


Information Retrieval Systems

Oreste Signore
(Oreste.Signore@cnuce.cnr.it)



Contenuto

- ❖ **Elaborazione automatica dell' informazione**
- ❖ **Aspetti tecnici e semantici**
- ❖ **I sistemi di Information Retrieval**
 - ◆ Caratteristiche generali e schema funzionale
 - ◆ Gli indici invertiti
 - ◆ Operatori di adiacenza e di contesto
 - ◆ L' indicizzazione
 - ◆ Richiamo e Precisione
- ❖ **I thesauri**
- ❖ **OPAC, MultiOPAC, MetaOPAC**
- ❖ **Gli standard**
- ❖ **Z39.50: questo sconosciuto**
- ❖ **Conclusioni**

© 2001 - Oreste Signore

Information Retrieval

2



Elaborazione automatica dell'informazione

❖ L'epoca dell'informazione

❖ Tipi di informazione:

- ◆ testi
- ◆ voce
- ◆ grafici e immagini

❖ Aspetti da considerare:

- ◆ **tecnico**
(rappresentazione, manipolazione, memorizzazione, trasferimento, fruizione)
- ◆ **semantico**
(efficacia di trasmissione del messaggio)



Aspetti tecnici

❖ Tecnicamente, l'informazione è un insieme di *elementi discreti* (caratteri, parole, pixel, etc.)

❖ Le dimensioni:

- ◆ il linguaggio naturale è unidimensionale
- ◆ la voce è bidimensionale (onde)
- ◆ le immagini sono bidimensionali (pixel).
Alcune linee possono essere rappresentate sotto forma di equazione (grafica vettoriale)

❖ Elaborazione (per voce e immagini)

- ◆ sintesi *facile*
- ◆ riconoscimento *difficile*



Aspetti tecnici (cont.)

❖ *Un' immagine vale più di mille parole*

❖ **Ingombri fisici**

Tipo di informazione	Spazio occupato
Testo	500 parole/pag • 6 bytes/parola = 3000 bytes/pag
Voce	240 s/pagina • 9600 bit/s = 2.3 Mbit/pag (circa 290Kb/pag)
Immagine	Scansione a 500 bpi (equivalenti a 25×10^4 bit/inch ²) Una pagina 16x25 equivale a 400 cm ² (circa 62 inch ²) $62 \text{ inch}^2/\text{pag} \cdot 25 \cdot 10^4 \text{ bit}/\text{inch}^2 = 1.55 \cdot 10^7 \text{ bit}/\text{pag}$ (circa 2Mb/pag)



Aspetti semantici

❖ **Testi**

- ◆ occorre **comprendere** il significato del testo
- ◆ base di conoscenza **comune** a estensore e utilizzatore
- ◆ risultati soddisfacenti in settori ben definiti

❖ **Voce e immagini**

- ◆ **individuare** il contesto
- ◆ **formalizzare** il contesto
- ◆ inflessioni e carattere
- ◆ **sensibilizzazione ad aspetti particolari**
(latrato, pianto, scarabocchio)



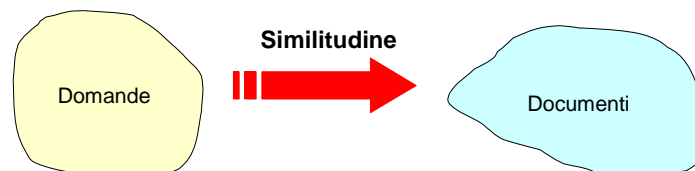
IRS: caratteristiche generali

- ❖ **I**nformation **R**etrieval **S**ystems
- ❖ Ricerca su testo libero (*similitudine*)
(sull'informazione strutturata la ricerca è per *corrispondenza diretta*)
- ❖ Contenuto identificato da:
parole chiave o *termini indice* o *descrittori*
- ❖ Operatori di ricerca su frase o documento
- ❖ Procedimento per raffinamenti successivi
- ❖ Informazione poco strutturata



