AA 2004-2005 - Università di Padova
Facoltà di Ingegneria - Corso di Laurea in Ingegneria Informatica
Insegnamento di “Basi di dati”

La gestione di basi di dati testuali

Maristella Agosti

Dipartimento di Ingegneria dell’Informazione

Gruppo di ricerca di Sistemi di gestione delle informazioni (IMS)

Indice dei contenuti

- caratteristiche distintive dei dati testuali/multimediali rispetto ai dati strutturati
- tipi di sistemi di gestione dei dati e dell’informazione
- il processo di recupero dell’informazione
- indicizzazione e indicizzazione automatica
- modelli di recupero dell’informazione
- funzioni di pesatura dei termini
- valutazione del processo di recupero
- interrogazione di un IRS attraverso un *browser* Web
- il motore di ricerca (*search engine*) e le sue componenti
- cenni sui sistemi di gestione di biblioteche digitali.

Basi di dati testuali e multimediali

Caratteristiche di gestione dati delle basi di dati:

- capacità di gestire dati che possono assumere valori elementari, quali numeri o stringhe, cioè dati strutturati;
- i dati gestiti possono essere recuperati in base a condizioni formulate sui valori dei componenti.

Quando i dati da gestire non sono riconducibili a dati numerici o stringhe, perché sono testuali, immagini, audio e/o video, spesso si utilizza un sistema specializzato; ad es, per i dati testuali, un sistema di reperimento informazioni (*Information Retrieval System: IRS*).

Molti DBMS prevedono la possibilità di gestire dati non strutturati unicamente come dati di grandi dimensioni “non interpretati”, spesso chiamati *BLOB: Binary Large OBject*.

Caratteristiche dei dati testuali e multimediali che li differenziano dai dati strutturati (1/2)

- sono generalmente dati di grandi dimensioni; i DBMS sono stati specializzati per la gestione di grandi quantità di dati di piccole dimensioni;
- rappresentano informazioni prive di una struttura descritta separatamente;
- interessa ritrovarli in base alla loro semantica, cioè in base a cosa rappresentano (ricerca di testi che trattano un certo argomento, ricerca di immagini che contengono uno specifico oggetto/particolare);
- il risultato di una ricerca sono oggetti (testi, immagini, …) che soddisfano almeno in parte l'esigenza informativa dell'utente che ha formulato la sua richiesta (*query*), quindi vanno presentati per ordine di rilevanza alla query;

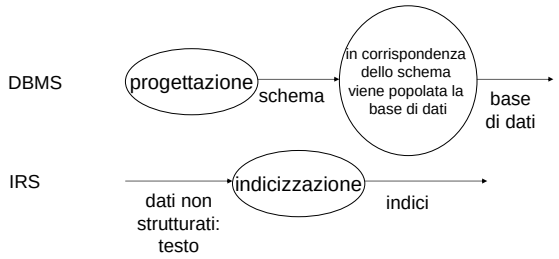
Caratteristiche dei dati testuali e multimediali che li differenziano dai dati strutturati (2/2)

- una ricerca non si esaurisce con un'unica interazione, spesso si svolge con un ciclo del tipo:
  - formulazione richiesta,
  - recupero di un insieme di dati rilevanti,
  - analisi dei dati recuperati,
  - formulazione di una nuova richiesta;
- quindi il processo di recupero dell'informazione si sviluppa in un ciclo iterativo; è di fondamentale importanza l'interazione utente-sistema (importante: interfaccia utente);
- l'immissione di questi dati richiede modalità e strumenti più complessi, perché questi dati spesso devono essere pre-elaborati (ad es. digitalizzazione).

Problemi che devono essere risolti per la gestione dei dati testuali e multimediali

- come rappresentare questi dati nel modello dei dati; esiste un modello dei dati?
- quali operatori rendere disponibili su di loro per formulare le richieste;
- come estendere il sistema con nuove strutture dati per agevolarne il recupero;
- come gestire in modo efficiente l'immissione e la visualizzazione.

