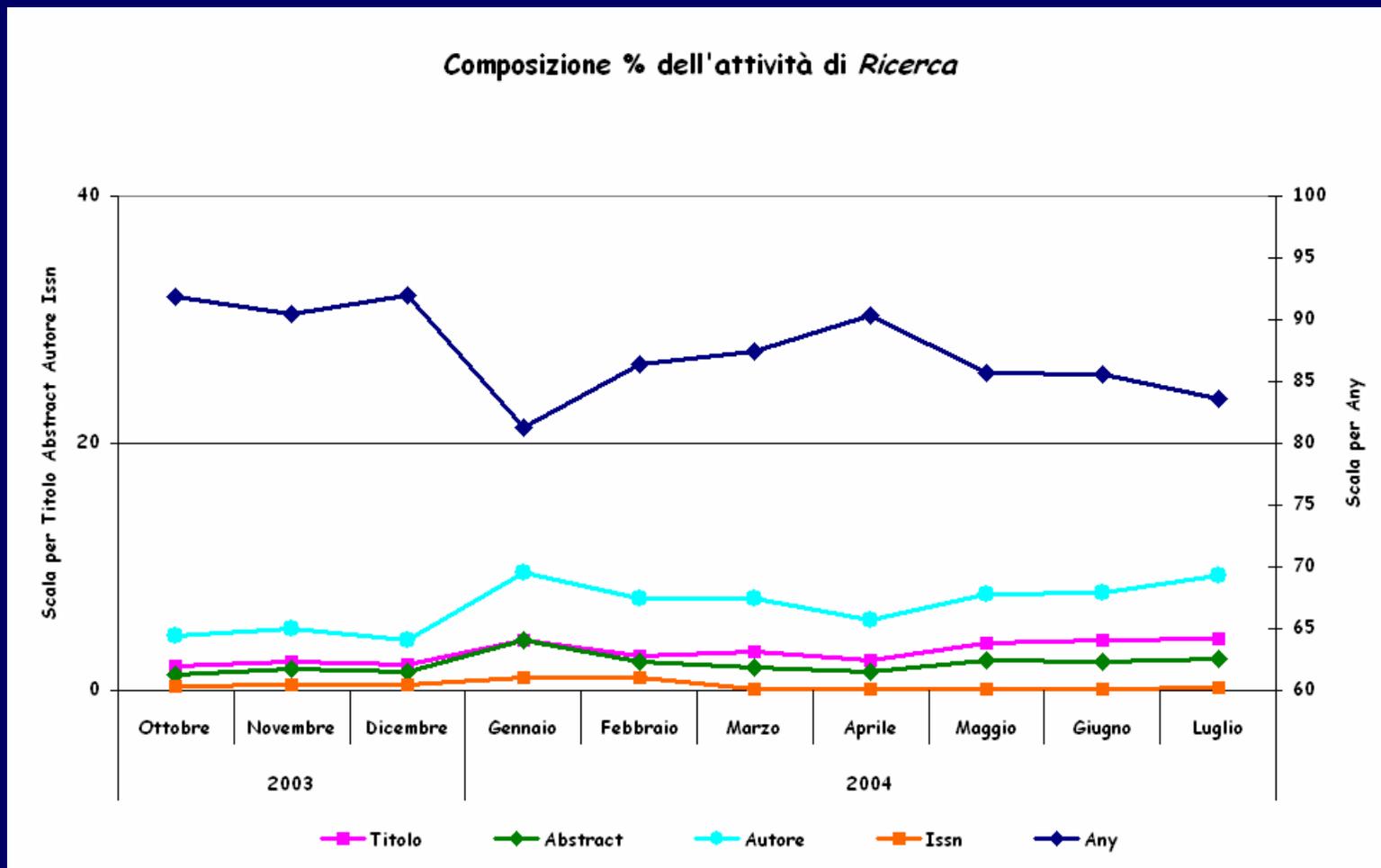
# In particolare nel browsing:

(solo una ricerca su 4 cominciata tramite browsing prosegue con il browsing stesso)



