



## Fondamenti di Informatica

Nicola Orio

### Il reperimento dell'informazione

- Motori di ricerca
- Localizzazione di pagine Web
- Indicizzazione: analisi, stop-words, stemming e pesatura
- Il ruolo dell'utente, ciclo valutazione/presentazione
- Il modello booleano: esempi di interrogazioni
- Il problema della rilevanza: valutazione, richiamo e precisione
- Altri approcci alla ricerca di informazioni nel Web

Parte di queste trasparenze è basata sul lavoro della Prof. Maristella Agosti

## Reperimento dell'informazione - 1

*"Information is the currency of democracy"* - T. Jefferson

La *disponibilità* di informazioni, anche in formato digitale, non implica che gli utenti possano averne *accesso*

- Gli utenti devono poter sapere
  - *Quali* informazioni sono disponibili, ovvero se sono presenti informazioni utili
  - *Come* raggiungere queste informazioni

Il problema di come reperire informazione aumenta con la *mole dei dati* messi a disposizione

- Ad esempio, nelle biblioteche esiste un *indice catalografico* ed un *ordinamento* per argomento, autore e così via
  - La catalogazione viene di solito svolta *manualmente* ed è quindi molto *lunga* e passibile di *errori*

## Reperimento dell'informazione - 2

La catalogazione *describe*, in maniera sintetica e di rapido accesso, il *contenuto informativo* dei documenti

E' possibile automatizzare l'estrazione del contenuto informativo, operazione che viene definita *indicizzazione*

- E' necessario *creare un modello* che consenta di estrarre l'*informazione rilevante* in modo automatico
  - Nei documenti testuali l'informazione è contenuta nella *semantica delle parole* che compongono i documenti
  - E' più difficile definire il contenuto semantico di documenti in *formati non testuali*, ad esempio musica o immagini

Una volta indicizzati i documenti è possibile effettuare delle ricerche nei *soli indici* dei documenti

- La ricerca negli indici è meno *onerosa computazionalmente*

## Reperimento dell'informazione - 3

L'utente ha un ruolo *cruciale* nel reperimento dell'informazione tramite mezzi informatici

- Una ricerca viene svolta più *efficacemente* se l'utente:
  - Sa cosa sta cercando e può indicare chiaramente la propria *esigenza informativa*
  - Conosce il funzionamento del sistema e la *sintassi del linguaggio* di interrogazione
  - Sa valutare le risposte del sistema e, in base a queste, formulare eventualmente una nuova richiesta *più precisa*

La ricerca è un processo *iterativo e interattivo*

- Una sola ricerca *non è di norma sufficiente* ad ottenere le informazioni desiderate
  - L'utente deve *interagire* con il sistema, valutandone le risposte, e *iterare* la propria richiesta variandone il contenuto

# Reperimento dell'informazione nel Web

Il Web non è gestito in maniera unitaria e coerente

- Chiunque può *creare una pagina o un sito Web*, nel quale può mettere qualsiasi genere di informazione
  - Non è possibile *effettuare controlli* sul contenuto della gran parte dei siti e quindi sulla loro attendibilità

Ogni giorno nascono *migliaia di nuovi siti* e altrettanti scompaiono

- Milioni di singole pagine sono continuamente modificate
  - Per tener traccia di questa *continua evoluzione* servono strumenti informatici potenti ed efficaci

Il numero di pagine Web supera le *centinaia di miliardi*

- Potenzialmente il Web contiene *ogni tipo di informazione*, basta che qualcuno abbia deciso di aggiungerla nel proprio sito
  - Il problema principale è scoprire *dove* si trova l'informazione

## I motori di ricerca - 1

La rapida espansione del Web ha reso necessaria la scrittura di programmi che *aiutino l'utente* a reperire l'informazione

- Ci si riferisce al sistema di programmi per il reperimento dell'informazione con il termine di *motori di ricerca*

### Definizione di motore di ricerca

- Un motore di ricerca, in inglese *search engine*, è un sistema in grado di *localizzare, indicizzare e ricercare* le pagine Web

Alcuni motori di ricerca:

- [www.google.com](http://www.google.com)
- [www.altavista.com](http://www.altavista.com)
- [www.yahoo.com](http://www.yahoo.com)
- [www.virgilio.it](http://www.virgilio.it)
- E inoltre Lycos, HotBot, Excite, Infoseek, Arianna, ...

## I motori di ricerca - 2

Un motore di ricerca, d'ora in poi SE (da Search Engine), opera in *tre fasi* distinte

- *Localizzazione* delle pagine Web (semiautomatica)
  - Il Web si modifica continuamente e vengono continuamente create nuove pagine, il SE *deve trovarle*
- *Indicizzazione* delle pagine localizzate (automatica)
  - Il SE estrae per ogni pagina le informazioni e le organizza in modo da *riaccedervi rapidamente*
- *Ricerca* (interattiva)
  - Quando un utente formula una richiesta al SE, questo recupera le pagine Web che ritiene più *rilevanti* per le esigenze informative espresse dall'utente
  - Il risultato della fase di ricerca è una *nuova pagina Web*, o una serie di pagine, contenenti i *link* ai documenti rilevanti

## Localizzazione delle pagine Web - 1

La componente dei SE demandata alla localizzazione delle pagine Web è denominata *Web Search Agent (WSA)*

- I WSA sono anche denominati alternativamente: *spider*, *crawler*, *wanderer* e raramente anche *worm*

Il WSA localizza le pagine Web, e in generale i documenti in formati diversi dall'HTML, lavorando *ricorsivamente*

- Parte da una *lista di URL noti*, forniti dai gestori del SE
- *Analizza* i documenti per vedere se questi *contengono link* a nuovi URL al di fuori della lista
- *Aggiorna* la propria lista di URL e visita i documenti agli URL aggiunti al fine di *trovare ancora nuovi link*
- Ad ogni iterazione *aggiunge nuovi URL* e visita i documenti associati per identificare *ancora nuovi URL*

## Localizzazione delle pagine Web - 2

Il WSA può localizzare solamente i documenti e le pagine Web che sono *raggiungibili a partire dalla lista iniziale* di URL

