

I SISTEMI DI RECUPERO DELL'INFORMAZIONE

- Sistemi specializzati nella gestione di documenti di testo e nel recupero in base al loro contenuto

Sempre più spesso i documenti nascono direttamente in forma elettronica dando vita a enormi banche dati che contengono oltre ad una sintetica descrizione dei documenti anche il testo in forma integrale.

- Grossa collezione di documenti (banca dati)
- Collezioni *full-text*
- Pagine Web e motori di ricerca
- Applicazioni multimediali:
immagini, filmati, musica,....

DIFFERENZE CON I SGBD

- Documenti con molto testo piuttosto che dati strutturati.
- Le richieste sono espressioni imprecise del bisogno informativo
- Le risposte sono riferimenti a documenti “che potrebbero contenere le risposte” piuttosto che direttamente le risposte

Domanda tipica a un SGBD

```
SELECT Nome, Ufficio  
FROM Impiegati  
WHERE AnnoAssunzione > 1970  
AND Stipendio > 3000
```

Nome	Ufficio
Mario Paletti	Amministrazione
Guido Carlesi	Amministrazione
Sandra Merlini	Vendite

Domanda tipica a un SRI

FIND architett*

AND (cad OR (progetto AND calcolatore))

“... l’impiego del calcolatore per lo sviluppo di progetti **architett**onici riguarda il campo di applicazioni dell’informatica conosciuto con il nome di **CAD**, ovvero progetto assistito da calcolatore...”

“... nell’affrontare il **progetto** dell’**architettura** di un **calcolatore** bisogna tener conto del settore di applicazione in cui verrà utilizzato ...”

Sintesi delle differenze

	SGBD	SRI
Tipologia dei dati	Strutturati	Testo
Richiesta	Completa precisa	e Incompleta e vaga
Criterio di scelta	Corrispondenza esatta	Corrispondenza parziale
Risultato	Dati richiesti	Documenti probabilmente rilevanti

Un sistema di recupero dell'informazione gestisce i documenti. Per agevolare l'accesso ai dati il sistema costruisce delle rappresentazioni sintetiche dei documenti (dati ausiliari).

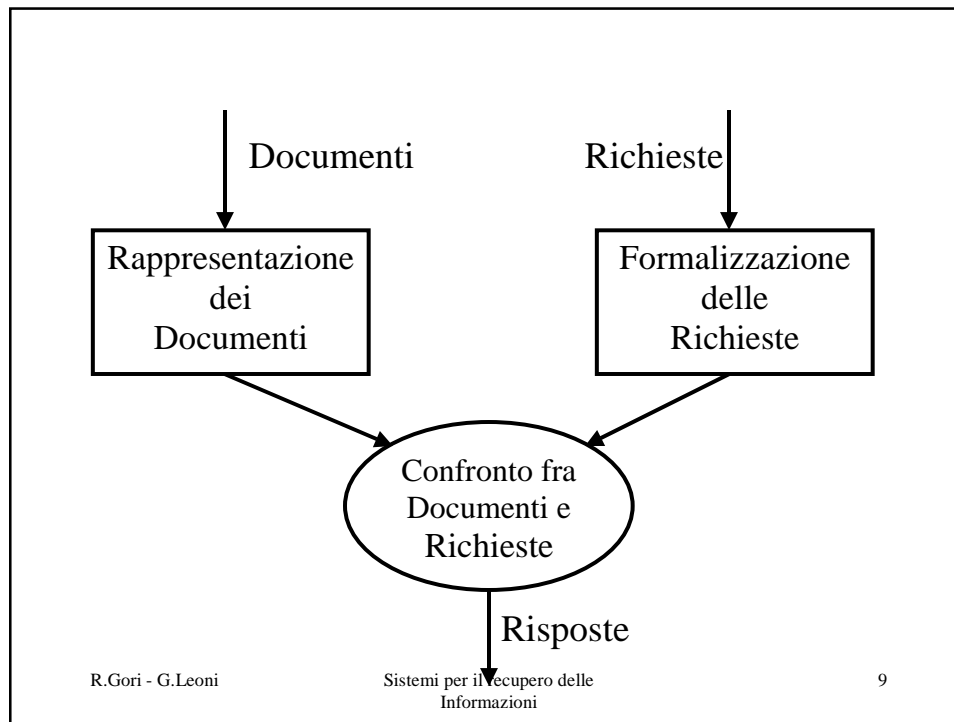
Il processo mediante il quale si associa ad un documento una sua rappresentazione sintetica viene detto *indicizzazione*.

