NOZIONI DI INFORMATION RETRIEVAL

a cura di Riccardo Ridi (Università Ca' Foscari di Venezia)
aggiornato a Luglio 2026

- **PERTINENZA = attinenza oggettiva/assoluta**

"assonanza concettuale tra argomento del testo reperito e argomento o materia della lista da costruire" Rino Pensato, *Corso di bibliografia*, Bibliografica, 4a ed. 1998, 1a rist. 2000, p. 80.

- **RILEVANZA = attinenza soggettiva/relativa (effettiva utilità personale)**

"concetto quantitativo e qualitativo insieme, poggia sul grado di innovazione e originalità e sulla quantità di nuove conoscenze, nuovi dati, nuovi modi di approccio forniti da un testo in rapporto all'argomento oggetto di studio o di repertorio" Rino Pensato, *ivi*.

Ma altri autori (fra cui A. C. Foskett, *Il soggetto*, Bibliografica, 2001, p. 36-38) utilizzano i due termini con significato inverso.

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

RAPPORTO INVERSO FRA RICHIAMO E PRECISIONE

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

$$\text{richiamo} = \frac{\text{documenti attinenti recuperati nella ricerca}}{\text{documenti attinenti esistenti nella banca dati}}$$

$$\text{precisione} = \frac{\text{documenti attinenti recuperati nella ricerca}}{\text{documenti (attinenti o no) recuperati nella ricerca}}$$

o no = rumore

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

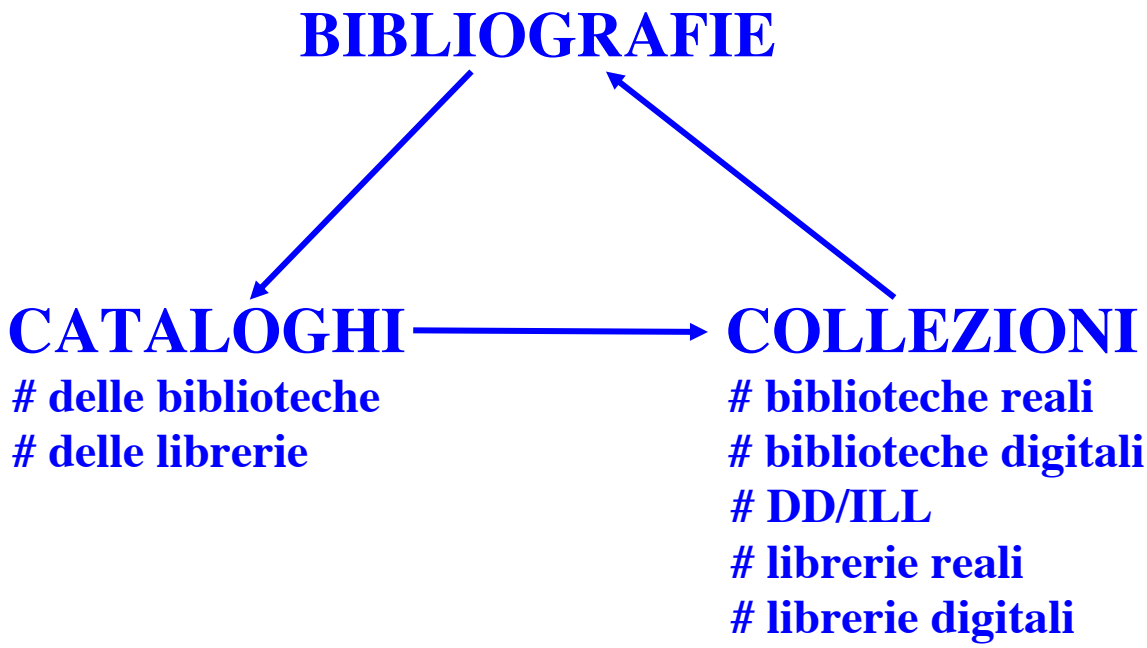
- **PUNTO DI FUTILITA' ---> ranking**

"TRASPARENTE":