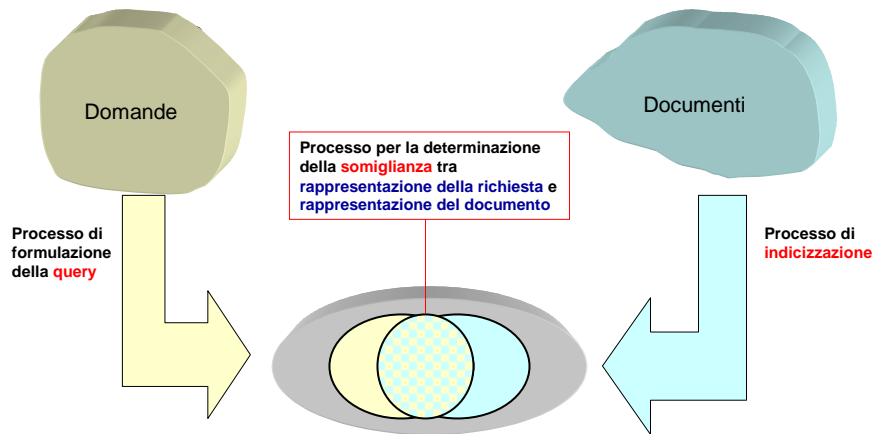
IRS: descrizione funzionale

- ❖ Analisi della *corrispondenza* tra l'insieme delle **domande** e l'insieme dei **documenti**, mediante un linguaggio intermedio di rappresentazione, detto *linguaggio di indicizzazione* o *linguaggio di classificazione*





IRS: descrizione funzionale



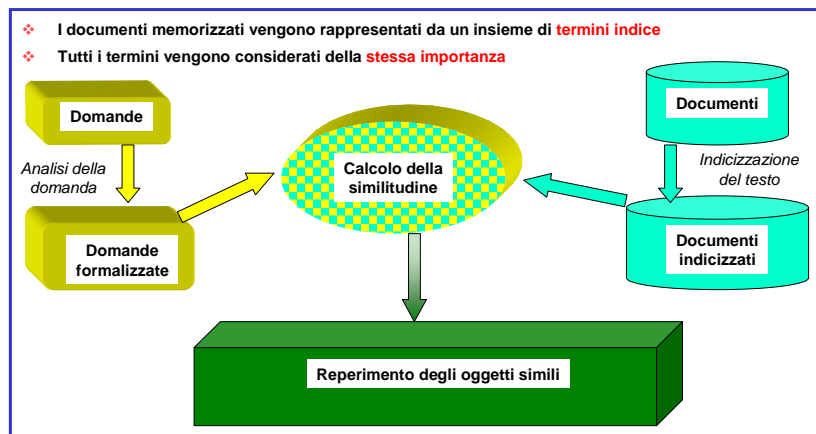
© 2001 - Oreste Signore

Information Retrieval

9



I vari passi



© 2001 - Oreste Signore

Information Retrieval

10



Gli operatori booleani

❖ I termini di ricerca vengono connessi mediante gli operatori booleani:

- ◆ AND
- ◆ OR
- ◆ NOT



Gli indici invertiti

❖ **Indice**= meccanismo veloce di accesso ai dati

❖ **Gli indici invertiti** sono il meccanismo adottato più frequentemente

- ◆ Archivio rappresentato come array di record indicizzati
- ◆ Trasposizione della matrice
- ◆ Manipolazione degli indici per determinare la risposta alla query

	T ₁	T ₂	T ₃	T ₄	T ₅
D ₁	1	1	0	1	0
D ₂	1	0	1	0	1
D ₃	0	0	1	1	1
D ₄	0	0	1	1	0
D ₅	1	1	0	1	0
D ₆	0	1	1	1	0

	D ₁	D ₂	D ₃	D ₄	D ₅	D ₆
T ₁	1	1	0	0	1	0
T ₂	1	0	0	0	1	1
T ₃	0	1	1	1	0	1
T ₄	1	0	1	1	1	1
T ₅	0	1	1	0	0	0



List merging (OR)

```
Inizializza le variabili
Prendi Ri da Lista1
Prendi Rj da Lista2
DOWHILE nessuna delle liste è finita
  IF (i<j) THEN
    trasferisci Ri in Listaoutput
    leggi il prossimo elemento da Lista1
  ELSE
    IF j non corrisponde a documento già trasferito
      trasferisci Rj in Listaoutput
    ENDIF
    leggi il prossimo elemento da Lista2
  ENDIF
  memorizza identificatore documento trasferito
ENDDOWHILE
IF (una delle liste in input non è finita) THEN
  trasferisci tutti i suoi dati in Listaoutput
ENDIF
Termina il programma
```



