Sistemi di Elaborazione delle Informazioni

Informatica II

Ing. Mauro Iacono

Seconda Università degli Studi di Napoli

Facoltà di Studi Politici e per l'Alta Formazione Europea e Mediterranea "Jean Monnet"

-

PARSeC Research Group

Information Retrieval

(Fadini-Savy cap. 11)

Information Retrieval

- Tecniche e sistemi per gestire efficientemente l'archiviazione e il reperimento di *informazioni non strutturate* ovvero contenute in *documenti*:
  - sui quali sono note soltanto informazioni di tipo *semantico*
  - non organizzati in maniera schematica come i DB

Documenti

- Possono essere:
  - di tipo *testuale* (libri, articoli, riviste, leggi, sentenze, documenti tecnici di descrizione di un prodotto)
  - di tipo multimediale (disegni, stampe, foto, filmati, testi musicali, discorsi, dipinti, sculture, edifici, scavi archeologici)
- Problema: archiviazione
  - Documento completo o surrogato?
- Problema: ricerca dei dati
  - Catalogazione: richiede un esperto

Esempi

- Archivi storici/testi antichi
  - schede complete per tutti i materiali? trascrizione? copia fotografica?
- Letteratura di evasione
  - sola catalogazione e solo per autore e titolo
- Articoli scientifici
  - catalogazione per argomento, autore, parole chiave; archiviazione di copia in formato elettronico
- Opere d'arte
  - impossibile memorizzarla: uso di schede di catalogazione (quali criteri?)
- Testi rari/deperibili
  - Copia fotografica per conservazione e consultazione

Organizzazione dati e IR

- E' necessario estrarre dai documenti le due componenti (da separare):
  - informazioni strutturate
  - informazioni in senso semantico
- Dal punto di vista informatico un documento viene descritto dai suoi attributi che possono a loro volta essere in alternativa
  - strutturati
  - testuali
- Tali attributi vengono usati per le ricerche
  - strutturati: DB
  - testuali: necessaria una *standardizzazione dei termini*

Thesaurus

- La standardizzazione dei termini permette di costruire anche nel caso testuale un apposito limitato e predefinito dominio di definizione per ogni attributo detto *thesaurus* o *dizionario predefinito* dell'attributo
- Il thesaurus assume il ruolo di:
  - standardizzazione dei termini
  - ausilio alla catalogazione
  - ausilio alla ricerca
- Attenzione: l'uso della standardizzazione aumenta la potenza espressiva dei sistemi di archiviazione ma non può formalizzare tutto il contenuto di un documento

Fasi dell'IR

- Il processo di IR consiste in due fasi:
  - archiviazione
    - inserimento dei documenti originali in un *archivio dei documenti* (fisico) ordinato per *posizione* o *collocazione*, analisi e creazione del *termine di individuazione* coerente con un *linguaggio di individuazione* da inserire nell'*archivio di indagine* (elettronico)
  - indagine
    - formulazione della *richiesta*, analisi e formulazione del *termine di indagine*, confronto con l'archivio di indagine, individuazione ed elencazione dei documenti coerenti con relativa collocazione

Esempio: Autore e titolo

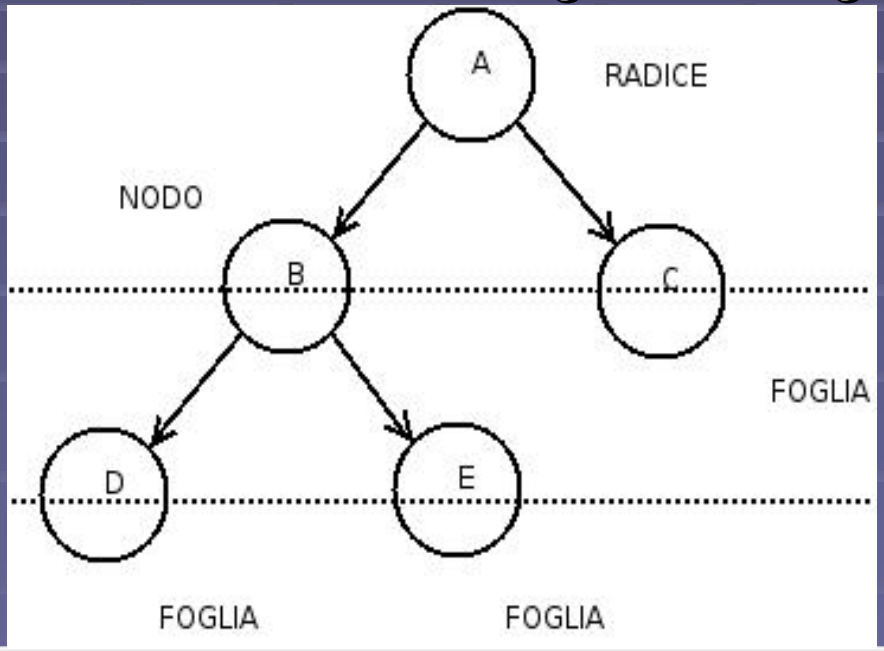
- Classificazione per Autore e titolo (forma base per le biblioteche)
  - Termine di individuazione: autore e titolo
  - Archivio di indagine: schedario
  - Archivio dei documenti: la biblioteca
  - Linguaggio di individuazione: regole per la redazione delle schede bibliografiche per lo schedario
  - Fase di archiviazione: schedatura
  - Fase di indagine: ricerca per autore
- Svantaggio: bisogna conoscere autore e titolo