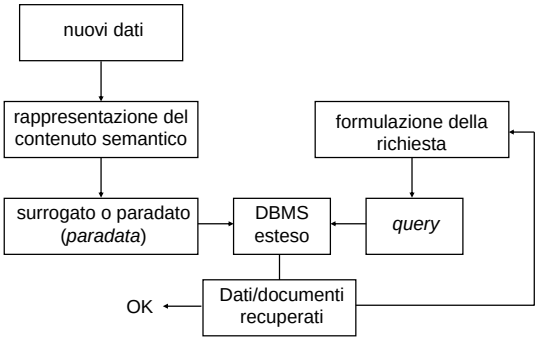
Tipi di sistemi di gestione dell'informazione



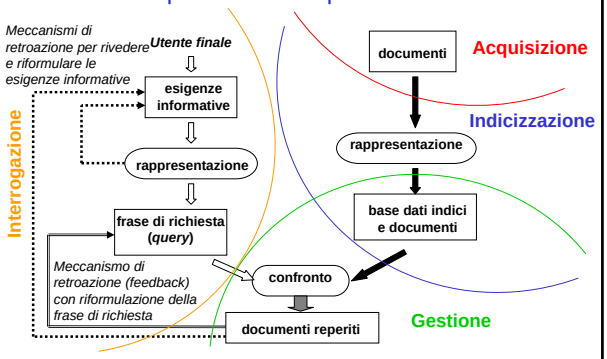
Tipi di sistemi di gestione dell'informazione

	IRS	DBMS
dati	non strutturati	strutturati
interrogazioni	libere	con uso dello schema
navigazione	poca	nessuna
collezione	grande	grande
recupero	probabilistico	deterministico
interattività	media	bassa

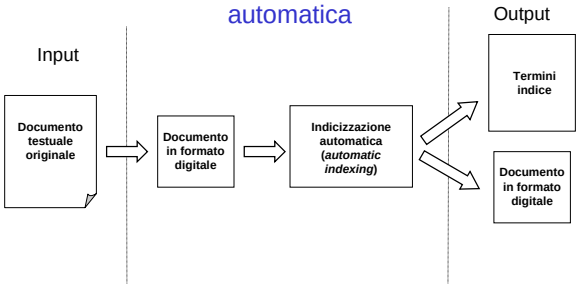
Attività di gestione dei dati testuali e multimediali



Schema del processo di reperimento dati testuali



Schema del processo di indicizzazione automatica



Indicizzazione

- Finalità: preparazione di un "surrogato" del contenuto del documento attraverso l'assegnazione al documento di descrittori che descrivono il suo contenuto semantico
- Tipologie: manuale, semi-automatica o automatica
- Modalità: estrazione dal documento o utilizzo di fonti esterne (es. dizionari)

un caso importante:

indicizzazione del testo ed estrazione automatica di parole (ad esempio: parole chiave)

Indicizzazione automatica di documenti testuali

L'indicizzazione automatica (*automatic indexing*) di un documento testuale è il processo che esamina automaticamente gli oggetti informativi che compongono il documento e, utilizzando appositi algoritmi, produce una lista di

termini indice (*index terms*)

che può essere utilizzata per la rappresentazione del contenuto informativo del documento di partenza.

I termini indice sono utilizzati come surrogati per la rappresentazione del documento originale, quindi, in sostituzione del documento originale.

Fasi del processo di indicizzazione automatica

Fasi del processo di indicizzazione automatica del testo ed estrazione automatica di termini indici (*index terms*) o parole chiave (key words); le fasi del processo devono essere attuate in sequenza:

1. analisi lessicale o selezione delle parole
2. rimozione delle *stop-word* o parole molto comuni (sinonimi: *stop list*, *negative dictionary*)
3. *prima parte:* estrazione delle radici (*stemming*), cioè riduzione delle parole originali alle radici
4. *seconda parte:* creazione dell'indice
5. pesatura.

Analisi lessicale e rimozione delle parole molto comuni

Analisi lessicale

Con analisi lessicale si identifica il processo di trasformazione di un flusso di caratteri di input in un flusso di parole o *tokens*, dove con *token* si intende identificare un gruppo di caratteri portatore di uno specifico significato.

Rimozione delle parole molto comuni (*stop words*, *stop list*, *negative dictionary*)

Le parole presenti con frequenza alta nel documento di partenza servono poco ad identificare il significato del documento stesso, quindi vanno eliminate e non considerate oltre.

Per la lingua inglese è stata preparata nel tempo una lista di circa 250 parole che sono considerate da tutti *stop words*.

Analisi lessicale e/o selezione delle parole

Documento	Contenuto/testo del documento
D_1	"shipment of gold damaged in a fire"
D_2	"delivery of silver arrived in a silver truck"
D_3	"shipment of gold arrived in a truck"

Esempio tratto da: D.A. Grossman, O. Frieder. *Information Retrieval: algorithms and heuristics*. Kluwer Academic, 1998.