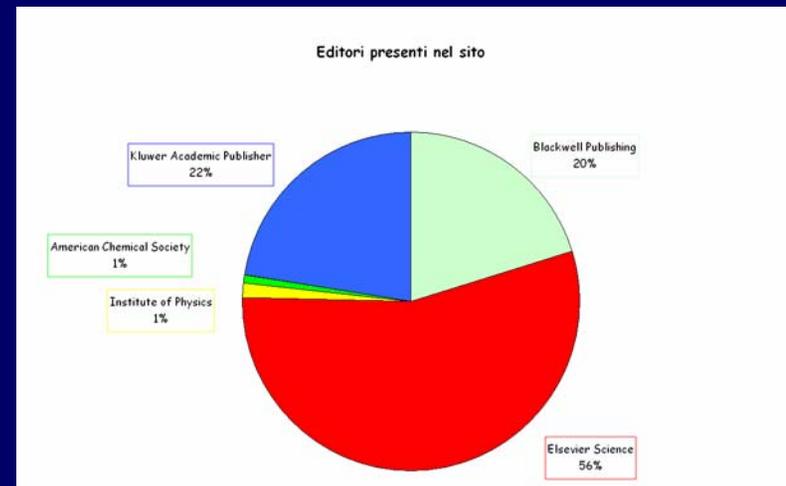
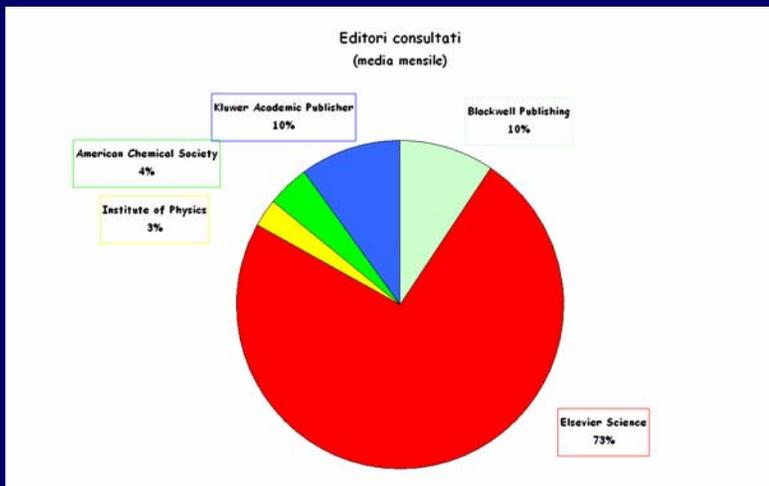
# ... e con le ricerche per parola chiave:

(quasi tutte le ricerche per chiave proseguono con..... *ricerche per chiave*)



## Problema metodologico:

le analisi possono essere riferite o al valore assoluto delle azioni effettuate o alle possibilità che l'emeroteca offre..  
...esempio..



*Elsevier Science o American Chemical Society?*

Base: attività di browsing per editore nel periodo

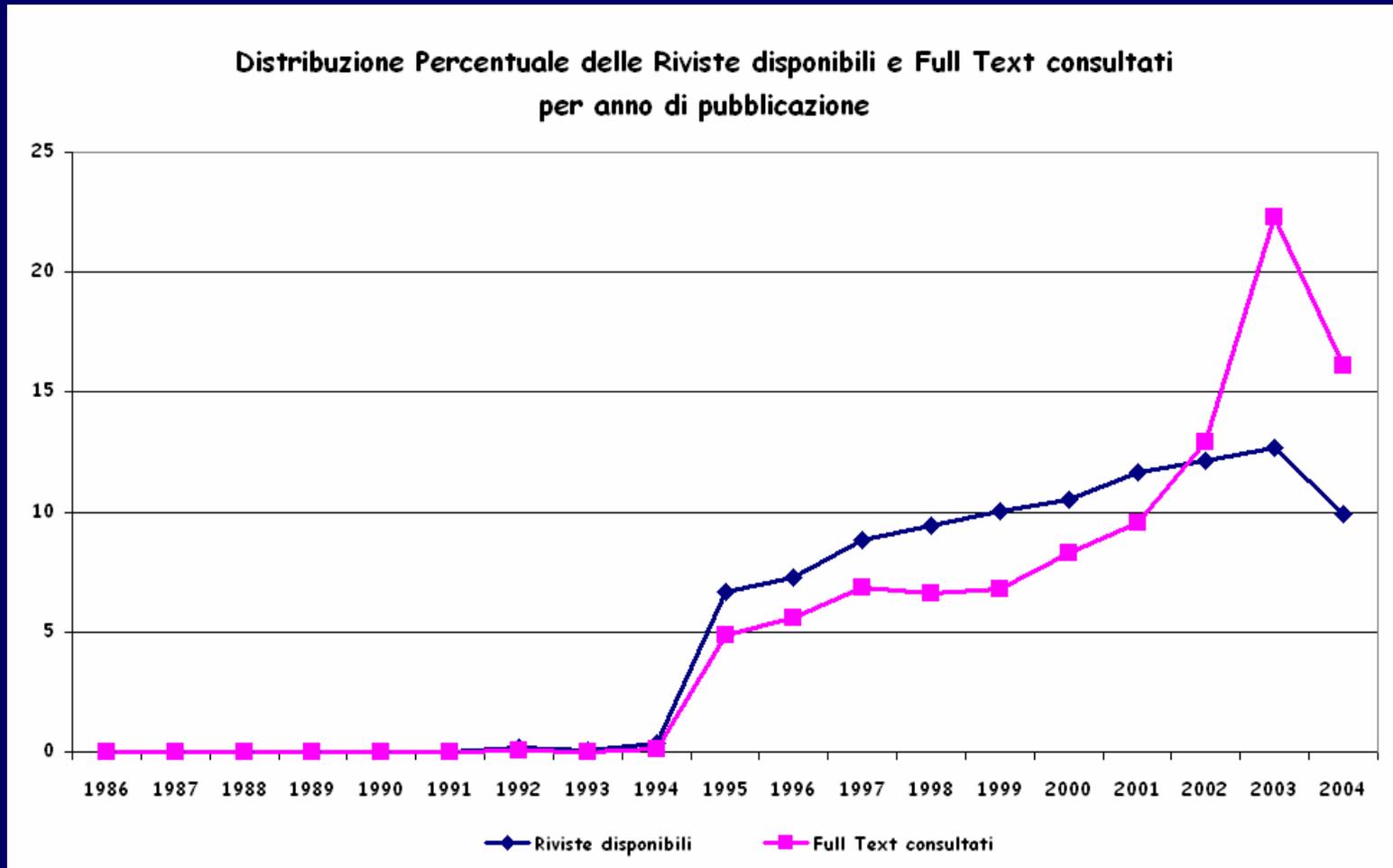
# Curiosità..