- La porzione di Web localizzata da un WSA è *molto piccola* rispetto al Web intero
  - Non è possibile sapere con precisione quanta sia la parte del Web “sommersa”, ovvero *invisibile* ai motori di ricerca
- Le pagine che non sono puntate da nessun'altra, non potranno *mai essere localizzate*
  - I SE consentono ai creatori di pagine Web di *pubblicizzarle*, ovvero che i loro URL siano inseriti nella lista iniziale di URL
- Le pagine protette da password sono *irraggiungibili* dai WSA
- Il processo di localizzazione richiede diversi giorni
  - I SE hanno una visione del Web che è *già vecchia*, per questo molte volte forniscono *dead-link*

## Indicizzazione

L'indicizzazione consente di *rappresentare il contenuto semantico* di un documento

- Il documento viene rappresentato da *descrittori*, chiamati appunto *indici*
  - Un caso molto importante è l'indicizzazione dei documenti *testuali*, tramite l'estrazione di *parole chiave* o *keywords*

L'indicizzazione può essere svolta

- *Manualmente* o in modo *automatico* o *semiautomatico*
- Estrahendo l'informazione *direttamente dal documento* o utilizzando *altre fonti*, come dizionari o metainformazioni

L'indicizzazione fornisce una rappresentazione *più compatta* del contenuto informativo del documento

- Gli indici sono utilizzati come *surrogati* del contenuto informativo del documento durante la fase di ricerca

## Indicizzazione automatica di testi - 1

L'indicizzazione automatica (*automatic indexing*) di un documento testuale è il processo che:

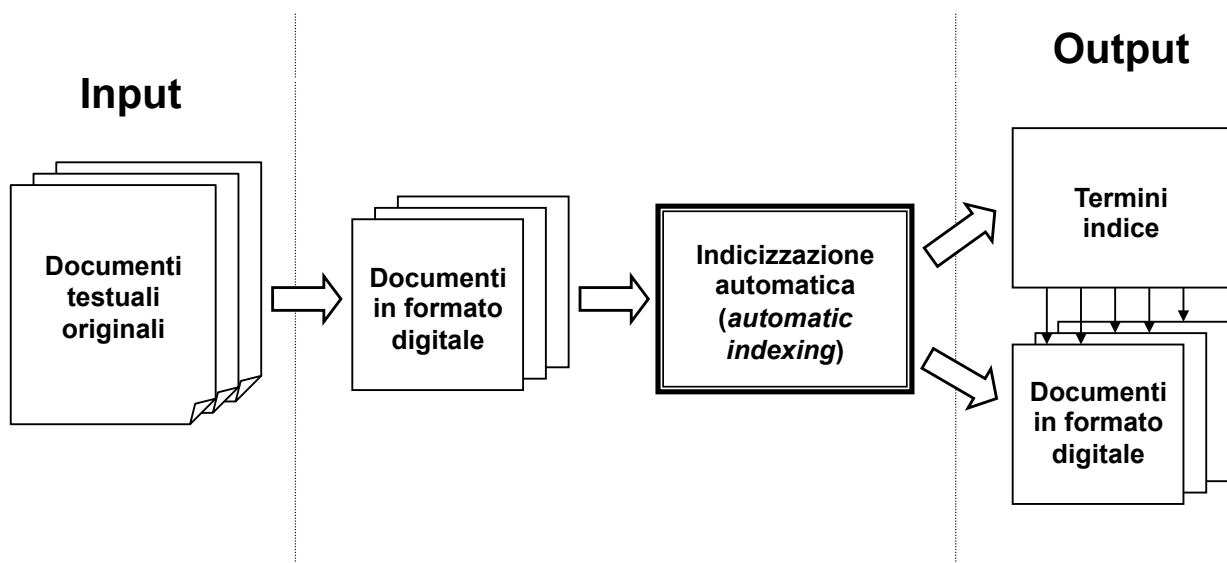
- *Esamina* automaticamente gli oggetti informativi che compongono il documento
  - Gli oggetti sono le *parole*, o le *frasi*, che compongono il testo
- Produce una lista dei *termini indice* (index terms) presenti nell'intera collezione di documenti
  - L'estrazione di termini indice viene fatta da appositi *algoritmi*
  - I termini indice sono *collegati* ai diversi documenti che li contengono
  - Durante la ricerca sarà quindi sufficiente fare riferimento alla *sola lista dei termini indice*, e non all'intera collezione

L'uso degli indici *semplifica ed accelera* la ricerca

- L'*indice analitico* di un libro ne è un esempio

## Indicizzazione automatica di testi - 2

Esempio di passaggio da una collezione di documenti testuali alla lista dei termini indice



## Indicizzazione automatica di testi - 3

L'indicizzazione automatica di documenti testuali viene eseguita in *più fasi*, che devono essere attuate in sequenza

- *Analisi lessicale* e selezione delle parole
- Rimozione delle parole molto comuni, o *stop-words*
- Riduzione delle parole originali alle rispettive *radici semantiche*
- Creazione dell'*indice*
- Eventuale *pesatura* degli elementi dell'indice

I SE disponibili in rete, e i sistemi commerciali in genere, *non implementano* necessariamente tutte queste funzionalità

- Ogni funzionalità necessita di *calcoli aggiuntivi*, il cui costo può non essere compensato da un effettivo miglioramento
- La ricerca nel settore del reperimento dell'informazione (*information retrieval*) si occupa anche di trovare nuove *metodologie* per l'indicizzazione automatica

## Esempio di collezione di documenti

**D1** L'enorme quantità di informazioni presenti nelle pagine Web rende necessario l'uso di strumenti automatici per il recupero di informazioni...

**D2** I presenti hanno descritto le fasi del recupero dell'enorme relitto ma le informazioni non concordano su tipo e quantità di strumenti in uso...

**D3** E' stato presentato nel Web un documento che informa sulle enormi difficoltà che incontra chi usa uno strumento informativo automatico...

