

Introduzione all'Information Retrieval

Argomenti della lezione

- Definizione di Information Retrieval.
- Information Retrieval vs Data Retrieval.
- Indicizzazione di collezioni e ricerca.
- Modelli per Information Retrieval.
- Cenni sugli altri argomenti del corso.

Cos'è l'Information Retrieval

- L'Information Retrieval (IR) si occupa della rappresentazione, memorizzazione e organizzazione dell'informazione, al fine di rendere agevole all'utente il soddisfacimento dei propri *bisogni informativi*.
- Data una collezione di documenti e un bisogno informativo dell'utente, lo scopo di un sistema di IR è di trovare informazioni che potrebbero essere *utili*, o *rilevanti*, per l'utente.
- Rispetto alla teoria classica delle basi di dati, l'enfasi non è sulla ricerca di dati ma sulla *ricerca di informazioni*.

Perché è interessante parlare di IR?

- Il settore dell'Information Retrieval è stato studiato fin dagli anni '70.
- Negli anni '90, l'esplosione del Web ha moltiplicato l'interesse per IR.
- Il Web infatti non è altro che un'enorme collezione di documenti, sui quali gli utenti vogliono fare ricerche informazionali.
- Il problema principale è che non è semplice caratterizzare esattamente i bisogni informativi dell'utente.

Un esempio di bisogno informativo

- *Trova tutti i documenti che contengono informazioni sulle squadre di calcio partecipanti a campionati di prima divisione e che:*
 - *Provengono da organismi calcistici ufficiali;*
 - *Contengono informazioni sui risultati raggiunti nei tornei nazionali negli ultimi tre anni;*
 - *Forniscono l'indirizzo e-mail o il numero di telefono della società.*

Information Retrieval vs Data Retrieval

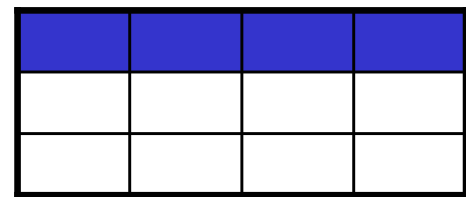
- Un sistema di Data Retrieval (ad esempio un DBMS) gestisce dati che hanno una struttura ed una semantica ben definita.
- Un sistema di Information Retrieval gestisce testi scritti in linguaggio naturale, spesso non ben strutturati e semanticamente ambigui.
- Di conseguenza:
 - Un linguaggio per Data Retrieval permette di trovare tutti gli oggetti che soddisfano esattamente le condizioni definite. Tali linguaggi (algebra relazionale, SQL) garantiscono una risposta *corretta e completa*.
 - Un sistema di Information Retrieval, invece, potrebbe restituire, tra gli altri, oggetti non esatti; piccoli errori sono accettabili e probabilmente non verranno notati dall'utente.

