

NOZIONI DI INFORMATION RETRIEVAL

a cura di Riccardo Ridi

Ottobre 2011

- **RILEVANZA** = oggettiva
- **PERTINENZA (INTERESSE)** = soggettiva

- **RICHIAMO** = recuperare TUTTI i doc. rilevanti
- **PRECISIONE** = recuperare SOLO i doc. rilevanti

- **RUMORE** = doc. recuperati ma non rilevanti
- **SILENZIO** = doc. rilevanti ma non recuperati

- **PUNTO DI FUTILITA'** ---> ranking

RAPPORTO INVERSO FRA RICHIAMO E PRECISIONE

aumentando il richiamo
diminuisce la precisione
(e l'inverso)

$$\text{richiamo} = \frac{\text{documenti rilevanti recuperati nella ricerca}}{\text{documenti rilevanti esistenti nella banca dati}}$$
$$\text{precisione} = \frac{\text{documenti rilevanti recuperati nella ricerca}}{\text{documenti (rilevanti o no) recuperati nella ricerca}}$$

o no = rumore

"TRASPARENTE":

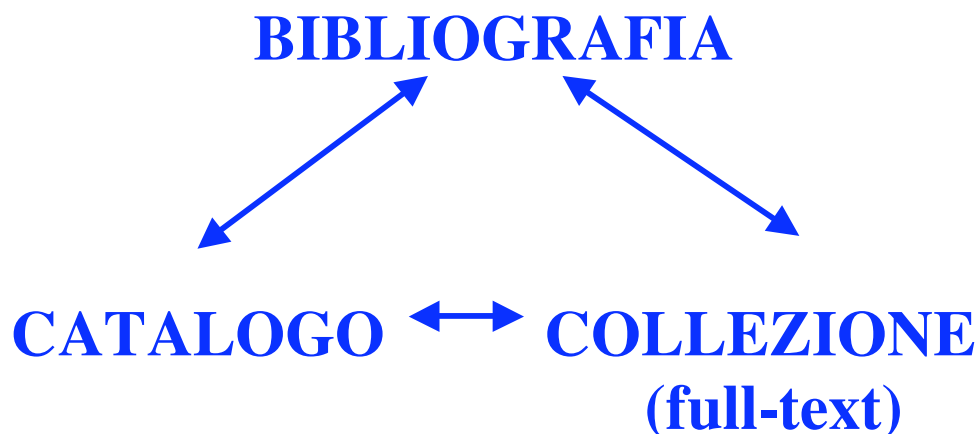
- **MECCANISMI INVISIBILI**
BLACK BOX, AUTOMAGIC, GUI (graphical user interface)
- **MECCANISMI VISIBILI**
ACQUARIO, COMPrensione & RIPETIBILITA', CUI
(character user interface)

RICERCA DI INFORMAZIONI:

- **NEI METADATI**
(IN CAMPI NORMALIZZATI)
- **NEI DATI PRIMARI**
 - **TESTUALI (RICERCA FULLTEXT)**
 - **NON TESTUALI (MULTIMEDIA INF. RETR.)**

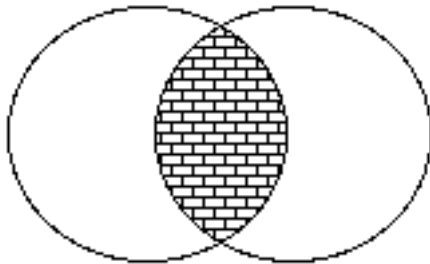
RICERCA BIBLIOGRAFICA:

- **NEI METADATI**
- **NEI DATI PRIMARI**
- **NELLE CITAZIONI**
 - **VERSO IL PASSATO (bibliografie di libri e articoli)**
 - **VERSO IL FUTURO (citation indexes)**