2003		
Ottobre	Novembre	Dicembre
"TOURISM"	"HEAVY OIL"	"PLAUSIBILITY"
"STYRENE AND BUTADIENE"	"PRIME NUMBER"	"SUBMODULAR"
"EBOLA"	"GEOGRAPHY"	"TESI "
"TABU SEARCH"	"MAYERS"	"DEPRESSION AND MIGRAINE"
"JOURNAL OF THERMAL ANALYSIS"	"HEAT ISLAND"	"MONTE CUCCO"
"MORTALITY"	"ADOPT"	"CU-AL-NI"
"MERCURY POLLUTION"	"HEALTH SERVICE RESEARCH"	"ISLAM"
"BOUCHAUD"	"HASSANEIN"	AUTHOR=(WARD)
"BITE"	AUTHOR=(HENSEL)	"SYRIA"
"MANHATTAN"	"DOPING"	"BREAKWATER"

2004						
Gennaio	Febbraio	Marzo	Aprile	Maggio	Giugno	Luglio
"TRENDS IN IMMUNOLOGY"	"SCHIZOPHRENIA RESEARCH"	"BLINDNESS"	"USABILITY"	"VERTEBRAL OSTEOMYELITIS"	"BURNOUT"	"RORSCHACH"
"THERMAL DEGRADATION"	"E-COMMERCE "	"BECK"	"CORPORATE GOVERNANCE"	"LANGEREIS"	"IMPACT FACTORS"	"HYDROGEL CONTACT LENSES"
"APPENDECTOMY"	"GAS HYDRATES "	"TURNER"	"BIOELECTROCHEMISTRY"	"CHLAMYDIA PNEUMONIAE"	"PROCEEDINGS OF THE SP"	"JOURNAL OF ANIMAL ECOLOGY"
CELL	"CHLAMYDIA"	"CARBON/CARBON"	"GIORDANO S"	"HEPATOCELLULAR CARCINOMA"	"EMDR"	AUTHOR=(("Valsecchi"))
"OSMOSIS"	"AMERICAN ECONOMIC REVIEW"	"ONCOGENE"	"PEGYLATION"	"INDIAN J. CHEM."	"REED"	"BREAST UNIT"
	"EMAIL"	"NEUROLOGY"	"MIKLAVCIC"	"BLOOM SYNDROME"	"DONATIIL"	"PNAS"
	"CAPSICUM"	"ATP"	"MONTELEONE F"	"GDNF"	"ROWING"	"NURSING"
	"ALCHEMY"	"CANCER EPIDEMIOLOGY"	"CITATION INDEX"	"DARK MATTER"	"ATKINSON"	"ENDORIBONUCLEASES"
	"CERAMIC GLASS"	"ASTHMA"	"CIRCULATION"	"NEPHRON"	"THE USE OF TRANSFORMM"	"X-RAY FLUORESCENCE SPECTROSCOPY"
	"NEWTON"	"CURRENT ACCOUNT"	"ANTICANCER"	"MINERVA"	"DIAPHRAGM WALLS"	