# Analisi lessicale e selezione delle parole

Un testo è rappresentato da una *successione di simboli*

- L'analisi lessicale è il processo di trasformazione del flusso di simboli in un *flusso di parole* (dette *tokens*)
  - Le parole vengono *facilmente identificate* grazie alla presenza di spazi, a capo e segni di interpunzione
  - Le parole hanno un *significato a prescindere dal loro ordine*
- Nell'esempio, l'analisi lessicale porterebbe:
  - **D1**: automatici di di di enorme il informazioni informazioni l' l' necessario nelle pagine per presenti quantità recupero rende strumenti uso web
  - **D2**: concordano del dell' descritto di e enorme fasi hanno i in informazioni le le ma non presenti quantità recupero relitto strumenti su tipo uso
  - **D3**: automatico che che chi difficoltà documento è enormi informa informativo incontra nel presentato sulle stato strumento un uno usa web

## Rimozione delle stop-words - 1

Le parole *molto frequenti* nell'insieme di tutti i documenti portano *poca informazione* sul contenuto dei singoli documenti

- In una collezione di documenti sull'informatica, la parola "computer" *non serve a discriminare* i diversi documenti
- Alcune parole, oltre ad essere molto frequenti, *non hanno* un proprio *significato* semantico

Articoli, preposizioni, verbi ausiliari sono un esempio

Tali parole, denominate *stop-words*, possono essere *eliminate* dalla lista dei token

- Le *stop-words non sono utilizzate* per indicizzare i documenti

Nel Web, che contiene documenti su *qualsiasi argomento*, le *stop-words* sono le parole *molto frequenti* nella lingua in cui i documenti sono scritti



## Rimozione delle stop-words - 2

Se le stop-words sono *note a priori*, è possibile creare una *lista* che le contiene (detta *stop-list*)

- Ogni parola estratta dall'analisi lessicale viene confrontata con quelle nella stop-list e, se presente, viene *scartata*

Nell'esempio, una possibile lista di stop-words è:

*che chi del dell' di e i il in l' le ma nel nelle per su sulle un uno*

Nell'esempio, le parole restanti sarebbero:

- **D1**: automatici enorme informazioni informazioni necessario pagine presenti quantità recupero rende strumenti uso web
- **D2**: concordano descritto enorme fasi hanno informazioni non presenti quantità recupero relitto strumenti tipo uso
- **D3**: automatico difficoltà documento è enormi incontra informa informativo presentato stato strumento usa web

## Riduzione alle radici semantiche - 1

In molte lingue, parole che iniziano allo stesso modo, o che hanno delle parti in comune, possono avere la stessa *origine etimologica*

- Tali parole hanno spesso un contenuto informativo *molto simile*

E' possibile ridurre tutte le parole affini ad un'unica *radice semantica*

- L'operazione viene chiamata *stemming*, da “*stem*” che in inglese significa radice

In italiano, e in inglese, lo stemming si traduce spesso nell'*eliminazione della parte finale* delle parole

- Ad esempio, le parole musica, musicista, musicologo, musicale, musicante e il verbo musicare hanno la stessa radice

Esistono diversi algoritmi, la ricerca in questo fronte è molto attiva

## Riduzione alle radici semantiche - 2

L'operazione di stemming non viene sempre effettuata

- Le sole radici semantiche possono *non essere dei buoni indici* per un documento
  - “dentellato” e “dentificio” hanno la stessa radice “dent-”, ma significati e contesti molto diversi
- Lo stemming risulta comunque utile nelle lingue *molto inflesse* come l'italiano o il francese; è meno utile per l'inglese

Nell'esempio, le radici potrebbero essere:

- **D1**: autom enorm inform inform necessar pagin present quantità recuper rend strument us web
- **D2**: concord descr enorm fas ha inform no present quantit recuper relitt strument tip us
- **D3**: autom diffic document è enorm incontr inform inform present stat strument us web

## Pesatura dei termini indice

Non tutte le parole di un documento ne descrivono il contenuto semantico con la stessa *precisione*

- Si può associare un *peso* ai termini indice
  - Il peso indica l'*importanza di un indice per ciascun documento*

L'associazione di un peso ai termini di un documento viene effettuata utilizzando una *funzione di pesatura*

- La pesatura tiene normalmente conto della *frequenza* del termine nel *documento* e nella *collezione*

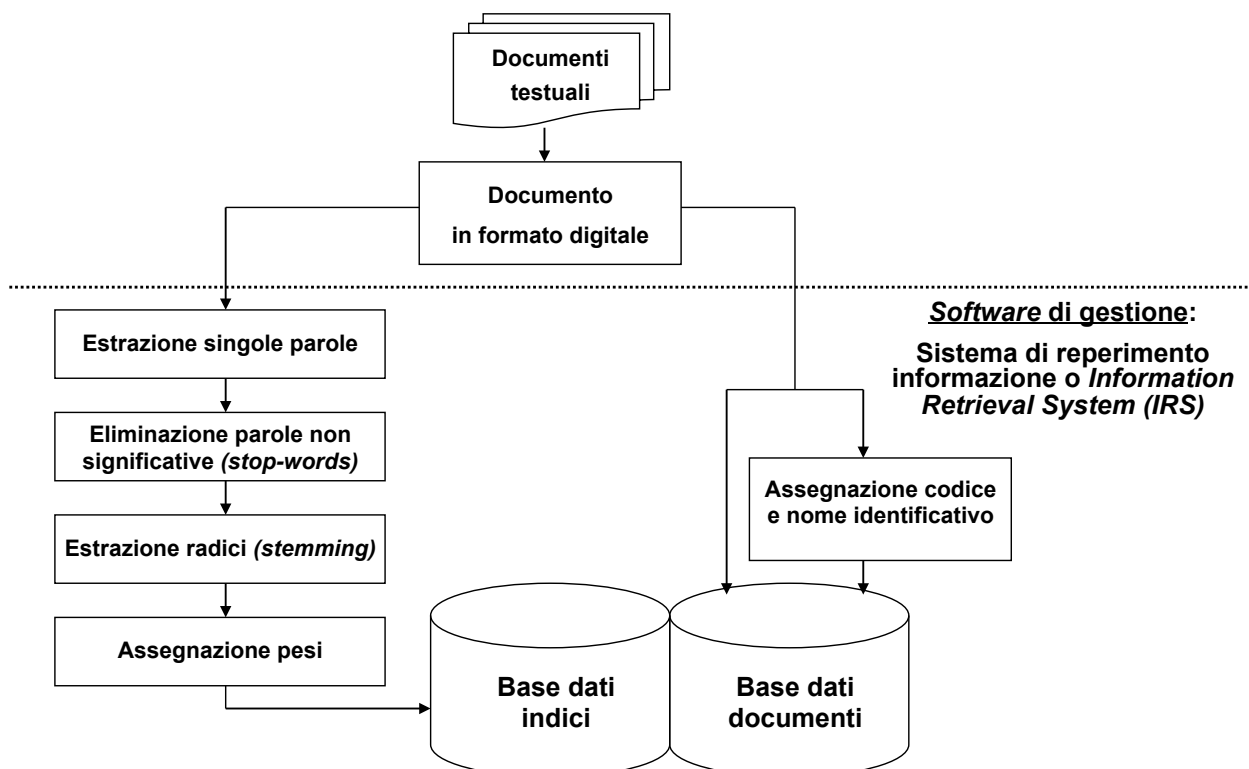
Sono possibili diversi sistemi di pesatura