Operatori di contesto

❖ Query: **repubblica** AND **presidenziale**

L' onorevole XY ha parlato ieri in un convegno, illustrando la posizione del suo partito per la costituzione di una **repubblica presidenziale**. Il pubblico ha espresso le sue opinioni in merito in un lungo e interessante dibattito

~~Il presidente della **repubblica** ha ricevuto ieri l' ambasciatore di Z. Dopo l' incontro, i due illustri personaggi si sono recati in visita alla tenuta **presidenziale** di S. Rossore, dove ha avuto luogo un pic-nic.~~

❖ Occorrono quindi particolari operatori

(implementati con *informazioni aggiuntive* nell' indice)

- ◆ **ADJ**(n) termini *adiacenti* (nell'ordine specificato)
- ◆ **NEAR**(n) termini *vicini* (in qualunque ordine)
- ◆ **WITHIN**(s|p|d) per identificare la *porzione di documento* (sentence, paragraph, document)



Termini pesati

❖ **In generale:** $R_i = \{T_j \cdot \partial_{ij}\}$
dove
 $\partial_{ij} = 1 \Leftrightarrow T_j \in D_i$
 $\partial_{ij} = 0 \Leftrightarrow T_j \notin D_i$

❖ **Ma se i termini non sono tutti della stessa importanza:**

$R_i = \{T_j \cdot p_{ij}\}$
dove
 $0 < p_{ij} \leq 1 \Leftrightarrow T_j \in D_i$
 $p_{ij} = 0 \Leftrightarrow T_j \notin D_i$



Termini pesati (cont.)

- ❖ Il valore di p_{ij} dipende da quanto il termine è rilevante per identificare il contenuto del documento. Es. $R_i = \{T_1 \cdot 0.8; T_2 \cdot 0.2; T_3 \cdot 0.5\}$
- ❖ Possibile **ordinare i documenti** in base alla loro pertinenza alla domanda
- ❖ L'assegnazione dei pesi deve essere **semplice**
 - ◆ in generazione dell'archivio
 - ◆ in formulazione della domanda



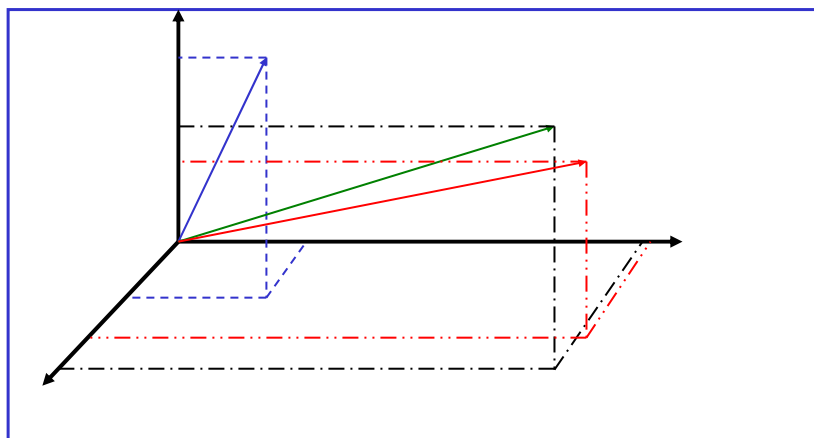
Spazio vettoriale dei documenti

❖ Ogni documento è rappresentato da un *vettore di termini*

	T_1	T_2	T_3	T_4	...	T_m
D_1	p_{11}	p_{12}	p_{13}	p_{14}	...	p_{1m}
D_2	p_{21}	p_{22}	p_{23}	p_{24}	...	p_{2m}
D_3	p_{31}	p_{32}	p_{33}	p_{34}	...	p_{3m}
...
D_n	p_{n1}	p_{n2}	p_{n3}	p_{n4}	...	p_{nm}



Termini pesati (cont.)