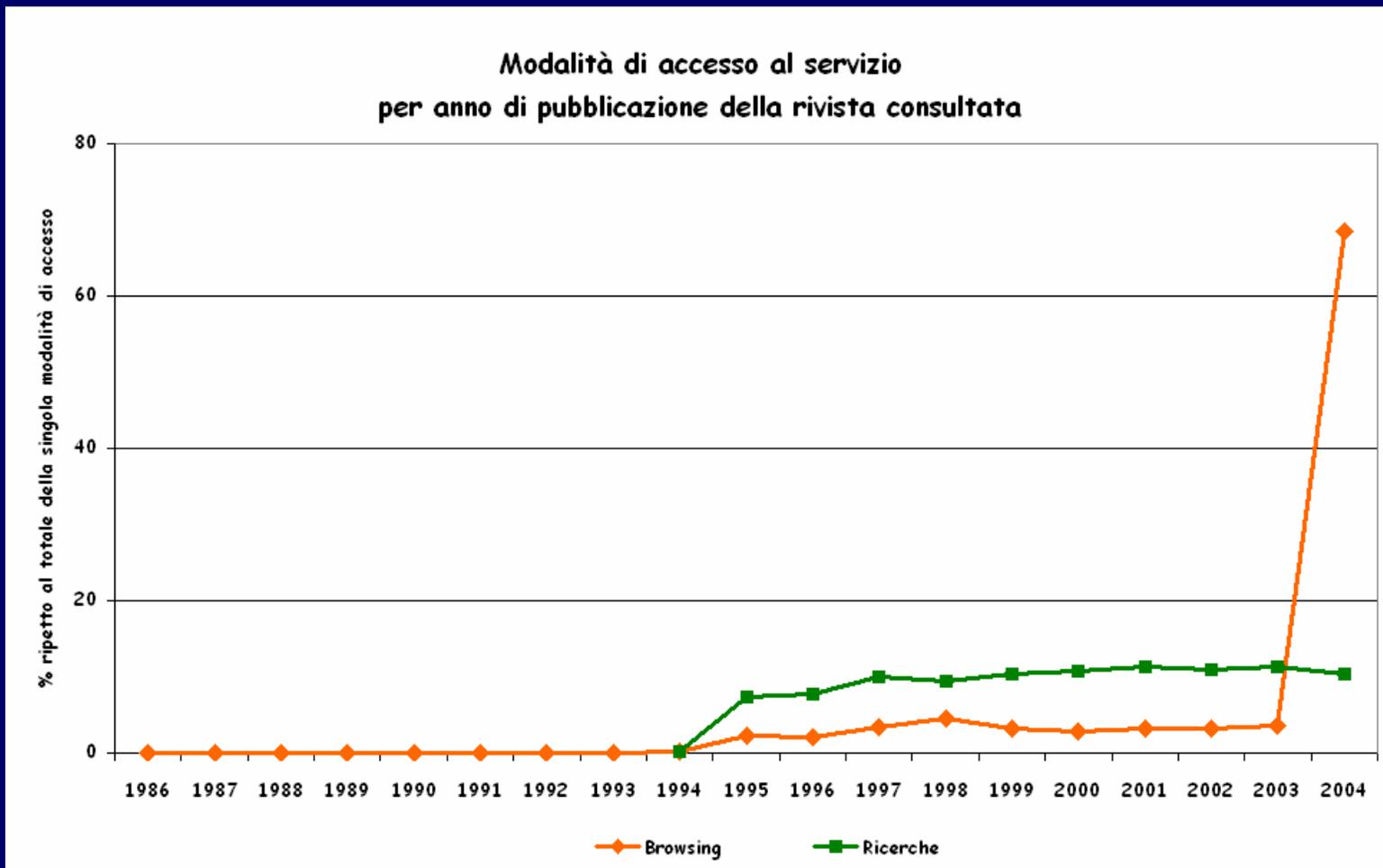
10 parole più cercate  
(sono state digitate più di 10,000 volte in un mese)