- **MECCANISMI INVISIBILI**
  black box, automagic, discovery tools
  GUI (graphical user interface), intelligenza artificiale
- **MECCANISMI VISIBILI**
  acquario, comprensione & ripetibilità,
  CUI (character user interface), OPAC

RICERCA DI INFORMAZIONI:

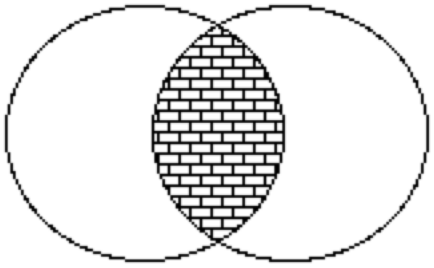
- **NEI METADATI**
  IN CAMPI NORMALIZZATI
- **NEI DATI PRIMARI**
  - **TESTUALI (RICERCA FULLTEXT)**
  - **NON TESTUALI (MULTIMEDIA INFORM. RETRIEVAL)**
- **ATTRAVERSO I RIFERIMENTI BIBLIOGRAFICI (vale solo per le informazioni bibliografiche)**
  - **VERSO IL PASSATO (bibliografie di libri e articoli)**
  - **VERSO IL FUTURO (citation indexes)**



OPERATORI BOOLEANI

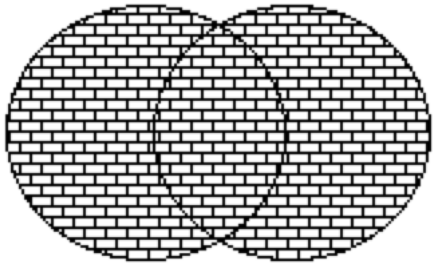
GEORGE BOOLE (1815-1864)

AND



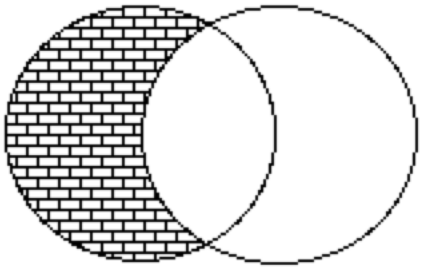
PROMESSI AND SPOSI

OR



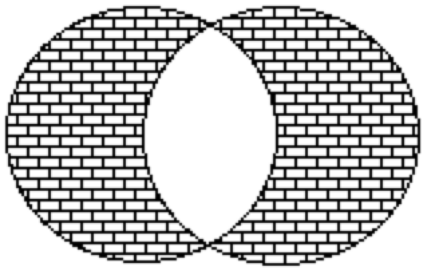
PROMESSI OR SPOSI

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*)**

$$7 \times 2 + 5$$

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

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

— , X & : , + & —

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**

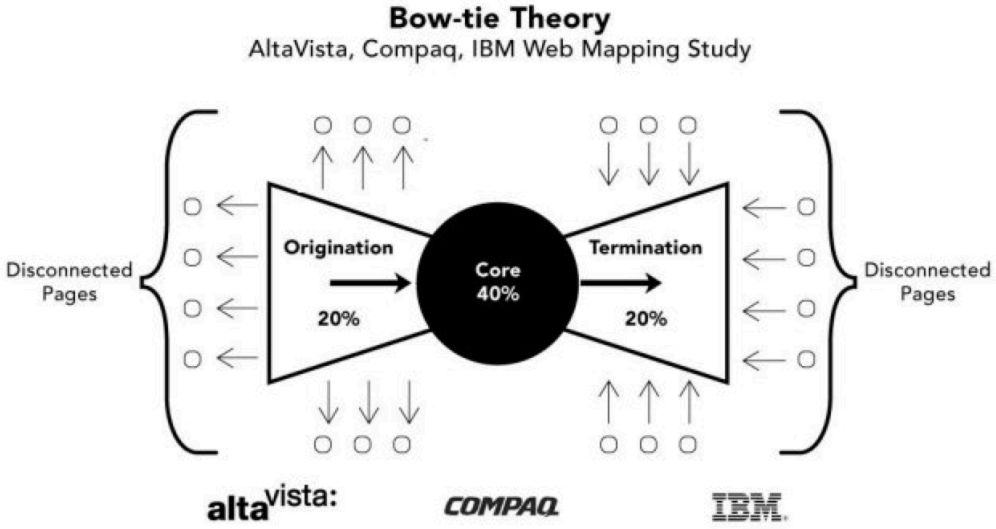
3: BROWSE -----> links ipertestuali da un record verso....