Sinonimi

- ❖ A volte il termine ricercato viene identificato con un termine che ha uno o più sinonimi
- ❖ Molti sistemi commerciali permettono di espandere automaticamente la query
- ❖ Se
 - ◆ il termine T_1 ha come sinonimi S_{11} e S_{12}
 - ◆ il termine T_2 ha come sinonimi S_{21} e S_{22}
 - ◆ Allora $(T_1 \text{ AND } T_3) \text{ OR } T_2$
diventa: $((T_1 \text{ OR } S_{11} \text{ OR } S_{12}) \text{ AND } T_3) \text{ OR } (T_2 \text{ OR } S_{21} \text{ OR } S_{22})$



Troncamento

- ❖ **Suffix truncation**
(di semplice implementazione con gli indici invertiti)
- ❖ **Prefix truncation**
(richiede la creazione di indici sui termini scritti in senso inverso)
- ❖ **Infix truncation**
(richiede la creazione di indici su tutte le "rotazioni" dei termini)
- ❖ **Stem**
(richiede l' utilizzo delle regole grammaticali)



Indicizzazione

- ❖ Il problema fondamentale: **identificare i contenuti dei documenti**
- ❖ Le realizzazioni sperimentali e commerciali si basano sul concetto di **rilevanza** del documento, cioè della misura di quanto un documento è pertinente alla domanda
- ❖ La caratterizzazione del documento consiste nell' assegnazione a ciascun documento di un insieme di termini, detti **parole chiave** o **parole indice**
- ❖ Noto come **approccio indiretto** in opposizione all' **approccio diretto** (lettura e comprensione del testo nella forma originaria)



Tipi di indicizzazione

- ❖ **Modalità**
 - ◆ **manuale**
(assegnazione di parola chiave da parte di esperti)
 - ◆ **automatica**
(assegnazione automatica dal computer)
- ❖ **Caratteristiche**
 - ◆ **esaustiva**
(considerati utili nell' indicizzazione anche argomenti marginali)
 - ◆ **specifica**
(al documento vengono assegnati termini molto precisi)



Indicizzazione automatica esaustiva

❖ Estrazione automatica dal testo di tutte le parole non comprese nella lista di “*parole da ignorare*”, o “*parole vuote*” o “*stopword*”

❖ Lista spesso:

- ◆ definita una volta per tutte
- ◆ include normalmente articoli, preposizioni, parti del discorso poco significative, ...
- ◆ non specifica per le varie sezioni del documento
- ◆ produce alcuni effetti indesiderabili

Gli **agli** coltivati nelle campagne meridionali hanno un sapore e un profumo nettamente più accentuati di quelli coltivati nelle regioni più fredde. Tale caratteristica è probabilmente da attribuire **agli** effetti delle diverse condizioni climatiche che caratterizzano ...

La convenzione generalmente accettata per accordare gli strumenti musicali è quella di far riferimento al **La** di una certa ottava ...

Datazione: sec I a.C.
Oggetto: fibula
Descrizione: I segni incisi ...



Linguaggio e vocabolario

❖ Linguaggio di indicizzazione:

- ◆ **non controllato** (*testo libero originale*)
- ◆ **controllato** (*identificazione manuale dei termini significativi*)
Introduce meno errori, ma comporta costi aggiuntivi

❖ Tipo di vocabolario:

- ◆ a **parole singole** (*apparecchiatura, impianto, fune, medicale*)
Richiede un processo di **post-coordinazione** al momento della formulazione della query
- ◆ a **frasi** (*apparecchiatura medicale, impianto a fune, ...*)
Richiede un processo di **pre-coordinazione** in fase di indicizzazione