Dati strutturati, semistruutturati, non strutturati

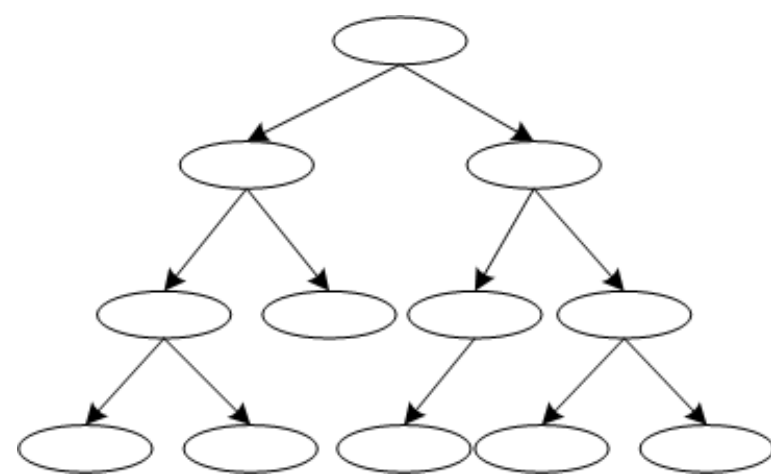
**Dati
strutturati**

**Dati
semistruutturati**

**Dati non
strutturati**



**DBMS
relazionali +
SQL**

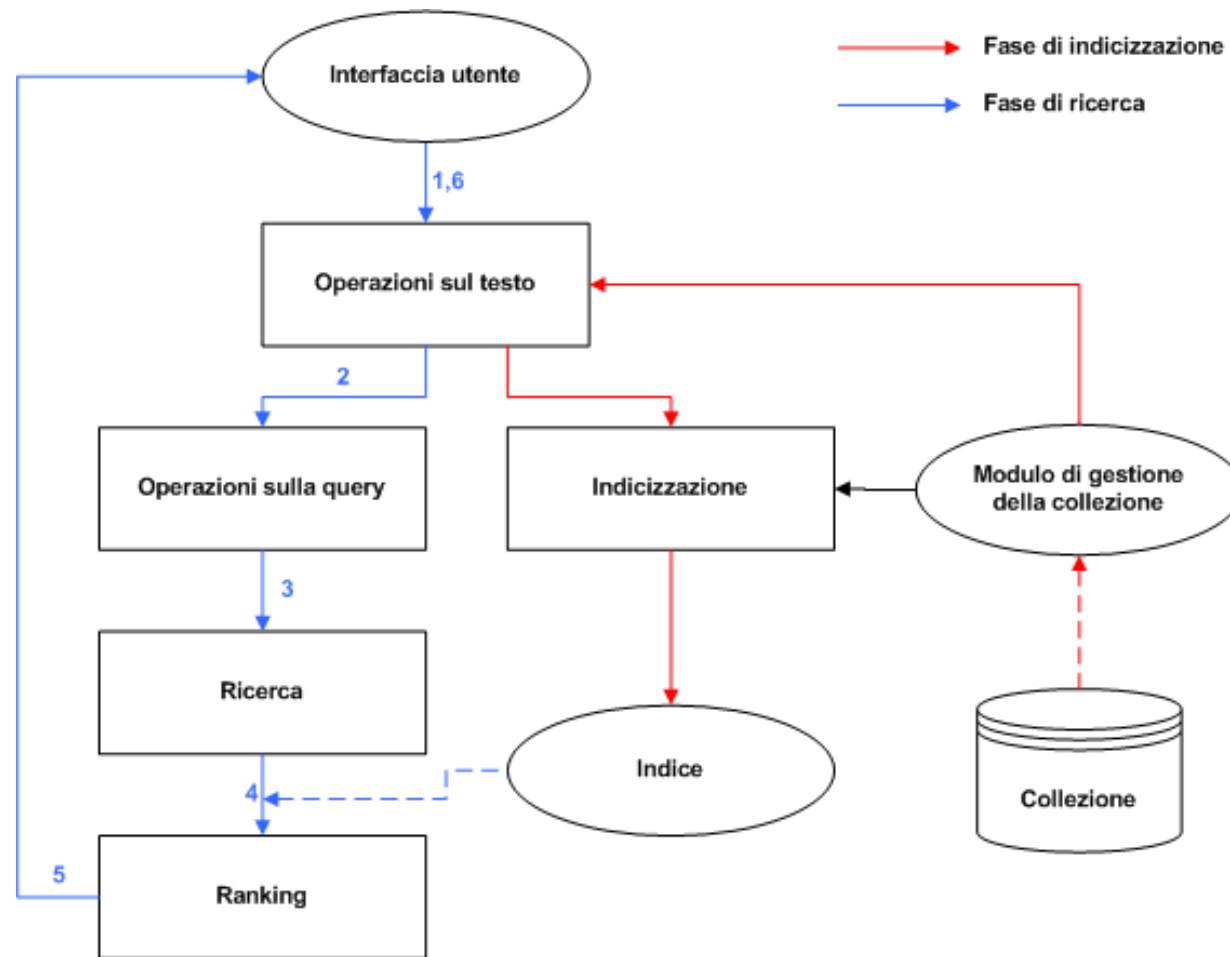


**DBMS XML
nativi o non
nativi +
XPath/XQuery**



?

Architettura di un tipico sistema di IR



Indicizzazione della collezione

- I sistemi di IR non operano sui documenti originali, ma su una *vista logica* degli stessi.
- Tradizionalmente i documenti di una collezione vengono rappresentati tramite un insieme di *keyword*.
- La capacità di memorizzazione dei moderni elaboratori permette talvolta di rappresentare un documento tramite l'intero insieme delle parole in esso contenute; si parla allora di vista logica *full text*.
- Per collezioni molto grandi tale tecnica può essere inutilizzabile; si utilizzano allora tecniche di modifica del testo per ridurre la dimensione della vista logica, che diventa un insieme di *index term*.
- Il modulo di gestione della collezione si occupa di creare gli opportuni indici, contenenti tali termini.

Tecniche di indicizzazione

- Le tecniche di indicizzazione studiate per le basi di dati relazionali (ad es. B-Tree) non sono adatte per i sistemi di Information Retrieval.
- L'indice più utilizzato è l'**indice invertito**:
 - Viene memorizzato l'elenco dei termini contenuti nei documenti della collezione;
 - Per ogni termine, viene mantenuta una lista dei documenti nei quali tale termine compare.
- Tale tecnica è valida per query semplici (insiemi di termini); modifiche sono necessarie se si vogliono gestire altre tipologie di query (frasi, prossimità ecc.).

Operazioni sul testo

- Il numero di termini indicizzati viene ridotto utilizzando una serie di tecniche, tra cui:
 - Eliminazione delle **stopword**: articoli, congiunzioni ecc.;
 - **De-hyphenation**: divisione in più parole di parole contenenti un trattino;
 - **Stemming**: riduzione delle parole alla loro radice grammaticale;
 - **Thesauri**: gestione dei sinonimi.
- L'utilizzo di tali tecniche è sicuramente positivo dal punto di vista dell'occupazione di spazio, ma non sempre migliora la qualità delle risposte ad una query.