- *Binaria*: il termine ha peso = 1 se presente e peso = 0 se assente
  - Non si tiene conto della frequenza ma della sola *presenza*
- In base alla *frequenza relativa*: si divide l'occorrenza del termine nel documento e per la sua occorrenza nella collezione

## Pesatura in base alla frequenza relativa

documenti parole	D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>
autom	1/2	0	1/2
concord	0	1	0
descr	0	1	0
diffic	0	0	1
document	0	0	1
è	0	0	1
enorm	1/3	1/3	1/3
fas	1/2	0	1/2
ha	0	1	0
incontr	0	0	1
inform	2/5	1/5	2/5
necessar	1	0	0
no	0	1	0
pagin	1	0	0
present	1/3	1/3	1/3
quantit	1/2	1/2	0
recuper	1/2	1/2	0
....		...	

## Il processo completo di indicizzazione



## La fase di ricerca

La ricerca consente di selezionare i documenti che sono *verosimilmente rilevanti* per le esigenze informative dell'utente

- I documenti possono essere reperiti grazie a delle *interrogazioni*, alla *navigazione* o con un *approccio misto*
  - Nelle interrogazioni in forma testuale l'utente deve fornire al sistema alcune *parole-chiave*, dette *keywords*

Le interrogazioni vengono normalmente dette *query*

L'interazione con l'utente avviene sotto forma di *ciclo presentazione/valutazione*

- *Presentazione*: il sistema mostra all'utente i documenti *ritenuti più pertinenti* per le sue esigenze informative
- *Valutazione*: l'utente *consulta* i documenti e decide se *soddisfano realmente* le sue esigenze informative

## Il ruolo dell'utente

Poiché la ricerca è il solo passo *interattivo*, l'utente ha un ruolo *determinante* per la sua efficacia

Gli utenti di un sistema di reperimento dell'informazione appartengono a *tipologie* molto diverse; ai due estremi vi sono:

- *Utente esperto*: è in grado di definire *esaustivamente* le proprie esigenze informative, utilizza dei *linguaggi avanzati* nelle interrogazioni; operatore in *biblioteche digitali*
- *Utente casuale*: *non conosce* esattamente cosa sta cercando, formula le interrogazioni in maniera *generica*, *affidandosi* alle potenzialità del sistema; navigatore *Web*

I sistemi per il reperimento dell'informazione sono nati per essere utilizzati da utenti *esperti*

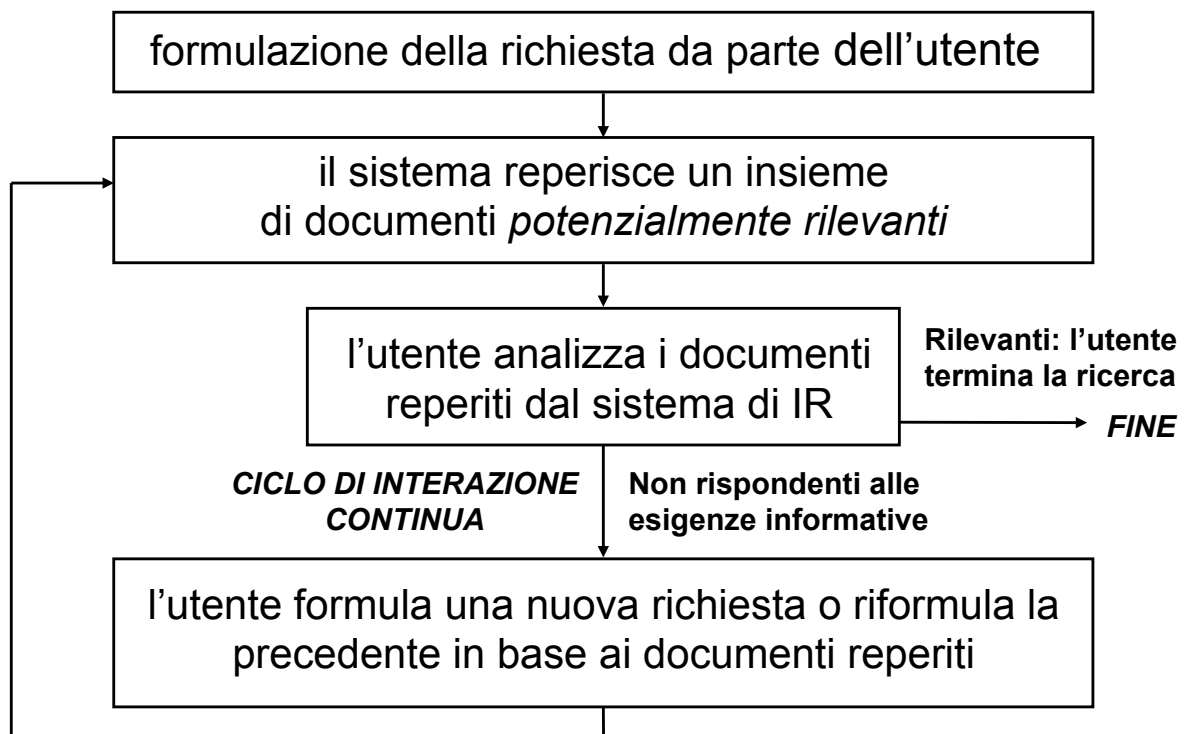
- Il numero di utenti è in costante *aumento*, e i sistemi devono quindi orientarsi verso utenti casuali

## Il ciclo presentazione/valutazione - 1

Ci si riferisce al modo in cui utente e sistema interagiscono con il termine di *ciclo presentazione/valutazione*; ad ogni iterazione:

- L'utente *interroga* il sistema formulando una query
  - L'utente deve utilizzare il *linguaggio* fornito dal sistema
- Il sistema *presenta* all'utente alcuni documenti ritenuti rilevanti
  - *Exact match*: solo i documenti che soddisfano *esattamente* la query vengono presentati all'utente
  - *Best match*: i documenti sono presentati all'utente in base ad una misura di *similarità* con la query (omettendo quelli troppo lontani)
- L'utente *valuta* i documenti presentati dal sistema
  - Uno dei maggiori problemi, specie nel Web, è l'*altissimo numero* di documenti normalmente reperiti
  - Se questi non soddisfano la sua esigenza informativa l'utente deve formulare una *nuova query*