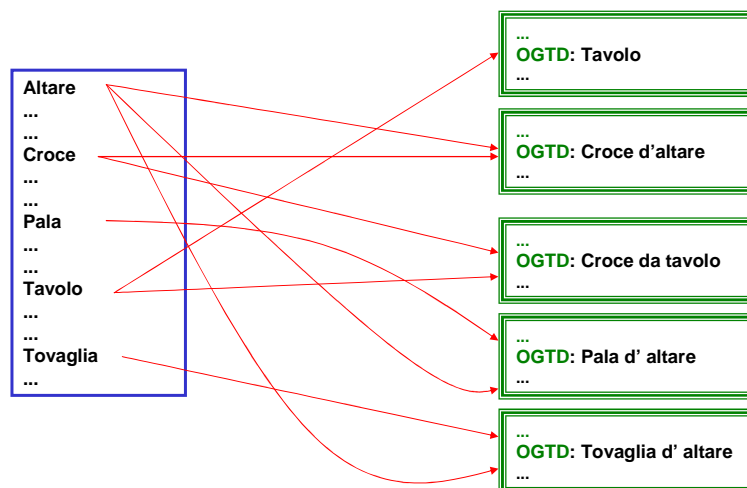


Tipi di indicizzazione

- ❖ Single term
- ❖ Terms in context
- ❖ A volte possibile la doppia indicizzazione

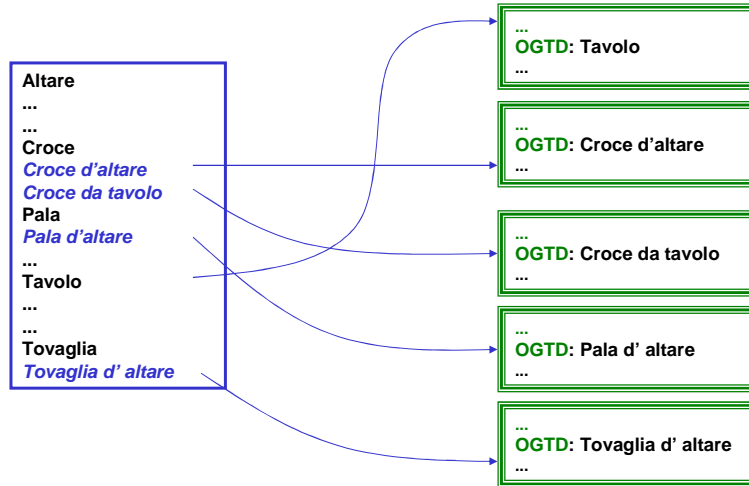


Indicizzazione single term

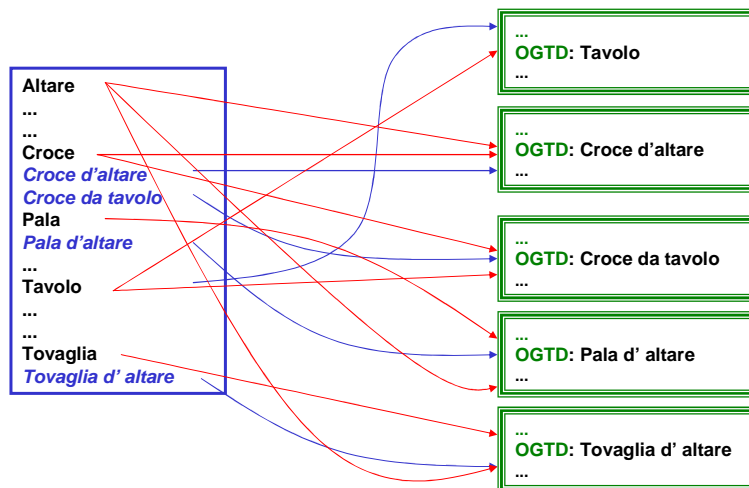




Indicizzazione per frase



Indicizzazione mista

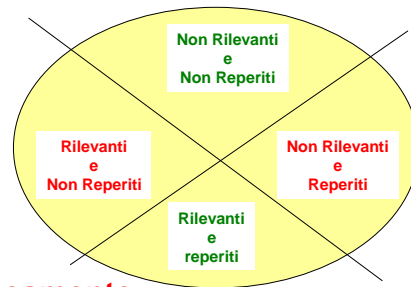




Richiamo e Precisione

$$R = \frac{\|Doc_{rilevanti} \cap Doc_{reperiti}\|}{\|Doc_{reperiti}\|}$$

$$P = \frac{\|Doc_{rilevanti} \cap Doc_{reperiti}\|}{\|Doc_{rilevanti}\|}$$



❖ Parametri da valutare

sempre contemporaneamente

❖ Indicizzazione esaustiva + linguaggio specifico = alto RICHIAMO e alta PRECISIONE



Richiamo e Precisione(cont.)