A questo punto si possono analizzare i comportamenti... e verificare se gli utenti si tengono aggiornati o se cercano rispetto interessi specifici



Periodo: Ottobre 2003- Luglio 2004

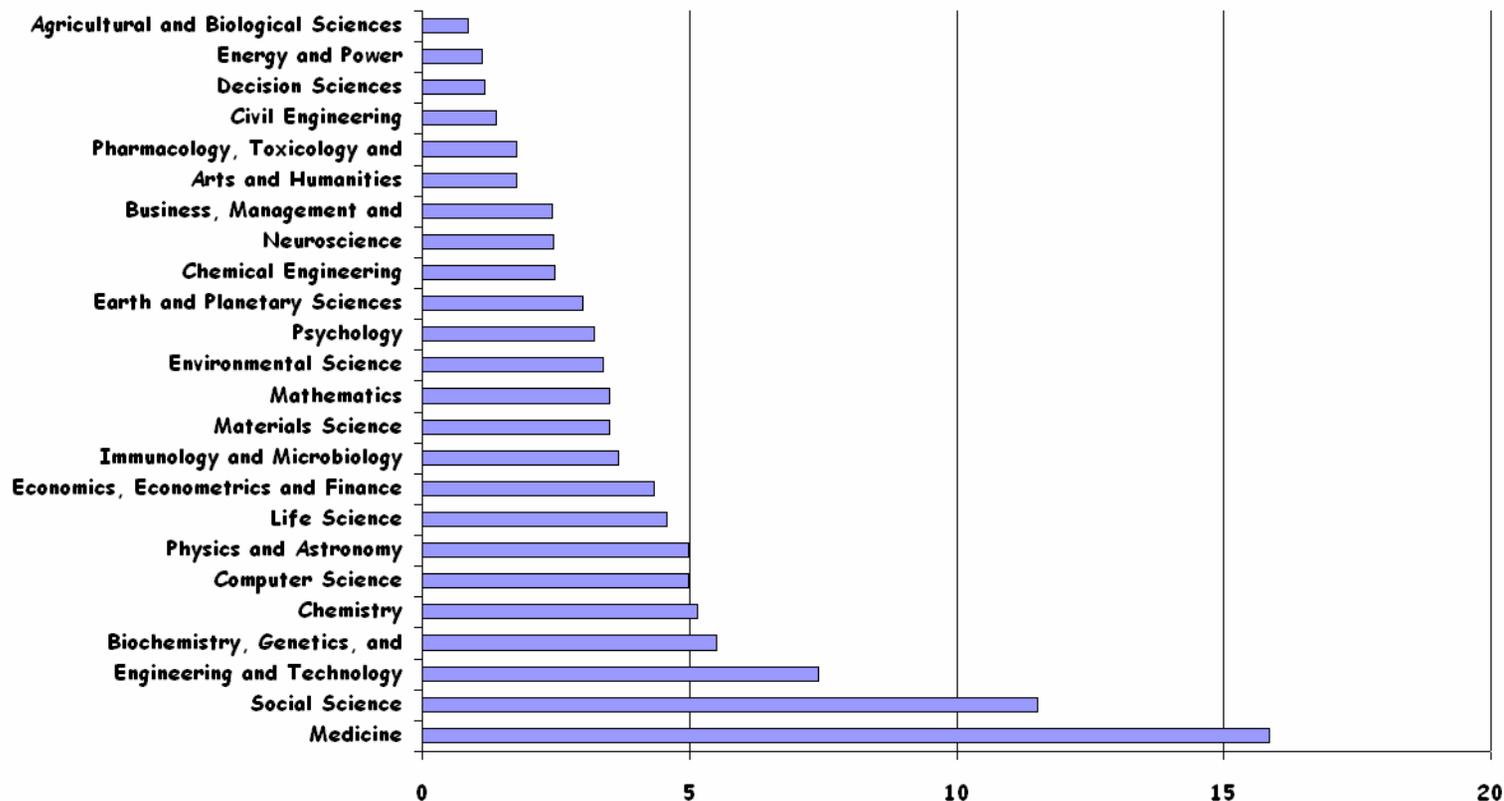
... anche rispetto alla modalità di ricerca...



