

Sistemi di Information Retrieval

Argomenti della lezione

- ◆ **Caratteristiche generali di un sistema di IR**
 - Informazioni gestite: informazione in un documento (contenuto, struttura, metadati)
 - Modalità di interazione con l'utente
 - Misure di efficacia: Precision/Recall
- ◆ **Architettura di un sistema di IR**
 - Componenti funzionali
 - Operazioni di trasformazione del testo
 - Indicizzazione
 - Modelli di IR

Sistemi di Information Retrieval

- ◆ **Definizione**
 - Sistemi specializzati nella gestione di grandi collezioni di documenti e nel recupero in base al contenuto
 - Information Retrieval (IR) = "recupero dell'informazione"
- ◆ **In generale IR tratta problemi e tecniche di**
 - rappresentazione, memorizzazione, organizzazione e accesso ad informazioni non-strutturate
- ◆ **Library:** *A warehouse of information which has been indexed and cataloged to make retrieval of information easier in the future.*

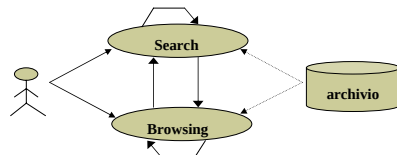
Informazioni gestite

- ◆ **Informazione non strutturata (o semi-strutturata)**
 - Tradizionalmente documenti testuali
 - IRS multimediali: gestione e recupero di documenti multimediali (trattati più avanti nel corso)
- ◆ **Informazione tipica associata ad un documento**
 - **Contenuto**
 - Termini
 - Concetti
 - **Struttura**
 - Stile/organizzazione di presentazione: capitoli, sezioni, riquadri,...
 - **Metadati non semantici**
 - Autore, data, ...

Servizi offerti all'utente da un sistema di IR

◆ Due modalità di accesso:

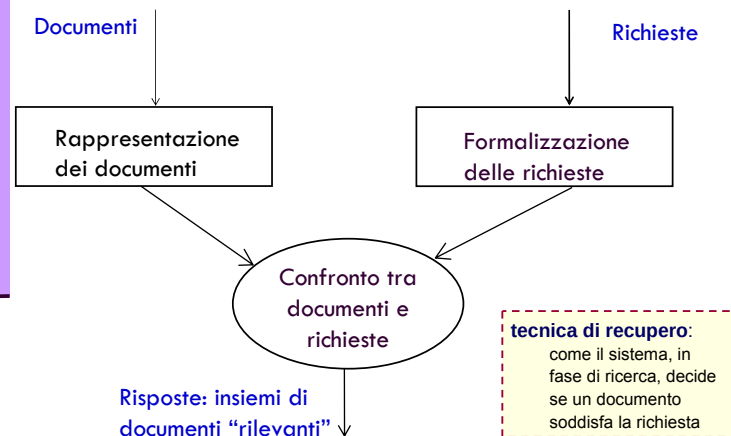
- **Search (ricerca):** una sessione di Search comporta la specifica degli interessi dell'utente e la sua trasformazione in una query (usualmente, un insieme di parole chiave o keywords)
- **Browsing (navigazione):** l'utente può usare un'interfaccia interattiva per visualizzare alcuni dei documenti proposti, e seguire hyperlink a partire da documenti che più lo interessano



◆ Interattività

- Se l'interesse dell'utente era mal specificato, o troppo vasto, è necessario raffinare le operazioni di browsing e search

Il processo di IR



Il processo di IR: il punto di vista dell'utente

◆ problema informativo: bisogno di informazione dell'utente

- specifica "approssimata" del contenuto: documenti che contengono una certa parola, frase, ...
- Es.: *documenti che (1) contengono informazioni sulle squadre di tennis dei college americani che hanno partecipato al torneo NCAA e (2) parlano dell'allenatore della squadra*

◆ query:

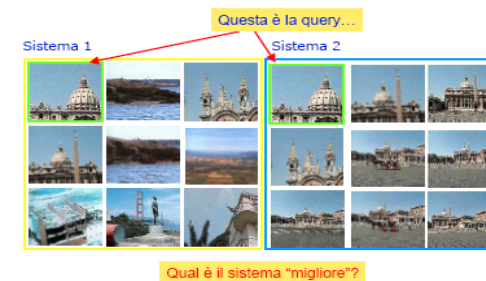
- codifica del problema informativo (in un linguaggio di interrogazione), adatta all'uso da parte di un motore di ricerca
- Es.: **SEARCH "tennis" AND "college" AND "NCAA" AND "trainer"**
- spesso, la query può essere vista come un piccolo documento di esempio che indica parte del contenuto dei documenti desiderati