L'idea è quella di associare a ciascun documento un insieme di termini significativi che saranno utilizzati per selezionare il documento.

Possiamo pensare ad un SRI come ad un sistema in cui da un lato entrano documenti che vengono sottoposti ad un processo di indicizzazione, per ottenerne una rappresentazione sintetica, dall'altro entrano le richieste dell'utente che devono essere codificate in modo analogo, cioè come un insieme di termini.

Risultato

- Binario (si/no) – il risultato soddisfa o non soddisfa la richiesta (corrispondenza esatta)
- Probabilistico – il risultato soddisfa la richiesta in una qualche misura (corrispondenza parziale)



Valutazione di un SRI

Richiamo: il numero di documenti rilevanti recuperati in rapporto ai documenti rilevanti presenti nella collezione

Precisione: il numero di documenti rilevanti recuperati in rapporto ai documenti recuperati

Modelli di SRI

Nel caso degli SRI un modello riguarda:

- lo stile di rappresentazione dei documenti;
- lo stile di rappresentazione delle richieste;
- la modalità del confronto tra rappresentazioni di documenti e richieste.

Il modello booleano

Rappresentazione dei documenti

- Un insieme di termini che ne rappresentano il contenuto (scelti durante l'indicizzazione)

Interrogazioni

- combinazioni booleane di termini, cioè termini combinati tra loro mediante AND, OR, NOT

Criterio di corrispondenza

- AND: i termini sono entrambi presenti
- OR: almeno uno dei due termini è presente
- NOT: il termine non è presente

Esempio

(film AND amore)

documenti che contengono “film” e “amore”

(dramma OR drammatico)

documenti che contengono “dramma” o “drammatico”

NOT (dramma OR drammatico)

... che **non** contengono “dramma” o “drammatico”

((film AND amore) AND NOT (dramma OR
drammatico))

Il modello booleano è a *corrispondenza esatta*

Esistono numerosi altri modelli:

- Modello vettoriale, a *corrispondenza parziale*
- Modello fuzzy, *probabilistico* ...
-

Indicizzazione

- **Indicizzazione:** processo di rappresentazione dei documenti mediante una descrizione sintetica (es: *catalogazione per soggetto* in ambito bibliotecario)
- Serve per costruire indici su collezioni di documenti

Un indice è costituito da:

- Una lista di termini
- Una lista di termini pesati

Linguaggio di indicizzazione

insieme dei termini scelti per indicizzare una collezione di documenti

Come sono scelte le parole del linguaggio di indicizzazione?

- **Linguaggio controllato:** limitato ad un vocabolario predefinito
- **Linguaggio libero:** termini estratti liberamente dal testo del documento e non definiti a priori

Processo di indicizzazione

- **Manuale:** è una persona che sceglie quali termini meglio caratterizzano il contenuto di un documento
 - Più “semantico” e quindi migliore
 - Soggettivo, costoso
 - Linguaggio controllato
- **Automatico:** fatto da un programma
 - Più sintattico, su base statistica e quindi “peggiore”
 - Economico, scalabile
 - Linguaggio libero

Qualità dell'indicizzazione

- **Esaustività:** il grado in cui tutti i concetti trattati sono presenti nella rappresentazione
- **Specificità:** il grado di specificità del linguaggio utilizzato
- **Indicizzazione profonda:** alto grado di esaustività e specificità, più costosa, precisione e richiamo più alti
- **Indicizzazione superficiale:** uso di alcuni termini generici, meno costosa, prestazioni inferiori