Periodo: Ottobre 2003- Luglio 2004

# Ancora una volta il problema metodologico: riviste disponibili

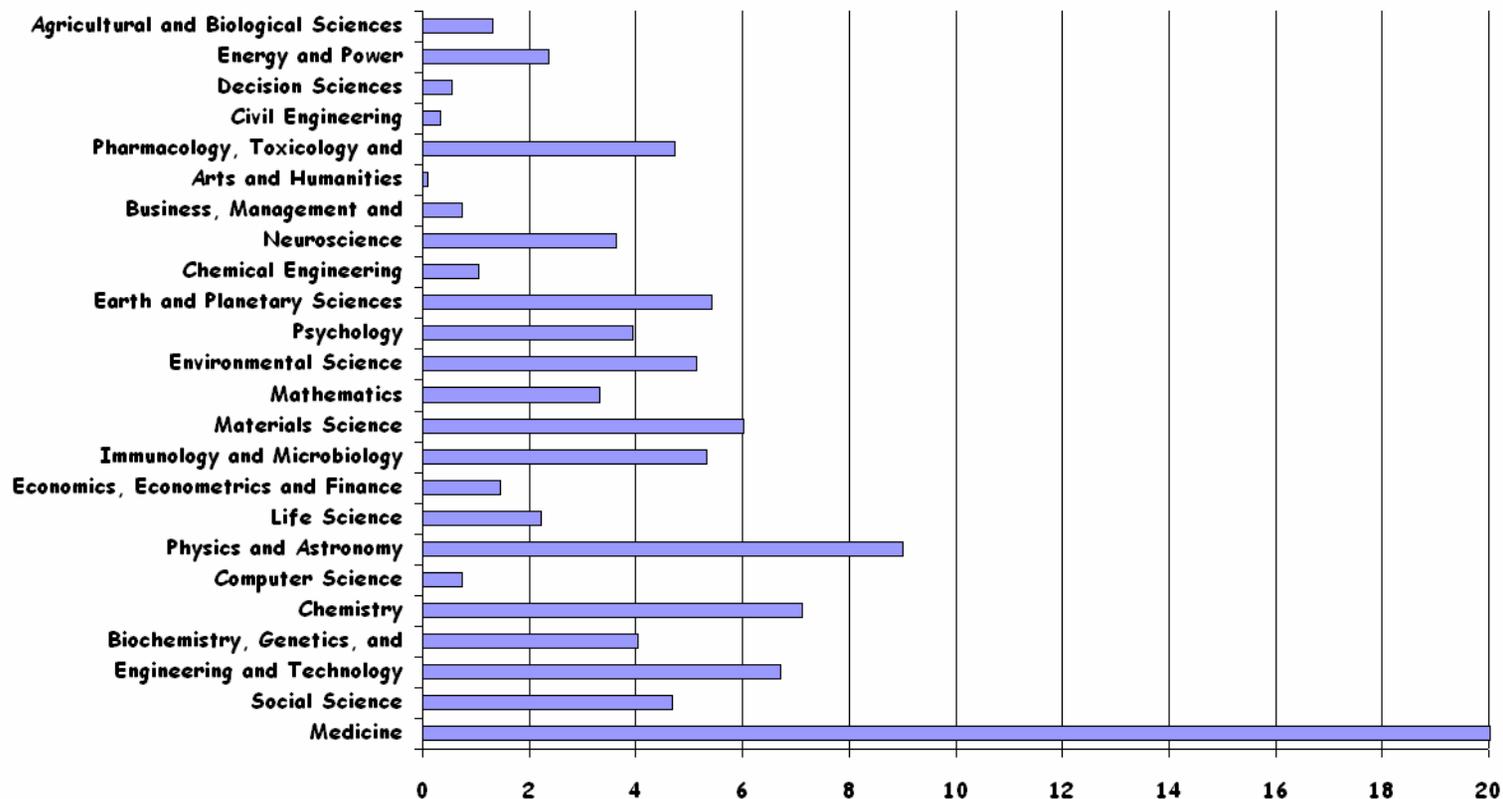
Distribuzione percentuale delle riviste disponibili per argomento di interesse