## Il ciclo presentazione/valutazione - 2



# Il reperimento dell'informazione testuale

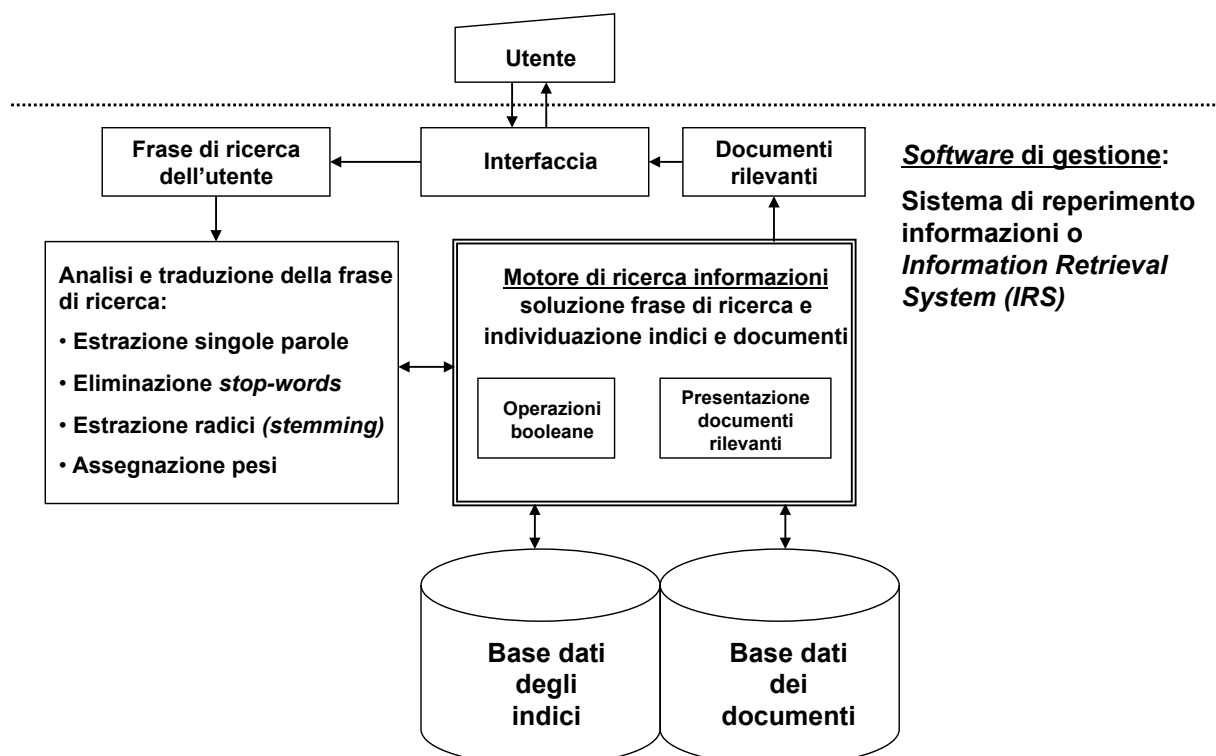
La fase di indicizzazione, eseguita *off-line prima dell'interazione* con l'utente, estrae degli indici dai documenti testuali

- Gli indici sono delle *parole*, che esprimono in modo sintetico il contenuto informativo dei documenti

La fase di ricerca, eseguita *on-line interagendo* con l'utente, si basa anch'essa sull'uso di *parole* che sintetizzano l'esigenza informativa

- L'utente formula la sua query utilizzando alcune parole, spesso indicate con il termine di *parole chiave* o *key-words*
  - Il sistema *indicizza* la query, così come ha fatto per i documenti, e calcola la *potenziale pertinenza* dei documenti in base al confronto tra gli indici della query e gli indici dei documenti
  - Sono possibili diverse *strategie* per il *calcolo della pertinenza*, la ricerca si occupa di trovare nuove soluzioni

# Il processo completo di reperimento



## Il modello booleano - 1

Un modello molto diffuso per i linguaggi di interrogazione è il modello *booleano*

- Il termine deriva dall'*algebra di Boole*, che è basata su *operazioni logiche* tra proposizioni, che possono essere *vere* o *false*

Il significato degli operatori booleani usati per *combinare coppie di parole della query* è il seguente:

- **AND**: *entrambi* i termini devono essere presenti
- **OR**: *almeno uno* dei termini deve essere presente
- **NOT**: il termine *non* deve essere presente

Alcuni esempi:

- musica **AND** pittura: documenti dove si parli di *entrambe*
- arte **OR** letteratura: documenti dove si parli di *almeno una*
- **NOT** scultura: documenti (tantissimi) che *non ne parlano*

## Il modello booleano - 2

Gli operatori booleani possono essere *combinati* in modo da creare interrogazioni anche *molto complesse*

- Se si cercano opere di Mozart che non siano per piano:  
Mozart AND (sonata OR concerto) AND (NOT piano)

Il modello booleano ha alcune caratteristiche:

- *Vantaggi*:
  - Implementazione software *intuitiva* e di *semplice* realizzazione
  - Efficace in ambienti *controllati* e con utenti ben *addestrati*
- *Svantaggi*:
  - Poco controllo sul *numero* dei documenti reperiti
  - *Impossibile l'ordinamento* per una qualche misura di similarità
  - Non c'è *pesatura* dei termini
  - La logica booleana *non è intuitiva* per gli utenti
  - Gli utenti devono *sapere con precisione* cosa cercano

## Esempi di interrogazioni booleane

$D_1$  = L'enorme quantità di informazioni presenti nelle pagine Web rende necessario l'uso di strumenti automatici per il recupero di informazioni

$D_2$  = I presenti hanno descritto le fasi del recupero dell'enorme relitto ma le informazioni non concordano su tipo e quantità di strumenti in uso

$D_3$  = E' stato presentato nel Web un documento che informa sulle enormi difficoltà che incontra chi usa uno strumento informativo automatico

recupero AND Web	→ $D_1$
recupero OR Web	→ $D_1, D_2, D_3$
recupero AND NOT relitto	→ $D_1$
(Web OR uso) AND strumenti	→ $D_1, D_2$
(Web OR uso) AND NOT strumenti	→ $D_3$
informazioni AND relitto AND studente	→ $\emptyset$
informazioni OR relitto OR Internet	→ $D_1, D_2$
<del>bologna OR NOT padova</del>	→ $D_1, D_2, D_3$