Indicizzazione manuale

In generale viene utilizzato un linguaggio controllato; questa scelta presenta diversi vantaggi:

- Semplificazione del processo di indicizzazione
- Indipendenza, o minor dipendenza, dal soggetto che effettua l'indicizzazione
- Semplificazione dell'uso da parte degli utenti (se conoscono il linguaggio di indicizzazione)

Struttura del linguaggio di indicizzazione

- **Dizionario:** termini ordinati alfabeticamente
Possono essere semplici o contestualizzati
- **Schema di classificazione:** codici che organizzano i termini gerarchicamente
es: DDC – Classificazione Decimale di Dewey
– usato nella classificazione bibliografica
- **Thesaurus:** termini organizzati in una “rete semantica”

Le relazioni previste in un thesaurus sono:

- *preferenza*: US (usa), UF (usato per)
- *gerarchia*: BT (generale), NT(specifico)
- *affinità*: COR (correlato), SIN(sinonimo)

e altre ancora

Indicizzazione automatica

Si basa sulla seguente ipotesi:

la frequenza di occorrenza di un termine in un documento è indicativo della sua importanza nel caratterizzarne il contenuto

“... Uno scrittore normalmente ripete certe parole nell’elaborare aspetti di un certo argomento. L’enfasi è considerata un indicatore di significatività ...”. [Luhn

Frequenza sì ma ...

- Le parole in assoluto più frequenti sono anche poco significative
 - avverbi, articoli, preposizioni ecc.
 - le 250 parole più comuni coprono in media il 40-50% di un testo
- Quello che conta non è la frequenza assoluta ma la frequenza relativa

Es. 'Computer' in una biblioteca di informatica

Processo di indicizzazione automatica - 1

Eliminazione di parole di uso comune (uso di *liste di esclusione* – STOP list)

- *Stralcio di una lista di esclusione per la lingua inglese:*

A	ALMOST	AMONGST	ANYWHERE
ABOUT	ALONE	AN	ARE
ACROSS	ALONG	AND	AROUND
AFTER	ALREADY	ANOTHER	AS
AFTERWORDS	ALSO	ANY	AT
AGAIN	ALTHOUGH	ANYHOW	BE
AGAINST	ALWAYS	ANYONE	BECAME
ALL	AMONG	ANYTHING	BECAUSE

Processo di indicizzazione automatica – 2

Riduzione delle parole alla radice

- Per aumentare il richiamo e ridurre le dimensioni del linguaggio di indicizzazione

- Si utilizzano liste di suffissi:

Es. calcol[are]
 calcol[atore]
 calcol[atrice]
 calcol[abilità]
 calcol[o]

Processo di indicizzazione automatica –3

Assegnazione di un peso a ciascun termine, basata sulla seguente ipotesi:

“l'importanza di un termine per caratterizzare il contenuto di un documento è tanto maggiore quanto più è alta la frequenza del termine nel documento e quanto più bassa è la frequenza del termine nei documenti della collezione”

- i termini con peso alto vengono assegnati al documento

Processo di indicizzazione automatica –4

Rappresentazione dei documenti, ad esempio come vettori di pesi.

Se ad es il linguaggio di indicizzazione è:

{Arbusto, Architettura, botanica, coltivazione, colonna, pianta, Rinascimento, Roma,}, il vettore

0	4	0	0	2	2	3	3	
---	---	---	---	---	---	---	---	--

rappresenta un documento in cui ‘arbusto’ ha peso 0, ‘architettura’ ha peso 4, ‘botanica’ ha peso 0,

Indicizzazione automatica o manuale?

- In media c'è un accordo del 60% con le due tecniche.
- L'accordo che esiste tra due indicizzatori “umani” d'altra parte non è molto più alto
- L'approccio manuale, anche se qualitativamente superiore, non è scalabile
- In certi domini (es. Web) l'indicizzazione automatica è l'unica possibile