Periodo: Ottobre 2003- Luglio 2004

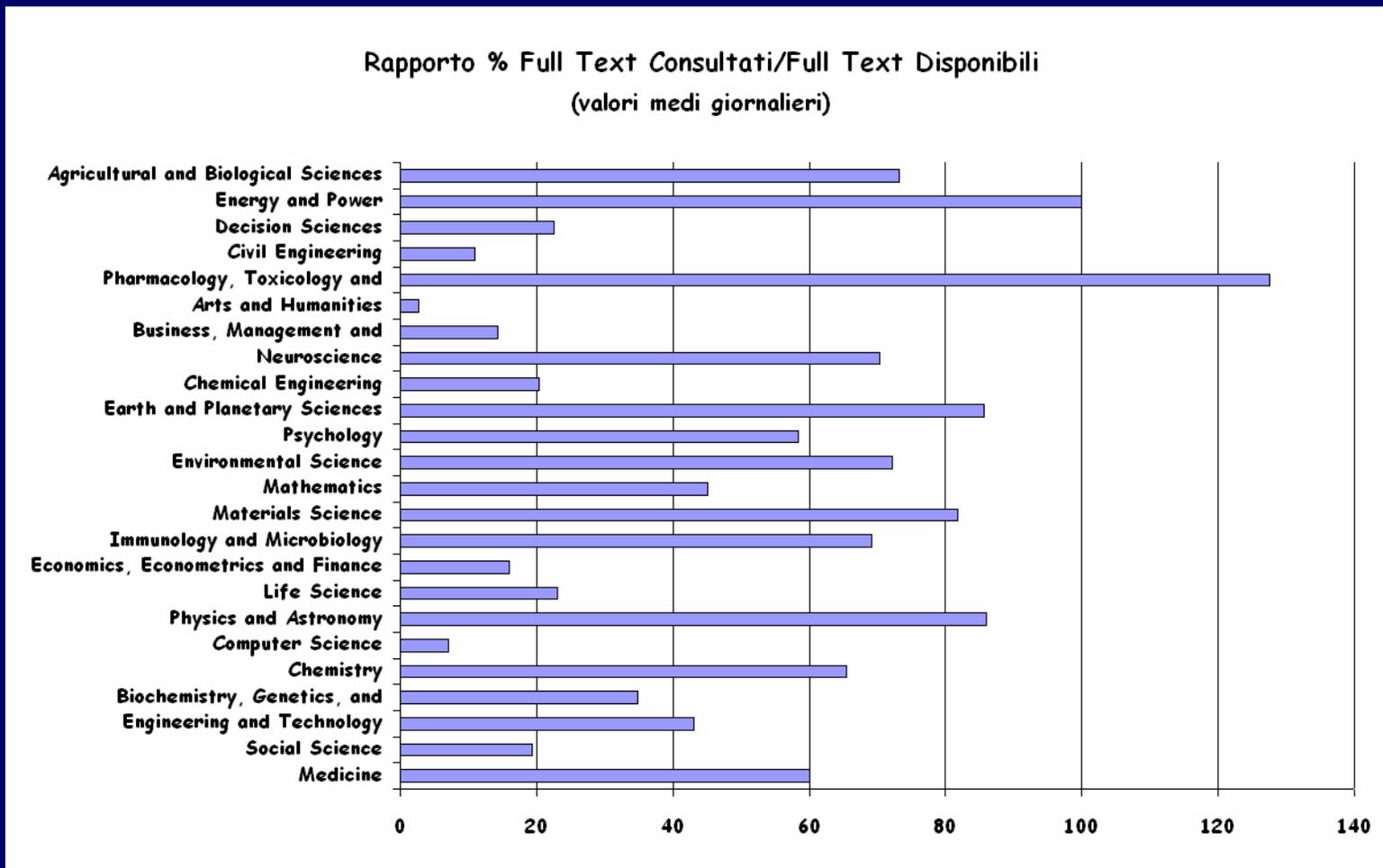
# ... che vengono così consultate...