◆ risultato:

- un insieme di documenti che hanno contenuto simile a quello specificato nella query

Query in un IRS multimediale

◆ Di solito la query è specificata tramite un esempio



Linguaggi di interrogazione

- ◆ Elementi di base sono termini/stringhe
- ◆ Operatori
 - operatori logici derivanti dall'algebra di Boole
 - operatori relazionali
 - operatori di prossimità
 - caratteri jolly

20 aprile 2004

L'informazione strutturata

9

Gli operatori logici (o booleani)

- ◆ AND (intersezione):
 - si usa per rintracciare documenti che soddisfano entrambi i criteri richiesti.
- ◆ OR (unione):
 - si usa per rintracciare documenti che soddisfano almeno uno dei criteri richiesti.
- ◆ NOT o AND NOT (complemento):
 - si usa per rintracciare documenti che soddisfano un determinato criterio escludendo quelli che però ne soddisfano un altro
- ◆ XOR o NOR (esclusione):
 - si usa per rintracciare documenti che soddisfano esclusivamente uno solo dei criteri richiesti.

20 aprile 2004

L'informazione strutturata

10

Operatori relazionali

- ◆ permettono di richiedere che il contenuto di un campo sia maggiore (>), minore (<) oppure uguale (=) ad un valore specificato.
- ◆ Questi operatori si applicano tipicamente alle date ed altri campi numerici;
 - altrimenti assumono il significato di "precedente/successivo nell'ordine alfabetico a -".

20 aprile 2004

L'informazione strutturata

11

Operatori di prossimità

- ◆ Gli operatori di prossimità sono una versione più precisa degli operatori AND e NOT:
 - permettono di richiedere che le parole in questione si trovino ad una determinata distanza fra loro ed eventualmente in un determinato ordine
- ◆ Operatori di prossimità tipici permettono di:
 - cercare due parole adiacenti e nell'ordine dato (l'operatore è spesso indicato con *ADJ*);
 - cercare due parole adiacenti in qualsiasi ordine (spesso indicato con *NEAR*)

20 aprile 2004

L'informazione strutturata

12

Caratteri jolly

- ◆ simboli che possono corrispondere a qualsiasi carattere, oppure a una sequenza di caratteri di qualsiasi lunghezza.
 - E' possibile ricercare tutti i testi che contengono una certa parola, una sua "variante" o un suo sinonimo
- ◆ Alcuni esempi :
 - troncamento: è un modo per evitare di dover digitare lunghe sequenze di termini alternativi aventi una stessa radice, separati dall'operatore OR (carattere jolly usato: asterisco, dollaro, due punti, o altro).
 - comput? --> computa, computo,...
 - comp* --> computer, compilato, compito, ecc.
 - mascheramento: si sostituisce un determinato numero di caratteri all'interno della parola (carattere jolly usato: soprattutto asterisco, o punto interrogativo).
 - comp*on? --> computation, comparison, competition...

20 aprile 2004

L'informazione strutturata

13

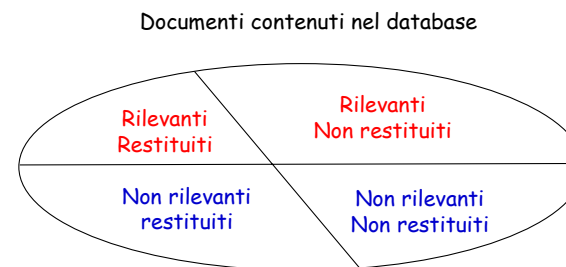
Problema: indeterminatezza

- ◆ Ambiguità del linguaggio naturale
 - L'interpretazione della richiesta da parte del sistema può divergere con le intenzioni dell'utente
- ◆ Richieste approssimate
 - L'utente potrebbe non sapere esattamente cosa cerca
- ◆ Approssimazione intrinseca al sistema
 - rappresentazione dei documenti incompleta
 - meccanismo di confronto approssimato