= navigazione ipertestuale

..... **SCAN** (meno diffuso)
..... **QUERY** (più diffuso)
..... **BROWSE** (singoli links "creativi" fatti "a mano")

- **RICHIAMO => copertura effettiva**
  - **intranet**
  - **internet invisibile**
  - **web invisibile (database, cms)**
  - **pagine html troppo lunghe**
  - **pagine html troppo recenti**
  - **immagini, filmati, suoni**
  - **teoria del papillon**

BUSINESS WIRE COMMERCIAL PHOTO



- **PRECISIONE => strumenti di raffinamento**
  - **operatori booleani e virgolette**
  - **ricerca per campi, filtri**

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

NUOVE FRONTIERE DELL'INFORMATION RETRIEVAL

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.

Multilinguismo. Approfondimenti delle tecniche di ricerca in ambienti multilinguistici, collegati anche alle ricerche sulla traduzione automatica.

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

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 spesso più vicini alla realtà.

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).

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

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

Open data (ovvero interoperabilità) e **linked data** (ovvero ipertestualità) come premesse del **web semantico**.

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 automatici 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.

Multiricerca. La possibilità, offerta dai siti di molte biblioteche universitarie e di ricerca ai propri utenti registrati, di interrogare contemporaneamente sia l'opac locale che tutte le altre fonti informative digitali ivi disponibili, indipendentemente dalla rispettiva tipologia, attraverso strumenti dalle denominazioni varie e mutevoli (opac allargati, portali bibliotecari, **discovery tools***, metaricerche, ricerche federate, ecc.) che consentono talvolta anche varie forme di personalizzazione e di memorizzazione delle ricerche, come ad esempio la possibilità di ripetere automaticamente una certa interrogazione a cadenza regolare, ricevendone i risultati nella propria casella di posta elettronica.

*Il termine "discovery tool" (strumento per la scoperta) andrebbe più appropriatamente utilizzato per indicare solo quei software che, anzichè effettuare una ricerca bibliografica su una pluralità di fonti informative separate e indipendenti, interrogano una sola "base di conoscenza", precedentemente predisposta cumulando i contenuti delle varie fonti disponibili.

Reference linking. L'adozione sempre più diffusa, da parte delle stesse biblioteche, di software per il "reference linking" (detti anche "link resolver") che collegano automaticamente fra loro i documenti digitali e le relative descrizioni bibliografiche e catalografiche, facilitando gli utenti registrati nel passaggio da queste ultime ai relativi testi integrali (se disponibili) fino al punto di rischiare che essi non si rendano nemmeno più conto della distinzione - nonostante tutto ancora significativa - fra bibliografie, cataloghi e collezioni.

Filtraggi automatici, applicati ai documenti recuperati in base alle loro caratteristiche, ricavandone sottoinsiemi (**clusters, faccette**) più omogenei.

Suggerimenti (sia in fase di immissione della query che in fase di valutazione dei risultati) forniti dal sistema informativo, basandosi sulle precedenti scelte degli utenti.

Metadati aggiunti dagli utenti stessi (**social tagging**, **folksonomie**).

Apps per smartphones e tablets.

Intelligenza artificiale, utilizzata per la ricerca bibliografica soprattutto in 3 modi:

1) chiedendo informazioni bibliografiche a un software IA (come ChatGPT);
2) chiedendo informazioni bibliografiche a un motore di ricerca generalista (come Google) che fra le sue opzioni ha anche la “modalità IA”;
3) chiedendo informazioni bibliografiche a uno strumento di ricerca bibliografica (come un opac o un discovery tool) potenziato con IA.

Soprattutto nei casi 1) e 2) bisogna fare molta attenzione alle imprecisioni (o addirittura alle “allucinazioni”) che la IA tende a generare (perché in realtà non “cerca” gli item bibliografici, ma li “crea”), possibilmente “testandola” prima su un argomento che già si conosce.