Sistemi gerarchici

- I *sistemi di classificazione gerarchici* si basano sull'organizzazione di un certo numero predeterminato di argomenti (detti *categorie* o *classi* o *settori della conoscenza*) in una gerarchia
- Classi della conoscenza:
  - determinate per affinità
  - formate da documenti diversi che trattano tutti dello stesso argomenti allo stesso livello di generalità
  - concatenate ad albero per ordine di generalità decrescente
- Possibili ricerche su termini non presenti ma affini

Sistemi di tipo gerarchico puro

- Costruiti per sfruttare al massimo la gerarchia
- Principali:
  - Sistema di Classificazione Universale Decimale (UDC)
  - Sistema di Classificazione Decimale di Dewey (DDC)
  - Sistema della Libreria del Congresso degli USA (LC)



Esempio: sistema DDC

- Ampiamente usato in Italia (difetto: nuove discipline?)
  - Radice dell'albero: il *Sapere*
  - Al primo livello: 10 *classi* molto generali (000-900)
  - Al secondo livello: per ogni classe 10 *divisioni* meno generali (X00-X90)
  - Al terzo livello: per ogni divisione 10 *sezioni* più specifiche (XY0-XY9)
  - Ulteriori suddivisioni minori: XYZ.0-XYZ.9 e così via fino a 9 cifre
- Archiviazione: analisi per sottoclassi successive, collocazione su scaffali etichettati con il codice numerico, produzione di schede ordinate per codice
- Reperimento: ricerca sull'albero, posizioni adiacenti

Sistemi ad argomenti principali

- Detti anche *soggettari*
- Adottati dal sistema bibliotecario nazionale italiano, con responsabile la Biblioteca Nazionale Centrale di Firenze
- Un soggettario è un elenco di soggetti organizzati ad albero e indipendenti l'uno dall'altro (*foresta*)
  - Vantaggio: aperto a nuovi generi
  - Svantaggio: più difficile classificare documenti con nuove caratteristiche
  - Aggiunta di nuove voci: dietro apposita procedura di autorizzazione per evitare duplicazioni
  - Evoluzione: *rimandi tra argomenti* (“vedi” o “vedi anche”) che collegano tra loro più alberi (risolvono i sinonimi)

Esempio: soggettario BNI

- Soggettario delle Biblioteche Nazionali Italiane
  - Il soggetto può essere composto da più parole
  - Sottovoci separate da punti
  - Permesse schede di rimando
  - Ordinamento: soggetti a una parola, sottovoci del soggetto, soggetti a due parole separate da 'e', soggetti aggettivati, schede di rimando

```
DIRITTO
DIRITTO.Concetto
DIRITTO.Filosofia
DIRITTO E ECONOMIA
DIRITTO AMMINISTRATIVO
DIRITTO
AMMINISTRATIVO.Concetto
```

Sistemi analitico-sintetici

- Il metodo *analitico-sintetico* o *a faccette* prevede una classificazione mediante più elenchi di termini elementari ognuno relativo a una proprietà dell'oggetto da classificare (documento in senso lato)
- Le proprietà derivano da una analisi preliminare
- Adatto per oggetti d'arte
- Fase di archiviazione:
  - Analisi del documento e ricerca della n-pla più adatta, associazione dei dati ottenuti al documento
- Fase di reperimento
  - L'utente formula la n-pla più affine alla sua ricerca e ottiene tutti i documenti corrispondenti

Esempio: soggettario BNI

Oggetti da catalogare

Nome	Descrizione
A1	Vaso egizio in terracotta per rituali di tipo religioso
A2	Anfora di produzione sumera in argilla per l'oreficeria
A3	Coppa ornamentale babilonese in alabastro
A4	Monile aureo romano, per uso identificativo
A5	Moneta d'argento bizantina

Classificazione a faccette

	FORMA	MATERIALE	PERIODO	USO
A1	Vaso	Terracotta	Egizio	Religioso
A2	Anfora	Argilla	Sumerico	Oreficeria
A3	Coppa	Alabastro	Babilonese	Ornamentale
A4	Monile	Oro	Romano	Identificativo
A5	Moneta	Argento	Bizantino	Economico

Sistemi a chiavi

- Permettono di superare la necessità di un linguaggio di identificazione predefinito
- Idea: identificare in ogni documento alcune parole o espressioni significative che caratterizzino il documento (*chiavi* o *parole chiave*) e formino il thesaurus (problema degli omografi)
- Fase di archiviazione
  - individuazione delle chiavi (eventualmente fatta dall'autore in un thesaurus), creazione della tabella delle chiavi (documento -> chiavi relative), generazione dell'*indice inverso* (chiave -> documenti collegati)
- Fase di ricerca
  - automatica, sulle chiavi, con espressioni logiche

Reti semantiche

- I termini del thesaurus possono essere strutturati in una rete di collegamenti concettuali individuando:
  - componenti lessicali
  - relazioni semantiche
- Le componenti lessicali assumono forma di:
  - *descrittori* (termini che partecipano all'individuaz. della rete)
  - *non-descrittori* (considerati equivalenti ad altri descrittori)
  - *termini strumentali* (una sola parola, significato generico)
- Una relazione semantica può essere:
  - preferenziale
  - gerarchica
  - associativa

Relazioni semantiche

- Relazione semantica preferenziale
  - è il legame di equivalenza o sinonimità e si indica con USA (“Alunno USA Allievo”) o all'inverso SP (Sinonimo Preferenziale: “Allievo SP Alunno”)
- Relazione semantica di tipo gerarchico
  - Collega una due componenti lessicali, una più generica e una più specifica e si indica con TL: (Termine Largo, “Veicolo TL Auto”) o all'inverso TS (Termine Stretto, “Auto TS Veicolo)”, oppure TA (Termine più Ampio) per la radice
- Relazione semantica associativa
  - Indica un legame biunivoco qualsiasi e si indica con RT (Termine in Relazione, “vittima RT incidente” per casualità)