Valutazione del risultato di una query

- ◆ Risultato ideale: tutti e soli i documenti che interessano l'utente
 - Un sistema con tali caratteristiche non può però essere realizzato in pratica
 - Per tale motivo si valuta un sistema tanto più efficiente quanto più è in grado di avvicinarsi a tale requisito
- ◆ Due criteri di valutazione:
 - precisione (precision)
 - richiamo (recall)

Valutazione del risultato di una query



NB: si presuppone la conoscenza di quali siano i documenti rilevanti per la query

Recall (richiamo)

- ◆ La recall è la percentuale di documenti rilevanti restituiti rispetto al totale di documenti rilevanti presenti nel sistema

$$\text{Recall} = \frac{\text{Rilevanti Restituiti}}{\text{Totale Rilevanti}}$$

- ◆ Valore

- richiamo ideale è uguale ad uno (nei DBMS è garantito)
- In generale un valore inferiore ad uno perché il numero di documenti rilevanti restituiti è inferiore al numero di documenti rilevanti presenti nel sistema

Precision (precisione)

- ◆ La precision è la percentuale di documenti rilevanti sul totale dei documenti restituiti

$$\text{Precision} = \frac{\text{Rilevanti Restituiti}}{\text{Totale Restituiti}}$$

- ◆ Valore

- La condizione ideale è avere il 100% di precisione e richiamo
- In generale, aumentando il numero di documenti restituiti si aumenta il potere di richiamo a spese della precisione

Strategie di recupero: esatta

- ◆ Ipotesi:

- le informazioni specificate nella query siano esattamente contenute nella componente testuale del documento

- ◆ Svantaggi:

- molti documenti rilevanti sono ignorati, se il testo corrisponde solo parzialmente all'interrogazione
- non è possibile tenere in considerazione l'importanza relativa di concetti sia nell'interrogazione che nei documenti
- la logica del linguaggio di interrogazione risulta spesso complicata
- l'efficacia dipende dalla misura in cui le rappresentazioni da confrontare siano basate o meno sullo stesso vocabolario

Strategie di recupero: per similitudine

- ◆ Ipotesi:

- le informazioni specificate nella richiesta possano essere contenute parzialmente nel documento (corrispondenza parziale)

- ◆ Vantaggi

- maggiore flessibilità
- i documenti ritrovati possano essere ordinati per valori decrescenti di rilevanza

Ranking

- ◆ Il ranking è un ordinamento dei documenti restituiti da una interrogazione che riflette il grado di rilevanza dei documenti rispetto all'interrogazione
- ◆ I criteri per effettuare il ranking dipendono dal modello adottato per rappresentare documenti e query

Sistemi di IR vs. DBMS

- ◆ Rappresentazione delle informazioni.
 - Nei DBMS le informazioni si rappresentano come insiemi di dati strutturati (record) e relazioni fra insiemi
 - Negli IRS le informazioni si rappresentano come insiemi di documenti
- ◆ Rappresentazione delle richieste (query)
 - Nei DBMS l'utente descrive, in modo completo e preciso, ciò di cui ha bisogno
 - negli IRS si specifica il contenuto dei documenti da recuperare, mediante una "descrizione" abbreviata, e pertanto soggettiva e incompleta
- ◆ Risultato: cosa fornisce il sistema in risposta a una query
 - Nei DBMS sono forniti solo le tuple che soddisfano la condizione di ricerca;
 - negli IRS vengono forniti documenti probabilmente rilevanti

IRS vs. DBMS

	DBMS	IRS
Tipo di informazioni	Dati strutturati	Documenti
Richiesta	Completa e precisa	Incompleta e vaga
Criterio di scelta	Corrispondenza esatta	Corrispondenza parziale
Risultato	Dati richiesti	Documenti probabilmente rilevanti

IRS vs. DBMS (2)

- ◆ Query su DBMS: specifica esatta dei dati da restituire
 - Esempio: `SELECT Nome, Indirizzo
FROM Impiegati
WHERE AnnoAssunzione ≥ 1970 AND Stipendio > 2.000`
 - Risultato: tuple che soddisfano esattamente le condizioni di query
- ◆ Query IRS: indicazione sul contenuto
 - Es.: `SEARCH architett* AND (CAD OR (progetto AND calcolatore))`
 - Risultato = documenti che contengono i termini indicati
"... 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 ..."