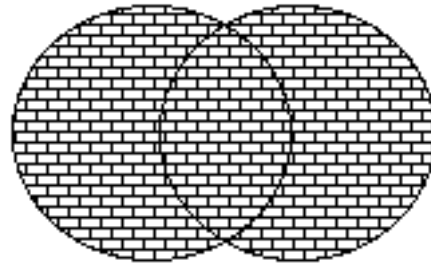
OPERATORI BOOLEANI

AND



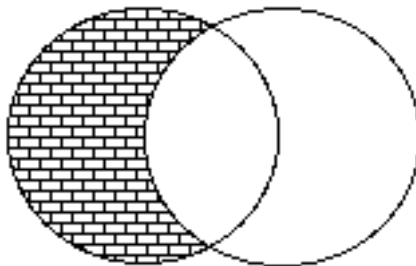
PROMESSI AND SPOSI

OR



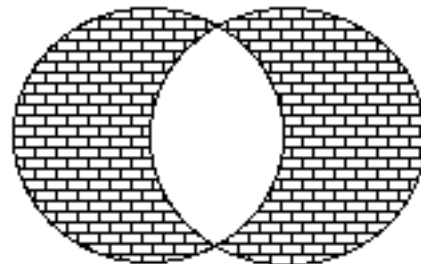
PROMESSI OR SPOSI

NOT



PROMESSI NOT SPOSI

XOR



PROMESSI XOR SPOSI

OPERATORI RELAZIONALI

>

<

||

FILTRI

**(per anno, per lingua,
per tipo di documento...)**

OPERATORI DI PROSSIMITÀ

(AND potenziato)

SAME

NEAR

ADJ /WITH

Gli **operatori di prossimità** costituiscono una sorta di **AND potenziato**, e rintracciano i termini cercati solo se presenti nello stesso campo o sottocampo:

- in qualsiasi ordine e a qualsiasi distanza fra loro (spesso **SAME**);
- solo se posti uno accanto all'altro o a una determinata distanza, specificabile, fra loro, in qualsiasi ordine (spesso **NEAR**);
- solo se posti uno accanto all'altro o a una determinata distanza, specificabile, fra loro, nell'ordine dato (spesso **ADJ** o **WITH**).

TRONCAMENTI

biblio?

?teca

MASCHERAMENTI

*to*o*

*to**o*

FRASE ESATTA

"annali della scuola normale superiore"

PARENTESI

promessi AND sposi OR manzoni

(promessi AND sposi) OR manzoni

promessi AND (sposi OR manzoni)

$$7 \times 2 + 5$$

$$(7 \times 2) + 5 = 19$$

$$7 \times (2 + 5) = 49$$

NOT, AND, OR, XOR

METODI DI RICERCA NEI SISTEMI INFORMATIVI DIGITALI

1: SCAN-----> look simile in TELNET e in WWW

= scorrimento di **liste** (ciascuna relativa ad un solo campo)

= ricerca per **liste** ----> elenchi

2: QUERY -----> look diverso in TELNET e in WWW

= estrazione, interrogazione, search, find

= ricerca incrociata in più **campi** ----> testo libero

----> mascherine

in TELNET ----> **TESTO LIBERO**

in WWW -----> **MASCHERINE**

O TESTO LIBERO

3: BROWSE -----> links ipertestuali da un record verso....

..... **SCAN** (meno diffuso)

..... **QUERY** (più diffuso)

..... **BROWSE** (singoli links
"creativi" fatti "a mano")

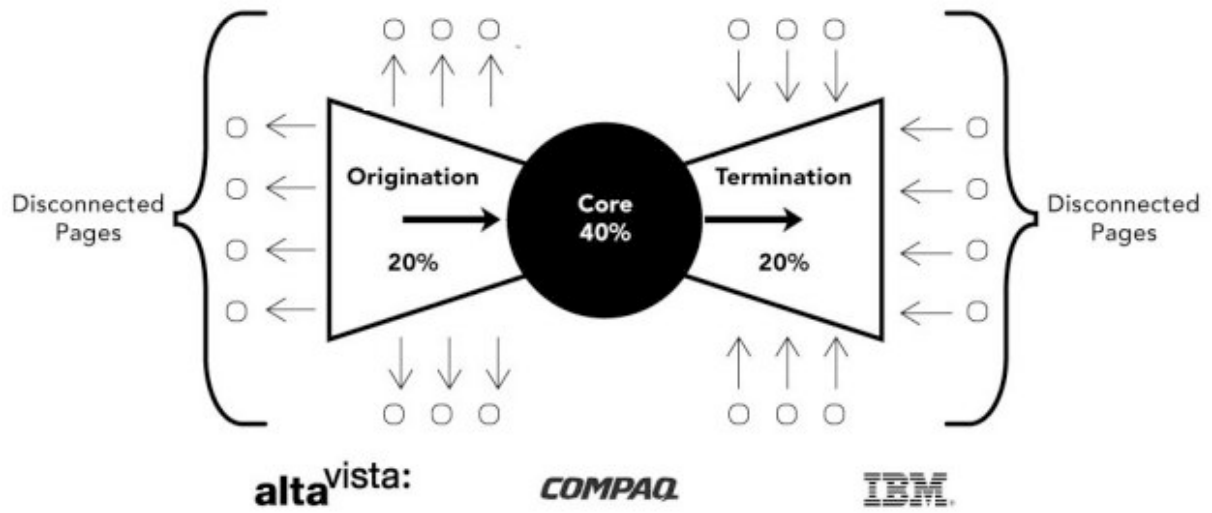
= navigazione ipertestuale

INFORMATION RETRIEVAL & SEARCH ENGINES (& DIRECTORIES)