Rimozione delle *stop-words* o delle parole molto comuni

Lista o dizionario delle *stop-words* = {of, in, a}

Documento	Parole che vengono considerate
D_1	shipment gold damaged fire
D_2	delivery silver arrived silver truck
D_3	shipment gold arrived truck

Rimozione delle *stop-words* o delle parole molto comuni

Lista o dizionario delle *stop-words* = {of, in, a}

Doc	Parole che vengono considerate							
D_1	-	damaged	-	fire	gold	shipment	-	-
D_2	arrived	-	delivery	-	-	-	silver(2)	truck
D_3	arrived	-	-	-	gold	shipment	-	truck

In totale vi sono 8 parole diverse.

Riduzione delle parole originali alle radici o estrazione delle radici (*stemming*)

Questa fase della indicizzazione rimuove dalle parole la parte finale riducendo tutte le parole affini che hanno in comune la stessa radice ad un'unica "radice".

Doc	Radici
D_1	shipment gold damag fir
D_2	deliveri silver arriv silver truck
D_3	shipment gold arriv truck

Riferimento bibliografico per l'algoritmo di *stemming* più utilizzato per la lingua inglese:
M.F. Porter. An algorithm for suffix stripping. *Program*, 14 (3), 1980, 130-137.

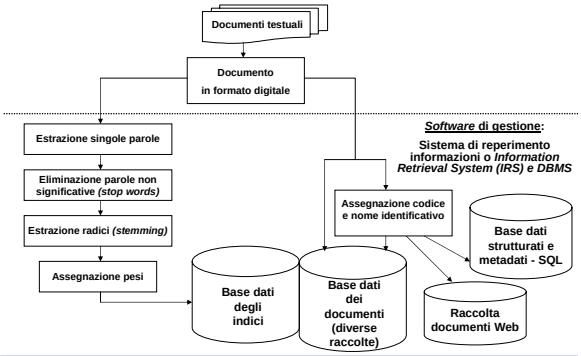
Creazione dell'indice

shipment	D_1		D_3
gold	D_1		D_3
damag	D_1		
fir	D_1		
deliveri		D_2	
silver		D_2 D_2	
arriv		D_2	D_3
truck		D_2	D_3

Pesatura

Documenti / Radici	D_1	D_2	D_3
arriv	0	1	1
damag	1	0	0
deliveri	0	1	0
fir	1	0	0
gold	1	0	1
shipment	1	0	1
silver	0	2	0
truck	0	1	1

Rappresentazione e indicizzazione



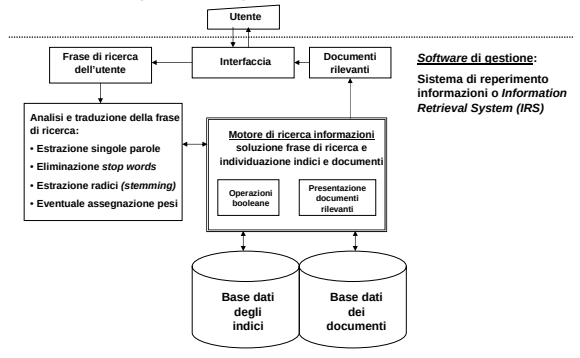
Recupero

- selezione dei documenti che sono verosimilmente rilevanti all'interrogazione
- dipende dal modello di recupero (attenzione, diverso dal modello di BD)
- l'utente gioca un ruolo determinante
- combinazione di diverse strategie interattive, ad esempio: interrogazione e navigazione

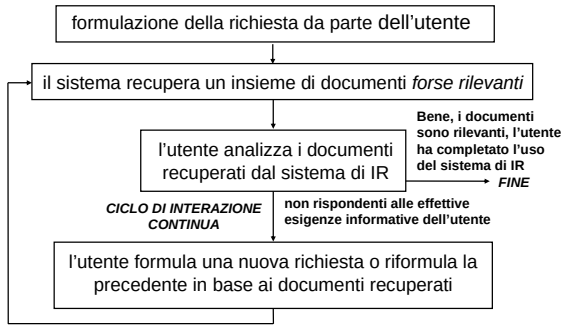
Possibilità:

- *exact match:* la collezione è divisa in due: i documenti che soddisfano l'interrogazione, e gli altri (*booleano*, basi di dati)
- *best match:* i documenti sono ordinati utilizzando una misura di "similarità" e sono proposti quelli sopra una soglia (vettoriale, probabilistico)