## Criteri di ordinamento dei documenti

La pesatura dei termini consente di ordinare i documenti

- Termine frequente nel documento = peso maggiore
- Termine frequente nella collezione = peso minore

Ci sono altre strategie che dipendono dalla struttura del testo

- Termini nel tag <title> o nell'URL = peso maggiore
- Termini nei tag H1-H6 = peso maggiore
- Termini all'inizio del testo = peso maggiore
- Termini vicini tra loro = rinforzo reciproco del peso

E' inoltre importante l'*autorevolezza* della pagina

- Misurata in base al numero di pagine che hanno un link
  - Ogni pagina ha una sua autorevolezza che ridistribuisce alle pagine verso cui ha un link
  - Bisogna essere puntati da pagine a loro volta autorevoli



## Il problema della rilevanza

La bontà di un sistema di reperimento dipende da quanti documenti reperiti sono *effettivamente rilevanti* per le esigenze informative

Le *prestazioni* di un sistema di information retrieval possono essere *calcolate*, per confrontare diversi sistemi

- Si deve conoscere *a priori* l'insieme dei documenti che *rispondono alle esigenze informative* dell'utente
  - E' praticamente impossibile conoscere la *rilevanza di milioni di documenti*, o miliardi se ci si riferisce al Web
  - La rilevanza è *soggettiva* e può *variare nel tempo*
  - Il giudizio sulla rilevanza di un documento *influisce* sul giudizio dei *successivi*

Sono state sviluppate delle *metodologie di sperimentazione*

- Utenti con *banche dati reali, collezioni sperimentali* in laboratorio

## Valutazione - 1

E' auspicabile che un sistema per il reperimento dell'informazione presenti *tutti e soli* i documenti rilevanti per l'utente

- Se così fosse, l'utente non avrebbe bisogno di valutare i documenti, e la ricerca *si esaurirebbe in un unico ciclo*

Vi sono due possibili *comportamenti negativi*, che rendono difficile la valutazione e onerosa la fase di ricerca

- *Effetto rumore*
  - Il sistema reperisce *anche* documenti *non rilevanti*; la valutazione e la consultazione sono *più onerose* perché i documenti rilevanti sono *diluiti*
- *Effetto silenzio*
  - Il sistema non reperisce *alcuni* documenti che sarebbero invece *rilevanti*; l'utente *non può accedere* ad una parte dell'informazione

## Valutazione - 2

Si definiscono alcune *misure* per valutare le *prestazioni* di un sistema di information retrieval

Dato un insieme di documenti e un'interrogazione, è possibile individuare quattro sotto-insiemi:

- **A**: documenti *correttamente reperiti* in quanto rilevanti
- **B**: documenti *erronamente reperiti* anche se non rilevanti (effetto rumore)
- **C**: documenti *correttamente omessi* in quanto non rilevanti
- **D**: documenti *erroneamente omessi* anche se rilevanti (effetto silenzio)

Questi sottoinsiemi possono essere individuati solo se si *conosce a priori* l'insieme dei documenti rilevanti

- La rilevanza viene di solito stabilita da un *gruppo di esperti*

## Richiamo e precisione

	dati rilevanti	dati non rilevanti
dati reperiti	<b>A (corretti)</b>	<b>B (inesatti)</b>
dati non reperiti	<b>D (omessi)</b>	<b>C (da omettere)</b>

In base ai quattro sottoinsiemi, è possibile calcolare due parametri molto usati per calcolare le prestazioni di un sistema di IR

- *Richiamo*
  - Rapporto tra il numero di documenti rilevanti reperiti (insieme A) e il totale dei documenti rilevanti (insiemi A e D)
  - In formula: **Richiamo** =  $A / (A + D)$
- *Precisione*
  - Rapporto tra il numero di documenti rilevanti reperiti (insieme A) e il totale dei documenti reperiti (insiemi A e B)
  - In formula: **Precisione** =  $A / (A + B)$

## Meta informazioni nel Web

Normalmente l'indicizzazione e la ricerca vengono svolte sul *contenuto* del documento

- L'utente può *accedere e leggere* le stesse parole usate per la creazione degli indici e per la ricerca
  - Il documento *describe sé stesso*, in base alle parole che lo formano

Le pagine HTML possono contenere dati, *non visibili* all'utente, che *descrivono esplicitamente* il contenuto della pagina Web

- Ci si riferisce a queste informazioni aggiuntive con il termine di *metainformazioni*
  - Le metainformazioni possono riguardare: descrizione con *parole chiave*, *lingua* utilizzata, *autore*, *data* di creazione e modifica

I motori di ricerca nel Web possono *utilizzare le parole chiave* contenute tra le metainformazioni, per migliorare le prestazioni

## Caratteristiche dei motori di ricerca

I motori di ricerca disponibili nel Web hanno alcune caratteristiche

- *Ricerca semplice*, per gli utenti inesperti, nella quale inserire semplicemente delle parole chiave
  - Il motore di ricerca *combina opportunamente* le parole chiave, il modo in cui questo viene fatto dipende dal SE
  - L'utente può comunque utilizzare il *linguaggio booleano* all'interno della ricerca semplice
- *Ricerca avanzata*, con delle funzionalità aggiuntive
  - La ricerca avanzata fornisce un'*interfaccia grafica* per aiutare l'utente a creare la query usando il *linguaggio booleano* e ad accedere ad alcune *funzionalità aggiuntive*
- Molti SE consentono anche di cercare *intere frasi*
  - Il testo cercato va normalmente messo tra *doppi apici*  
Ad esempio "information retrieval" cerca la frase intera, ed è diverso da information AND retrieval

## Altri approcci alla ricerca nel Web

L'utilizzo dei SE non è intuitivo per gli utenti inesperti, perciò sono disponibili altri modi per reperire informazioni nel Web

- *Directory*
  - Gestisce solo pagine che sono state *scelte* attraverso un processo di *selezione/catalogazione* editoriale o sottoposte dagli stessi utenti  
<http://www.yahoo.com>
- *Meta Search Engine*
  - Uno strumento che interroga *contemporaneamente diversi SE* e/o directory e *riassume* i risultati all'utente  
<http://www.metacrawler.com>
- *Portali*
  - Non sono dei reali sistemi per il reperimento dell'informazione, ma si presentano come *punti di partenza* per la navigazione  
<http://www.virgilio.it>