❖ Sulla stessa banca dati può essere utile avere, di volta in volta:

- ◆ alto R per argomenti trattati in modo marginale
- ◆ alto P per argomenti trattati in modo estensivo

❖ In letteratura riportati casi di studio in cui, per mancanza di un vocabolario normalizzato:

- ◆ P circa 0.8
- ◆ R circa 0.2

❖ Alcune sperimentazioni in (SMART):

- ◆ linguaggio non controllato e vocabolario a parole singole ha fornito risultati migliori rispetto a linguaggio controllato e vocabolario a frasi



Valore discriminante

❖ Legge di Zipf (1949)

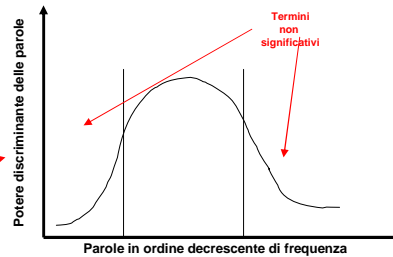
- ◆ si dispongono le parole in ordine decrescente di frequenza

- ◆ $Freq_{term} \cdot NumOrd_{term} \cong \text{costante}$

❖ Potere discriminante delle parole

❖ Trasformazioni:

- ◆ **di frase**
permette di scartare i termini con alta frequenza, incorporandoli in frasi di frequenza più bassa
- ◆ **di thesaurus**
i termini troppo specifici (di bassa frequenza) possono essere inclusi in classi di thesaurus



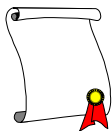
Indicizzazione automatica: un algoritmo

- Calcolare $FREQ_{ik} = \text{freq del termine } T_k \text{ in } D_i$
- Calcolare $TOTFREQ_k = \sum_{i=1, n} FREQ_{ik}$
- Ordinare le parole in ordine di frequenza decrescente
- Eliminare le più frequenti
Eliminare le meno frequenti
- Usare i termini di frequenza intermedia



Indicizzazione automatica: i problemi

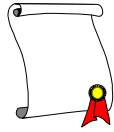
- ❖ **Identificare le soglie di frequenza minima e massima**
 - ◆ eliminare i termini molto frequenti abbassa il Richiamo
 - ◆ eliminare i termini poco frequenti abbassa la Precisione
- ❖ **Un buon termine indice:**
 - ◆ deve rendere reperibile il documento (**Richiamo**)
 - ◆ deve essere in grado di distinguere il documento all' interno dell' intera collezione (**Precisione**)
 - ◆ non può essere un termine presente in tutti i documenti
 - ◆ è molto frequente in alcuni documenti (ipotesi del minimo sforzo)
 - ◆ non è molto frequente nell' intera collezione di documenti



I thesauri

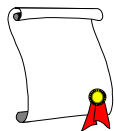
- ❖ **Problemi tipici dell' Information Retrieval**
 - ◆ scarsa precisione dei termini utilizzati
 - ◆ identificazione errata dell' elemento da ricercare
 - ◆ utilizzatore e indicizzatore **non** condividono la stessa **base di conoscenza**
- ❖ **Necessario un adeguato supporto all' utente, per esplicitare le *relazioni semantiche tra i concetti***
- ❖ **Un **thesaurus** è una lista di termini tra i quali sono definite delle **relazioni semantiche** di:**
 - ◆ gerarchia
 - ◆ preferenza
 - ◆ equivalenza

ISO 2788



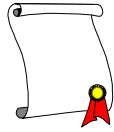
I thesauri (cont.)