Recupero o reperimento informazioni



Ciclo di interazione utente-sistema



Modelli di recupero o modelli di *information retrieval*

- costrutti per indicizzare e recuperare documenti
- i modelli si distinguono per la funzione di recupero

Alcuni dei modelli più utilizzati e/o diffusi:

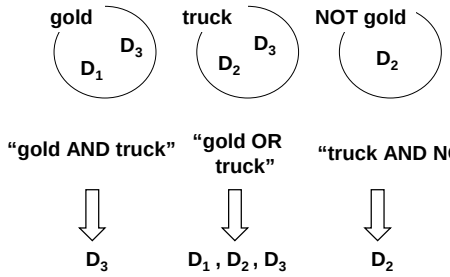
- booleano (*exact match*)
- vettoriale (*best match*)
- probabilistico (*best match*)

Modello booleano

Funzione di recupero:

- i termini dell'interrogazione sono collegati mediante operatori booleani (AND, OR, NOT)
- è possibile talvolta utilizzare operatori di prossimità o di troncamento dei termini

Modello booleano: esempio



"gold AND truck" → D_3

"gold OR truck" → D_1, D_2, D_3

"truck AND NOT gold" → D_2

Modello booleano

- è efficace in ambienti controllati e con utenti bene addestrati
- l'utente sa cosa chiede
- richiede l'addestramento dell'utente
- ha delle limitazioni dovute alla bassa "amichevolezza" della logica booleana
  - es. "information AND retrieval" oppure "information OR retrieval"
- poco controllo sulla dimensione dell'insieme dei documenti recuperati
- non è possibile l'ordinamento per una qualche misura di "similarità"
- non è possibile la pesatura dei termini.

Pesatura automatica dei termini indice

I termini indice possono essere assegnati ai documenti della collezione dei documenti di interesse sia senza che con un peso.

Quando il processo di indicizzazione è binario, ad un termine che viene assegnato ad un documento viene assegnato un peso implicito uguale a 1; questo, allora, avviene perché non si tiene conto della effettiva frequenza del termine nel documento.

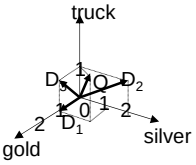
Quando il processo di indicizzazione associa ai termini un peso, vuol dire che si intende tenere conto nel processo in modo esplicito dell'importanza del termine nel documento; per assegnare il peso al termine si deve utilizzare una qualche funzione di pesatura. Molte e diverse funzioni sono state proposte negli anni, di seguito ne verrà presentata una molto utilizzata, perché ha fornito validi risultati sperimentali.

Modello vettoriale

- documenti e interrogazioni sono vettori in uno spazio multi-dimensionale
- la dimensione dello spazio vettoriale è il numero di termini
- la misura di "similarità" tra un documento e un'interrogazione è il coseno dell'angolo tra i due vettori

Esempio:

Q = "gold silver truck"



Modello vettoriale: schemi di pesatura

- lo schema *tf×IDF* e le sue variazioni si sono rivelate tra le più efficaci
- *tf* (*term frequency*)
  - la frequenza del termine nel documento
- *IDF* (*Inverse Document Frequency*)
  - è il logaritmo del reciproco del rapporto tra il numero di documenti e il numero di documenti indicizzati dal termine
- si basa su due assunzioni:
  - giacché l'insieme dei documenti rilevanti è piccolo, è probabile che i termini con *IDF* alto appaiano in documenti rilevanti,
  - quanto più un documento è rilevante, tanto maggiore è *tf* e, quindi, il peso del termine.

Schemi di pesatura: esempio con *tf×IDF*

	D_1	D_2	D_3
shipment	1×log(3/2)	0	1×log(3/2)
gold	1×log(3/2)	0	1×log(3/2)
damag	1×log(3)	0	0
fir	1×log(3)	0	0
deliveri	0	1×log(3)	0
silver	0	2×log(3)	0
arriv	0	1×log(3/2)	1×log(3/2)
truck	0	1×log(3/2)	1×log(3/2)

La valutazione

- la sperimentazione ha caratterizzato la ricerca in IR fin dagli inizi (anni '50)
- metodologie di sperimentazione
- le collezioni sperimentali
- misure di efficacia
- TREC (*Text REtrieval Conference*)
- CLEF (*Cross-Language Evaluation Forum*)
- valutazione di sistemi interattivi
- il problema della "rilevanza".