Distribuzione Percentuale Full text consultati mediamente in 1 giorno



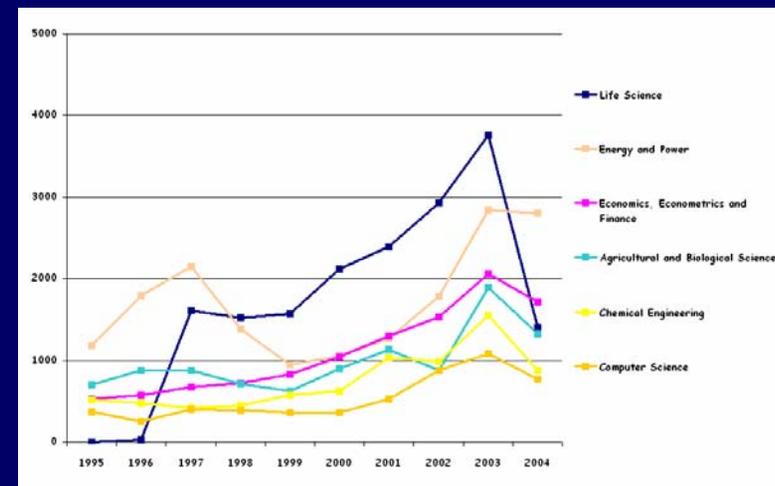
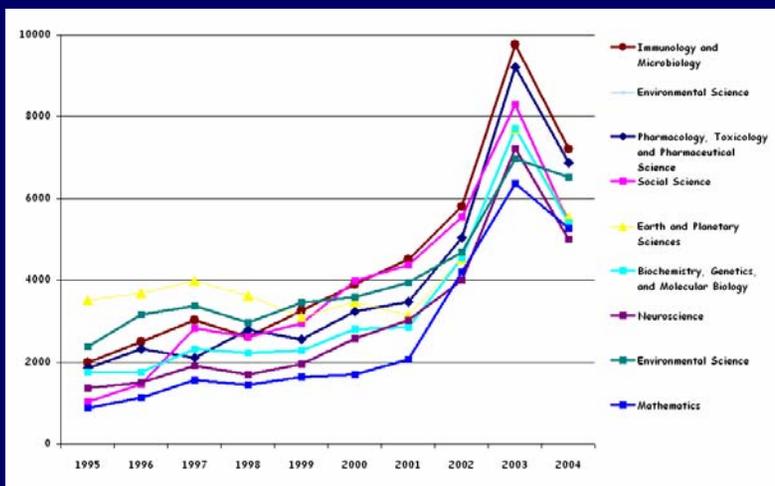
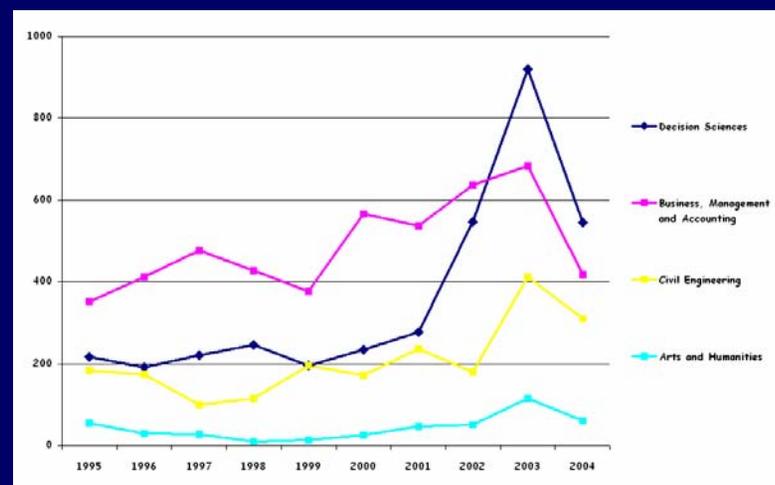
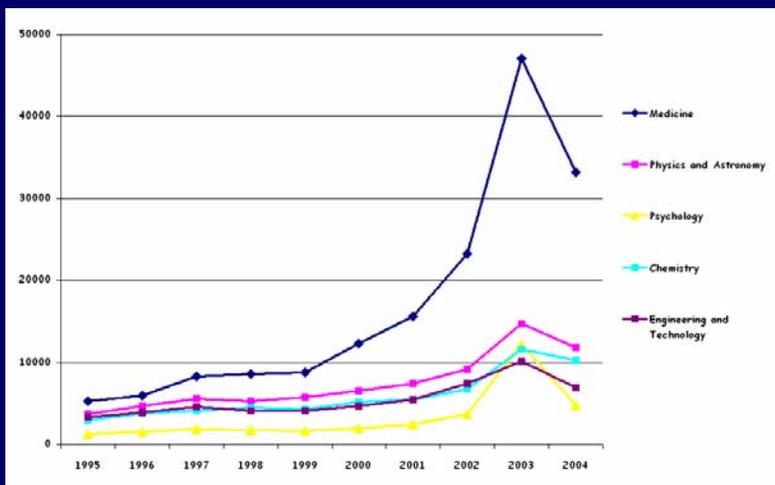
Periodo: Ottobre 2003- Luglio 2004

# ... consultazioni che potrebbero essere così sintetizzate:



Periodo: Ottobre 2003- Luglio 2004

# Inoltre come si comporta l'utente per categoria e anno di pubblicazione della rivista?



# Conclusioni

- Necessità di una *metodologia standard* di analisi dei dati:
  - Pre-trattamento
  - Sintesi e normalizzazione dei risultati
- Necessità di definire la *sessione di ricerca*