- **RICHIAMO => copertura effettiva**
 - intranet**
 - internet invisibile**
 - web invisibile (database, cms)**
 - pagine html troppo lunghe**
 - pagine html troppo recenti**
 - immagini, filmati, suoni**
 - teoria del papillon**
- **PRECISIONE => strumenti di raffinamento**
 - operatori booleani e virgolette**
 - ricerca per campi, filtri**
- **PUNTO DI FUTILITA' => relevance ranking**

Bow-tie Theory

AltaVista, Compaq, IBM Web Mapping Study



NUOVE FRONTIERE

Multimedia information retrieval (MMIR), ovvero la ricerca di documenti non testuali con tecniche basate prevalentemente sul loro contenuto multimediale, scavalcando l'intermediazione dei metadati testuali.

Logica fuzzy (sfocata), che permette di assegnare sia ai termini usati nella ricerca che agli operatori che li connettono valori percentuali diversi dai due casi estremi (100% e 0%) previsti dalla logica classica, producendo risultati meno netti ma, pare, più vicini alla realtà, che - come si suol dire - non è mai nè completamente bianca nè completamente nera.

Interfacce tridimensionali, sia del tipo “realtà artificiale” (non intrusive, sullo schermo del computer) che del tipo “realtà virtuale” (intrusive, con l'ausilio di caschi, occhiali, guanti e tute speciali per raggiungere il massimo del realismo).

Riconoscimento vocale dei comandi (input) e emissione vocale dei risultati (output).

Agenti intelligenti, sistemi esperti, collaborative filtering, folksonomies, knowbots (knowledge robots) e altri metodi basati sostanzialmente sulla capacità del sistema di ricerca di imparare dalla ricerca stessa e dagli input più o meno volontari provenienti da fonti diverse per perfezionare progressivamente la propria efficacia. Alcune delle ricerche in questo settore sono collegate a quelle sull'intelligenza artificiale e a sistemi per offrire agli utenti un *reference service* automatizzato.

Multilinguismo. Approfondimenti delle tecniche di ricerca in ambienti multilinguistici, collegati anche alle ricerche sulla traduzione automatica. A cavallo fra questo ambito di studi e quelli sull'intelligenza artificiale si collocano le ricerche più spinte sull'uso del linguaggio naturale in ambito di information retrieval.

Ipertestualità. Approfondimenti delle tecniche di ricerca ipertestuale, che hanno trovato nel web un fondamentale risultato ma non un punto di arrivo definitivo.

Integrazione coi motori di ricerca. Esposizione dei contenuti degli opac e delle biblioteche digitali agli strumenti di ricerca sul web di tipo generale, come *Google*, o specializzati in ambito accademico-scientifico come *Scirus*, sottraendoli ai recessi del web “invisibile”. Inversamente, viene approfondita anche l'applicazione dei metodi di ricerca fulltext dei “web search engines”

esclusivamente su raccolte di documenti primari omogenei come quelli presenti nelle biblioteche digitali, producendo così risultati più contestualizzati e focalizzati.

Relevance ranking. Accanto ai tradizionali metodi di ordinamento utilizzati nei cataloghi e in altre forme di offerta di servizi (bollettini delle nuove acquisizioni, disseminazione selettiva dell'informazione, ecc.), la biblioteca digitale offre la possibilità (che dovrebbe comunque sempre restare solo opzionale) di ordinamenti (ranking) alternativi basati su criteri complessi, mutuati anch'essi dai "web search engines", che cercano di produrre risultati più vicini possibile all'ipotizzata massima rilevanza per l'utente.

Mappe semantiche, che visualizzano graficamente la "distanza concettuale" fra i documenti recuperati.

Filtraggi automatici, applicati ai documenti recuperati in base alle loro caratteristiche, ricavandone sottoinsiemi più omogenei.

Natural user interface (NUI), ovvero l'interazione "naturale" col computer, senza nè linguaggi nè metafore da imparare.