Il problema della "rilevanza"

- è praticamente impossibile conoscere la rilevanza di milioni di documenti
- il giudizio su un documento influisce il giudizio su quelli visti successivamente
- la rilevanza sullo stesso documento dipende dall'utente e dall'istante di tempo.

Metodologie di sperimentazione

In ambiente operativo:

- banche dati operative
- utenti reali

In ambiente di laboratorio:

- collezioni sperimentali
- necessità di costruirle

Le collezioni sperimentali

Una collezione sperimentale è una tripla (*D*,*Q*,*R*) dove:

- *D* è una raccolta di documenti
- *Q* è una raccolta di interrogazioni estratte dai *log-file* di sistemi di IR effettivamente operativi oppure costruite appositamente
- *R* è la raccolta delle corrispondenze tra ogni interrogazione e i documenti giudicati rilevanti ad essa; questa raccolta di risultati alle interrogazioni è costruita analizzando effettivamente la raccolta di documenti *D*.

Valutazione del processo di reperimento

La presenza di un certo numero di dati non rilevanti fra quelli recuperati è considerato come un effetto rumore, mentre il non recuperare dati rilevanti viene chiamato effetto silenzio.

Ambedue sono effetti negativi che devono essere limitati.

Si definiscono, allora, alcune misure per valutare il comportamento di un sistema di reperimento di informazioni.

Dato un insieme di dati e una frase di richiesta (*query*), è possibile individuare quattro sotto-insiemi:

A: insieme dei dati correttamente recuperati in quanto rilevanti alla query;

B: insieme dei dati recuperati anche se non rilevanti;

C: insieme dei dati non rilevanti e giustamente omessi;

D: insieme dei dati non recuperati ma rilevanti.

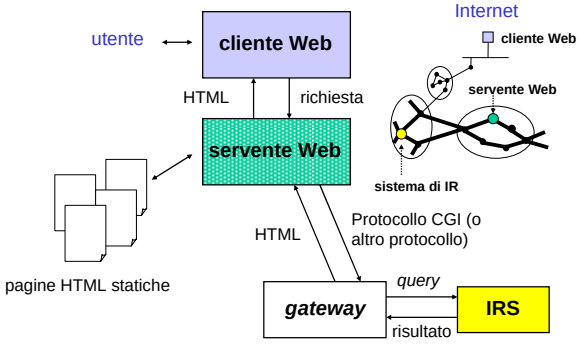
Richiamo e precisione

	dati rilevanti	dati non rilevanti
dati recuperati	***A (corretti)***	***B (inesatti)***
dati non recuperati	***D (omessi)***	***C (da omettere)***

Richiamo:
è il rapporto fra il numero dei dati rilevanti recuperati $|A|$ e il totale dei dati rilevanti che sono al momento gestiti dal sistema: $|A| + |D|$.

Precisione:
è il rapporto fra il numero dei dati rilevanti recuperati $|A|$ e il totale dei dati recuperati: $|A| + |B|$.

Accesso ad un IRS via Web



"Motori di ricerca": strumenti per il recupero di informazioni presenti in documenti Web

- *Search Engine-SE* (motore di ricerca): è uno strumento software dedicato alla individuazione e all'indicizzazione di pagine Web presenti in un sotto-insieme di interesse dei siti Web. Quando un utente utilizza un SE ottiene in risposta un elenco di link a pagine Web che il SE ha valutato come rilevanti alla domanda dell'utente. Gli indici che utilizza un SE sono generati e mantenuti in modo autonomo e automatico.
  Esempi di SE: AltaVista, Google, Lycos.
- *Directory*: sono degli strumenti che gestiscono solo pagine che sono state sottoposte da utenti/utilizzatori o scelte attraverso un processo di selezione/catalogazione.
  Esempio: Yahoo.

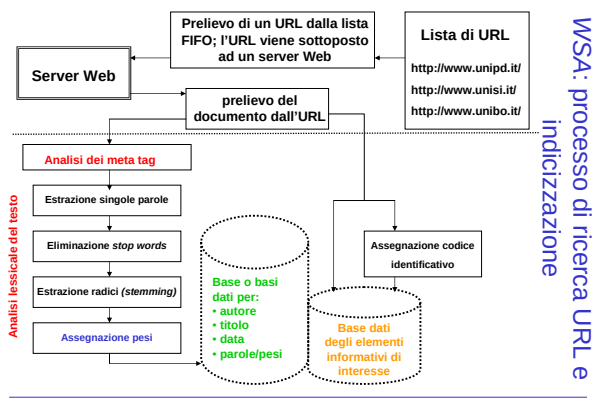
Componenti software di un *Search Engine-SE*

Web Search Agent (WSA) o agente di ricerca

- E' un programma che attraversa la struttura ipertestuale del Web in modo autonomo, automatico e ricorsivo recuperando informazioni dai documenti localizzati; ai documenti localizzati viene poi applicata una funzione che li trasforma o che ne ricava particolari dati. Alcune denominazioni di WSA: *spider, worm, crawler, wanderer*.