- ❖ L' uso di un thesaurus comporta un considerevole **aumento di Precisione**, sia a livelli bassi che, soprattutto, a livelli alti di Richiamo
- ❖ I thesauri sono essenziali per poter **raffinare** il concetto da ricercare
- ❖ L' utente deve essere messo in grado di **identificare** chiaramente le **relazioni** intercorrenti tra i vari termini
- ❖ La costituzione di un thesaurus è un' impresa di notevole **impegno**:
 - ◆ *However, the construction of useful thesauruses is an art rather than a science and requires extensive knowledge of the particular subject area under consideration and of the record collections to be processed* (G. Salton)



I thesauri (cont.)

- ❖ Un thesaurus è un insieme di **termini**, e di **relazioni** tra di essi, che costituiscono il **lessico specialistico** da usare per descrivere il contenuto dei documenti pubblicati in un certo **ambito disciplinare**
- ❖ Le relazioni possono essere di:
 - ◆ **preferenza**
 - ◆ **gerarchia**
 - ◆ **affinità semantica**
- ❖ Gli elementi base di un thesaurus sono quindi:
 - ◆ Indexing term
 - ◆ Entry term
 - ◆ Relazioni
 - ◆ Guida all' uso dei termini



Thesauri: i relatori standard

❖ Standard internazionale per i nomi dei relatori:

- ◆ **AF** abbreviazione di
- ◆ **BT** termine più ampio
- ◆ **NT** termine più ristretto
- ◆ **RT** termine affine
- ◆ **AB** forma abbreviata
- ◆ **UFA** usato per (remini combinati)
- ◆ **UF** usato per
- ◆ **LT** termine principale
- ◆ **SN** note esplicative
- ◆ **HN** note storiche
- ◆ **LE** equivalente in lingua
- ◆ **EQ** rinvio a termine principale
- ◆ **EQA** rinvio a termini principali (combinati)



Thesauri: un esempio

THESAURUS ITALIANO DI SCIENZE DELLA TERRA

Edizione gennaio 1998

[Cos'è](#) [Contenuto](#) [Architettura](#)

[Chi siamo](#) [Bibliografia](#)

Centro di Studio per la Geodinamica Alpina e Quaternaria - SEAL DNUCE

http://info.cnuce.cnr.it/THSGEO/thsgeo_prima.html

BT=Broader term US=Use
RT=Related term
NT=Narrower term UP=Use for

Search term Italiano English
(you can use % as wild-card)

cerca reset



Thesauri: un esempio

THESAURUS ITALIANO DI SCIENZE DELLA TERRA

Edizione ipermediale 1998

THESAURUS-SCIENZE DELLA TERRA

Termini trovati 23

litogeochimica (Geochemico generale) lithogeochemistry			
Geochimica generale	RT	geochimica	geochemistry
	RT	fugacità	fugacity
		ipartizione di elementi	partitioning
		anomalia geochimica	geochemical anomalies
		migrazione di elementi	migration of elements
Mineralogia generale	RT	inclusione fluida	fluid inclusion
		acqua di costituzione	water of hydration
Petrologia generale	RT	acqua di cristallizzazione	water of crystallization
		geotermometria	geologic thermometry
Petrografia delle rocce sedimentarie	RT	roccia sedimentaria	sedimentary rocks
		composizione ricca in ferro	iron-rich composition
Petrografia e petrologia del cristallino	RT	metamorfismo	metamorphism
		metasomatismo	metasomatism
		deveficazione	devefication
Petrografia statistica	RT	roccia ignea	igneous rocks
		roccia metamorfica	metamorphic rocks
Processi morfogenetici	RT	alterazione meteoica	weathering
Finca dell'atmosfera	RT	paleotemperatura	paleotemperature
Inciamentologia	RT	giacimento minerario	mineral deposits
Analisi chimiche e chimico-fisiche	RT	proporzione mineraria	mineral proportions
Altri termini	RT	metodo analitico	chemical analysis
		composizione	composition



Per una ricerca efficace...

- ❖ Conoscere la **struttura** della banca dati
- ❖ Conoscere le **caratteristiche** del sistema
- ❖ Condividere la **base di conoscenza** con l'indicizzatore