## Informazione in formato digitale - 1

- I supporti informatici presentano il vantaggio che lo *spazio fisico* necessario per contenere l'informazione è *irrisorio*
  - Teoricamente i bit *non hanno dimensione fisica*, anche se l'hardware occupa comunque spazio
    - L'informazione contenuta nei libri di un intero scaffale può essere contenuta all'interno di un disco di pochi centimetri cubi
- L'informazione rappresentata in formato digitale può essere *perfettamente replicata*, senza subire perdite
  - La copia fisica di un oggetto, ad esempio la fotocopia di un libro, non è mai identica all'originale
  - La sequenza di bit che rappresenta un oggetto informativo può, al contrario, essere riprodotta esattamente
  - Il formato digitale può essere utile anche per la *conservazione* nel tempo dell'informazione

## Informazione in formato digitale - 2

- Grazie alle reti di calcolatori, e a Internet, è possibile *trasferire* i dati in maniera pressoché *istantanea*
  - L'utente non deve trovarsi fisicamente nello *stesso luogo* dove sono immagazzinati i dati
  - L'informazione può essere distribuita su *supporti fisici diversi*
    - I supporti possono essere localizzati in luoghi anche *molto lontani*, purché siano collegati alla stessa rete
- Gli utenti possono usufruire dell'aiuto dei mezzi informatici per *organizzare e recuperare* le informazioni
  - Le informazioni possono essere *elaborate automaticamente* dal calcolatore per effettuare
    - La *catalogazione* (automatica o semiautomatica), l'*indicizzazione* e l'*archiviazione* di documenti
    - La *ricerca* e la *navigazione* tra oggetti informativi diversi

## Biblioteche digitali - 1

- Molte associazioni si occupano della creazione di strutture per la gestione delle informazioni dette *biblioteche digitali*, in inglese *digital libraries*
- Definizione di una biblioteca digitale:
  - Una biblioteca digitale è una *collezione organizzata* di documenti e informazioni in formato *digitale*
- Le biblioteche digitali combinano *raccolta* e *organizzazione* delle informazioni con le potenzialità delle tecnologie *informatiche*
  - Raccolta e organizzazione sono i passi principali delle biblioteche tradizionali
  - La raccolta delle informazioni nelle biblioteche digitali prevede normalmente l'*acquisizione* in formato digitale di documenti originamente su carta, pellicola o nastro magnetico

## Biblioteche digitali - 2

- Le biblioteche digitali presentano alcune differenze con le biblioteche tradizionali
  - *Vantaggi*
    - Facile *raggiungibilità* tramite i collegamenti in rete
    - I documenti *non deperiscono* nel tempo
    - I dati sono contenuti in uno *spazio* fisico molto *ridotto*
    - La *ricerca* e la prima consultazione sono molto *rapide*
    - I *costi* di mantenimento sono *ridotti*
  - *Svantaggi*
    - L'utente deve avere alcune *conoscenze informatiche* di base
    - Manca il *contatto diretto* con i responsabili della biblioteca
    - La *consultazione* di documenti digitali è più *faticosa* per l'utente, che deve ad esempio leggere dei testi da schermo
    - I documenti devono essere *acquisiti* per essere poi disponibili

## Acquisizione dei documenti

- Attualmente, la gran parte delle informazioni *non è direttamente disponibile* in formato digitale
  - Uno degli scopi di una biblioteca è la *conservazione* dei documenti prodotti dalla società
    - Fino a pochissimi anni fa, tutti i supporti erano di tipo *analogico*
- Per la creazione di una biblioteca digitale è necessario *acquisire* i documenti in *formato digitale*
  - L'acquisizione si basa, oltre al reperimento degli oggetti, su
    - L'utilizzo di particolari *periferiche di input*, che trasformano l'informazione del mondo reale in segnali adatti al computer
    - La *codifica* dell'informazione in sequenze finite di bit, con il conseguente utilizzo di *campionamento* e *quantizzazione*
    - In alternativa, l'acquisizione può essere ottenuta tramite l'*inserimento manuale* delle informazioni

## Acquisizione di documenti testuali - 1

- I documenti *testuali*, a meno che non siano stati prodotti da un word processor, sono in formato *cartaceo*
- Il modo più semplice per acquisirli è la *riproduzione manuale*
  - Consiste nel richiedere ad un operatore di *immettere* il testo nel calcolatore riscrivendolo con un *word processor*
    - Il procedimento è *lungo e costoso*
    - L'operatore può commettere *errori di battitura*, ed è quindi necessario *rileggere* completamente il documento
    - L'inserimento manuale si affida alla capacità umana di *interpretare correttamente* un testo scritto, anche se con caratteri a stampa di diverso tipo o scritto a mano
    - Documenti storici pregiati non corrono rischi di *danneggiamento*
    - L'operatore può riconoscere la *struttura organizzativa* dei documenti e riprodurla tramite il word processor

## Acquisizione di documenti testuali - 2

- L'*acquisizione automatica* viene eseguita in due fasi
  - *Digitalizzazione* dell'immagine del documento tramite una periferica di acquisizione, lo *scanner*
    - Le pagine che contengono il testo vengono trattate come delle *immagini* e acquisite una dopo l'altra
    - Il risultato è una sequenza di immagini, una per pagina, ad esempio di tipo *bitmap*
  - *Riconoscimento* del testo che compone l'immagine, ovvero dei singoli caratteri a stampa, tramite un programma di *Optical Character Recognition (OCR)*
    - Il programma cerca di *riconoscere* le diverse lettere dell'alfabeto e codificarle singolarmente
    - Il risultato è un documento finale in formato testuale, ad esempio *ASCII esteso* o *Unicode*

## Acquisizione di documenti testuali - 3

- Gli scanner possono essere di vario tipo, i più comuni sono
  - Del tipo *piatto*, in inglese *flatbed*, ovvero con uno schermo in vetro nel quale vanno appoggiate le pagine
  - Con *risoluzione* variabile, di solito di 400 o 600 dpi (dots per inch)
  - Con possibilità di scegliere il numero di sfumature di *colore*, di tonalità di *grigio* o di codificare direttamente in *bianco e nero*
- Gli OCR analizzano le immagini, interpretano i simboli grafici e assegnano ad ogni *simbolo* un *carattere alfabetico*
- L'acquisizione automatica:
  - E' *più veloce e meno costosa* dell'inserimento manuale
  - Necessita di *rilettura e correzioni* perché i programmi OCR effettuano diversi errori nel riconoscimento, specie di manoscritti
  - Porta alcuni rischi di *danneggiamento* degli originali