Motore di ricerca propriamente detto

- E' un sistema di recupero dell'informazione che permette all'utente di effettuare ricerche, mediante indici, nella base di dati di link ai documenti Web indicizzati dal SE; questi documenti Web probabilmente contengono informazioni di interesse per l'utente che ha formulato la richiesta.



WSA: processo di ricerca URL e indicizzazione

Dai documenti alla biblioteca digitale

La biblioteca digitale è una collezione organizzata di documenti digitali.

Biblioteca tradizionale

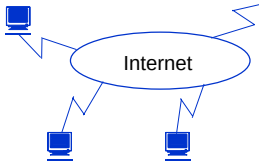
- Collezione di documenti
- Ricerca mediante cataloghi e indici
- Localizzazione e identificazione dei documenti

Biblioteca digitale

Biblioteca digitale: accesso via rete

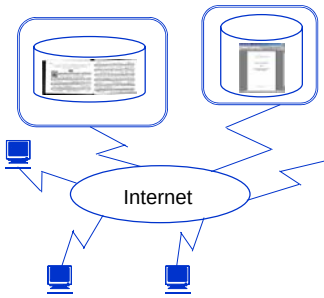
Ricerca documenti

utilizzando i cataloghi e gli indici da qualunque postazione connessa a Internet

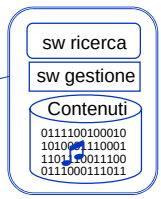


Il sistema della biblioteca digitale

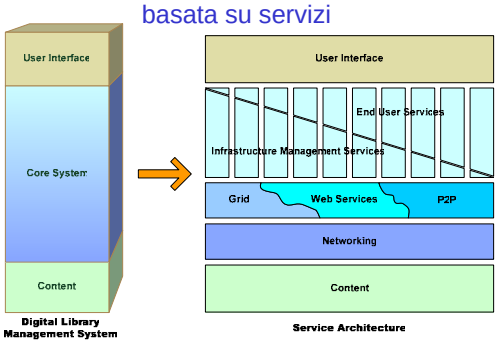
Accesso e biblioteca distribuiti



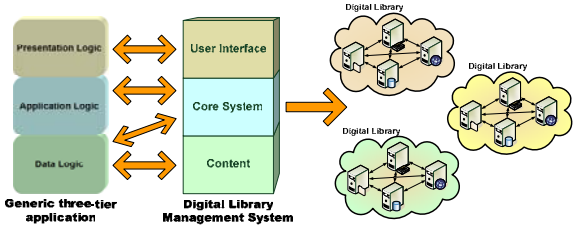
L'utente vede un'unica collezione virtuale



Da “un” sistema unico ad una architettura basata su servizi



Verso i nuovi sistemi di gestione di biblioteche e archivi digitali (DLAMS)



Alcuni riferimenti utili

- M. Agosti, F. Crestani, G. Pasi (Eds). *Lectures on Information Retrieval: Third European Summer-School, ESSIR 2000 Varenna, Italy, September 11-15, 2000*. Revised Lectures, Springer-Verlag, Berlin/Heidelberg, 2001.
- W. Frakes, R. Baeza-Yates. *Information Retrieval: data structures and algorithms*. Prentice Hall, 1992.
- D.A. Grossman, O. Frieder. *Information Retrieval: algorithms and heuristics*. Kluwer Academic, 1998.
- M.F. Porter. An algorithm for suffix stripping. *Program*, 14 (3), 1980, 130-137.
- G. Salton, M.J. McGill, *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.
- K. Sparck-Jones, P. Willett. *Readings in Information Retrieval*. Morgan Kaufmann, 1997.
- C.J. “Keith” van Rijsbergen *Information Retrieval (2nd Ed)*. Butterworths,1979. Disponibile anche in forma digitale all’URL: http://www.dcs.gla.ac.uk/Keith/Preface.html.
- I.H. Witten, D. Bainbridge. *How to Build a Digital Library*. Morgan Kaufmann, San Francisco, CA, USA, 2003.