Sistemi di IR

Sistemi di IR: qualità

◆ Aspettative dell'utente:

- Formulare le proprie richieste in modo semplice e intuitivo
- Trovare informazioni che soddisfino al meglio le proprie richieste, e in tempi rapidi

◆ Alcune misure di qualità per un IRS:

- **Efficacia**: il sistema recupera i documenti effettivamente rilevanti
- **Efficienza**: tempi rapidi per il calcolo dei risultati
- **Facilità di uso**: senza imporre all'utente vincoli rigidi nella specifica delle richieste

Implementazione di un IRS

Due aspetti:

- ◆ Sviluppo di modelli per rappresentare il problema del recupero di informazione: **modelli di IR**
- ◆ Sviluppo di **tecniche** per l'esecuzione delle query, e di **strutture** di supporto

Modelli di IR

- ◆ Un modello cerca di astrarre le caratteristiche salienti che stanno alla base di una classe di sistemi
- ◆ Nel caso degli IRS:
 - La strategia di rappresentazione dei documenti
 - La strategia di rappresentazione delle richieste
 - La modalità del confronto tra rappresentazioni di documenti e richieste

Modelli di IR: il modello booleano

- ◆ **Rappresentazione dei documenti**
 - Un insieme di termini che ne rappresentano il contenuto
- ◆ **Rappresentazione delle query**
 - espressioni booleane di termini, cioè termini combinati tra loro mediante AND, OR, NOT
- ◆ **Confronto: Il documento soddisfa esattamente l'espressione booleana:**
 - AND: i termini sono entrambi presenti
 - OR: almeno uno dei due termini è presente
 - NOT: il termine non è presente
- ◆ **Il modello booleano non consente il ranking**
 - Un documento soddisfa completamente la query o no

Il modello booleano: esempio

(film AND amore)
documenti che contengono "film" e "amore"

(dramma OR drammatico)
documenti che contengono "dramma" o "drammatico"

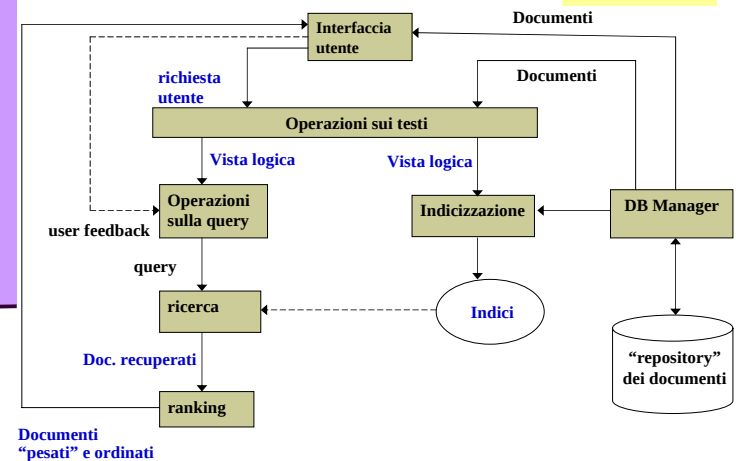
NOT (dramma OR drammatico)
... che non contengono "dramma" o "drammatico"

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

Operazioni fondamentali in un IRS

- ◆ **Trasformazione dei testi**
 - applicata a query e a documenti, sulla base di un modello (astratto) del testo
- ◆ **Costruzione e gestione di indici**
 - Strutture di puntamento per aumentare l'efficienza
- ◆ **Ricerca**
 - Confronto (matching)
 - Ordinamento (ranking)

Architettura di un IRS



Operazioni sui testi

◆ Scopo:

- generare una rappresentazione (*vista logica*), per un documento o per una query
- si ottiene una rappresentazione più generale e astratta, e più semplice da maneggiare in modo efficace

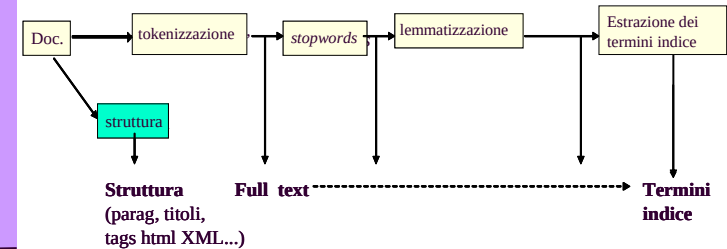
◆ Approccio full-text:

- la rappresentazione del documento consiste dell'intero insieme di parole che lo compongono

◆ Alternativa: estrazione di *termini indice* (*keywords*)

- efficienza: compressione (es. eliminazione di articoli, congiunzioni, lemmatizzazione...)
- efficacia: generalizzazione; si astrae da termini non rilevanti

Operazioni sui testi



Indicizzazione

◆ Indice

- Struttura di puntamento ai documenti della collezione, basata sul contenuto
- Ogni documento è puntato dai termini indice che contiene, o dai soggetti che tratta

◆ Motivazioni:

- Efficienza: minimizzare il tempo necessario per localizzare le informazioni
- Efficacia: ottenere tutte le informazioni "interessanti"

◆ Problema: come indicizzare in efficacemente i documenti?

- Bisogna modellare un documento sintetizzandone il contenuto ed estraendone solo alcune caratteristiche rilevanti
- Si vuole sempre buoni valori di recall e precision

Processo di indicizzazione

◆ Manuale:

- Un esperto sceglie quali concetti meglio caratterizzano il contenuto di un documento
 - I termini indice esprimono categorie semantiche
- Più "semantico" e quindi migliore
- Soggettivo, costoso (richiede tempo e competenze)

◆ Automatico:

- Fatto da un programma
- Più sintattico, su base statistica e quindi "peggiore"
- Economico, scalabile

Linguaggio di indicizzazione

◆ Linguaggio di indicizzazione:

- insieme dei termini scelti per indicizzare una collezione di documenti

◆ Basato su:

- *Dizionario*: termini ordinati alfabeticamente
- *Schema di classificazione*: codici che organizzano i termini gerarchicamente

Dizionario

◆ Nessuna relazione semantica fra i termini

- Le parole vicine sono "sintatticamente" simili (prefissi simili), ma non nel loro significato

◆ Esempio:

...	Acido ascorbico
Chimica	...
Chirurgia	...
Chemioterapia	...
...	
Medicina	...
...	Vitamina C

Schema di classificazione

◆ Esempio: schema di classificazione decimale di Dewey, usata in ambito bibliografico

15 psicologia
 152 psicofisiologia
 1521 percezioni sensoriali
 153 processi mentali
 154 subconscio

Sistemi di IR: esempi

◆ DIALOG Corporation

- offre più di 500 sistemi di IR su vari argomenti, quali scienze, medicina, economia e giornali elettronici (www.dialog.com)

◆ LEXIS-NEXIS

- ambito legale ed economico (www.lexis-nexis.com)

◆ OCLC (the online computer library center)

- offre l'accesso ad 1.5 milioni di articoli (www.oclc.org)

◆ H.W. Wilson

- offre più di 40 sistemi di IR per scuole ed istituzioni pubbliche (www.hwwilson.com)

Sistemi di IR: esempi

- ◆ **CA SEARCH: Chemical Abstract**
 - Contiene 14 milioni di documenti con una frequenza di aggiornamento di 11.000 documenti alla settimana
- ◆ **MEDLINE**
 - indicizza articoli provenienti da 3.700 riviste mediche
- ◆ **NewYork Times -- Fulltext**
 - contiene tutte le edizioni del NewYork Times dal 1981 ad oggi
- ◆ **PsycINFO: Psychological Abstract**
 - Contiene 1.5 milioni di documenti riguardanti psicologia, sociologia, psichiatria, linguistica ed antropologia dal 1887 fino ad oggi

IR e Web

- ◆ **Storicamente aveva una ristretta nicchia di interesse**
 - bibliotecari ed esperti dell'informazione (es. agenzie di stampa)
- ◆ **Nuovo interesse con l'avvento del Web**
 - Una sorgente di informazioni virtualmente illimitata
 - Accesso universale ed a basso costo
 - Non esiste un controllo editoriale centralizzato
 - Molti nuovi problemi si pongono: IR è vista come una area chiave per identificare soluzioni appropriate

IR nel Web