## Acquisizione audio - 1

- Le registrazioni audio analogiche sono presenti su *diversi formati*
  - Piano-roll perforati di pianoforti meccanici
  - Cilindri di cera per fonografi
  - Dischi 68, 33 o 45 giri
  - Nastri magnetici
- Tutti questi formati si *deperiscono* nel tempo, la loro acquisizione in formato digitale ha anche importanza per la *conservazione dei beni culturali*
  - Discorsi storici, interpretazioni di grandi musicisti, registrazioni di opere teatrali (anche radiofoniche)
- A differenza del testo, per l'audio l'importanza risiede anche nell'*interpretazione* di musicisti, cantanti o attori
  - La scelta del *formato digitale* da usare è cruciale, poiché ogni *campionamento e quantizzazione* introduce delle perdite



## Acquisizione audio - 2

- L'acquisizione viene di regola effettuata tramite riproduzione e digitalizzazione con una *scheda audio*
  - Nella scelta del tipo di campionamento e quantizzazione si tende a *privilegiare la qualità*, a discapito dello spazio
    - La tecnologia è in continua evoluzione, è quindi difficile scegliere i parametri ottimali
    - Normalmente si utilizza la *qualità CD*, ma spesso si cerca di andare oltre (campionamento più fitto, più bit per campione)
- L'audio viene acquisito per poterlo poi riprodurre, ma è possibile, anche se più difficile, effettuare un analogo dell'OCR
  - Nel caso del *parlato* è possibile effettuare delle *trascrizioni* automatiche, utilizzando tecniche di speech recognition
    - Allo stato dell'arte è però *impossibile estrarre la partitura* da brani di musica polifonica, ma è fattibile per brani monofonici

## Acquisizione di immagini - 1

- Se l'audio viene acquisito per preservare e conservare, le immagini vengono acquisite per scopi *documentaristici*
  - Le immagini in formato digitale sono normalmente *troppo diverse* dagli originali
    - I quadri hanno una terza dimensione data dalla stratificazione della successione di pennellate
    - Le dimensioni devono di solito venire notevolmente ridotte per poter essere rappresentate su di uno schermo di computer
    - La presentazione a schermo anche di fotografie è percettivamente molto diversa da quella su carta o, a maggior ragione, su tela
- L'acquisizione ha in ogni caso una *notevole importanza* per storici dell'arte, studiosi e appassionati in genere
  - Il mezzo informatico consente di analizzare opere d'arte in modo più approfondito che non la riproduzione fotografica

## Acquisizione di immagini - 2

- Le immagini necessitano di diversi *livelli di risoluzione*
  - L'immagine nel suo complesso deve essere *rappresentabile* sullo *schermo* del computer
    - A questo livello la risoluzione richiesta non è molto elevata
  - Anche i *dettagli importanti* di un'opera devono essere adeguatamente *rappresentati*
    - La risoluzione dipende dal livello di dettaglio che si vuole rappresentare
- L'acquisizione viene effettuata attraverso *fotocamere digitali*
  - Prendendo *diverse foto*, di tutto l'insieme e dei particolari
    - Processo laborioso e lento
  - Prendendo un'*unica immagine* alla massima risoluzione
    - Molto dispendioso in termini di risorse di memoria e di calcolo

## Organizzazione e catalogazione

- I documenti acquisiti sono memorizzati in forma di *file*
  - Nel file-system i file sono organizzati in base al loro *nome* e alla loro *posizione* nell'albero delle directory
    - E' però molto complicato *riaccedere* ai documenti
- Al momento dell'acquisizione i documenti vengono *catalogati* e *organizzati* da operatori esperti
  - Ai documenti vengono associati dei *descrittori testuali*
    - Titolo, autore e anno di uscita dell'opera
    - Descrizione con parole chiave prese da un vocabolario controllato
    - Descrizione estesa del contenuto del documento
    - Collegamenti ipertestuali con altri documenti correlati
- Inoltre, vengono utilizzati strumenti per l'indicizzazione automatica dei documenti

## Accesso ad una biblioteca digitale - 1

- Una volta creata, organizzata e classificata la collezione di documenti, è necessario fornire all'utente dei *metodi per accedere* alla biblioteca digitale
- L'obiettivo principale di una biblioteca digitale è consentire agli utenti di *selezionare gli argomenti di interesse*
  - Un utente consulta una biblioteca digitale poiché ha una determinata *esigenza informativa*
  - L'esigenza informativa può venire soddisfatta da un *sottoinsieme dei documenti* della biblioteca digitale, che sono *pertinenti* allo specifico argomento di interesse
  - Il sistema deve *individuare* questo sottoinsieme di documenti pertinenti, o rilevanti, e presentarlo all'utente
- Ci si riferisce a questa attività con il termine di *recupero* (o ricerca o reperimento) *di informazioni in linea*

## Accesso ad una biblioteca digitale - 2

- L'accesso ai documenti ha dei punti in comune con il recupero di informazioni nel Web, con alcune importanti differenze
  - La biblioteca digitale tratta, in maniera approfondita, alcuni *temi specifici*
  - La biblioteca digitale è *mantenuta* da un'organizzazione
    - I documenti sono *selezionati*, il loro contenuto è in genere *affidabile* e il formato è *coerente* tra i diversi documenti
    - E' *noto a priori* il numero e il contenuto dei documenti
  - All'estrazione automatica del contenuto è affiancata l'opera di *personale qualificato*
  - La creazione di una biblioteca digitale è frutto di uno *studio approfondito* delle tematiche trattate e delle tipologie di utente
- I documenti Web quindi *non costituiscono una biblioteca digitale*

## Accesso ad una biblioteca digitale - 3

- Il Web consente però di accedere alle biblioteche digitali, che sono *disponibili in Internet*
  - La tecnologia Web consente di utilizzare i browser e il protocollo di comunicazione HTTP per
    - *Collegarsi* al sito di una biblioteca digitale
    - *Interrogare* il sistema per reperire documenti pertinenti alle proprie esigenze informative
    - *Visionare* i documenti recuperati
- Dato il costo per la realizzazione di una biblioteca digitale, l'accesso può essere *a pagamento*
  - Gli utenti devono *registrarsi* prima di poter accedere ai documenti
  - Il controllo degli accessi viene fatto con un sistema di *login e password*, come per molti altri servizi nel Web
  - Si può accedere gratuitamente ad un *insieme ridotto* di funzioni