Processo di ricerca di informazioni

1. L'utente specifica un bisogno informativo...
2. Che viene analizzato e trasformato utilizzando le stesse operazioni sul testo applicate alla collezione;
3. La query viene eventualmente trasformata...
4. Per poi essere eseguita, utilizzando indici precedentemente costruiti, al fine di trovare documenti rilevanti;
5. I documenti trovati vengono ordinati in base alla probabilità che siano rilevanti e ritornati in tale ordine all'utente;
6. L'utente esamina i documenti ritornati ed eventualmente raffina la query, dando il via ad un nuovo ciclo.

Modelli di Information Retrieval

- In questo corso discuteremo dei due modelli classici di Information Retrieval:
 - **Modello Booleano;**
 - **Modello Vettoriale.**
- Formalmente un modello di Information Retrieval è una quadrupla (D, Q, F, R) , dove:
 - D è un insieme di viste logiche dei documenti della collezione;
 - Q è un insieme di viste logiche (dette query) dei bisogni informativi dell'utente;
 - F è un sistema per modellare documenti, query e le relazioni tra loro;
 - $R(q_j, d_j)$ è una funzione di ranking che associa un numero reale ad una query q_j e un documento d_j , definendo un ordinamento tra i documenti con riferimento alla query q_j .

Il modello booleano

- Il modello booleano è il modello più semplice; si basa sulla teoria degli insiemi e l'algebra booleana.
- Storicamente, è stato il primo ed il più utilizzato per decenni.
- I documenti vengono rappresentate come insiemi di termini.
- Le query vengono specificate come espressioni booleane, cioè come un elenco di termini connessi dagli operatori booleani AND, OR e NOT.
- La strategia di ricerca è basata su un criterio di decisione binario, senza alcuna nozione di *grado di rilevanza*: un documento viene considerato rilevante o non rilevante.

Il modello vettoriale

- Il modello vettoriale è giustificato dall'osservazione che assegnare un giudizio binario ai documenti (1=rilevante, 0=non rilevante) è troppo limitativo.
- Nel modello vettoriale ad ogni termine nei documenti o nelle query viene assegnato un peso (un numero reale).
- I documenti e le query vengono quindi rappresentati come *vettori* in uno spazio n -dimensionale (n = numero di termini indicizzati).
- La ricerca viene svolta calcolando il *grado di similarità* tra il vettore che rappresenta la query e i vettori che rappresentano ogni singolo documento: i documenti con più alto grado di similarità con la query hanno più probabilità di essere rilevanti per l'utente.
- Il grado di similarità viene quantificato utilizzando una qualche misura, ad esempio il coseno dell'angolo tra i due vettori.

Valutazione di un sistema di IR

- Come è possibile rispondere alla domanda “*quale di questi due sistemi di IR funziona meglio*”?
- Un sistema tradizionale di Data Retrieval può essere valutato oggettivamente, sulla base delle performance (velocità di indicizzazione, ricerca ecc.).
- In un sistema di IR tali valutazioni delle performance sono possibili, ma, a causa della soggettività delle risposte alle query, le cose si complicano...
- Quello che si vorrebbe in qualche modo misurare è la *soddisfazione* dell'utente.
- Vedremo che esistono delle misure standard per valutare la bontà delle risposte fornite da un sistema di IR.

Web Search

- Come detto, l'Information Retrieval è nata per gestire collezioni statiche e ben conosciute: testi di legge, enciclopedie ecc.
- Quando la collezione di riferimento diventa il Web, le cose cambiano completamente:
 - La collezione è dinamica, molto variabile nel tempo;
 - Le dimensioni sono enormi;
 - I documenti non sono sempre disponibili;
 - Le query degli utenti sono ancora più imprecise e vaghe.
- Vedremo le tecniche utilizzate dai passati, presenti e.. futuri motori di ricerca.

Nuove frontiere: Structured IR

- Abbiamo presentato (e studieremo) l'Information Retrieval come una scienza che ha a che fare con dati non strutturati.
- Tuttavia, negli ultimi anni sta emergendo la necessità di applicare tecniche di IR anche a dati semistrutturati, in particolare a documenti XML; si parla allora allora di **Structured Information Retrieval (SIR)**.
- Molte cose cambiano. Ad es. in IR la risposta ad una query è un elenco di documenti; in SIR, la risposta è un elenco di... ? Documenti XML? O frammenti di documenti XML? O singoli elementi?
- Ultimamente sono state avanzate proposte per estendere XQuery con capacità di ricerca *full-text*.

Argomenti della lezione

- Definizione di Information Retrieval.
- Information Retrieval vs Data Retrieval.
- Indicizzazione di collezioni e ricerca.
- Modelli per Information Retrieval.
- Cenni sugli altri argomenti del corso.

Risorse per questa lezione

- MIR, ch. 1