Accesso ad archivi sonori

Nicola Orio

Dipartimento di Ingegneria dell'Informazione

IV Scuola estiva AISV, 8-12 settembre 2008

Basi di dati
Biblioteche e archivi digitali

Sistemi informativi e basi di dati

- Ogni organizzazione ha bisogno di memorizzare e mantenere informazioni specifiche. Per esempio:
  - Conti correnti bancari
  - Studenti iscritti a un corso di laurea
  - Quotazioni di azioni nei mercati telematici
  - Archivi e biblioteche digitali

- I sistemi informativi organizzano e gestiscono le informazioni necessarie alle attività di un'organizzazione:
  - Inizialmente non automatizzati
  - Informatica → gestione automatica delle informazioni
    - → **basi di dati**
  - Informazione memorizzata e organizzata in modo rigoroso

Dato e informazione

- **Dato** = elemento di informazione, che di per sé non ha interpretazione, poiché privo di un contesto
  - Mario Rossi → *nome e cognome*
  - 10150 → *numero matricola? Numero di abitanti di una città? CAP? Numero di telefono?*

- **Informazione** = dato + interpretazione
  - Domanda: “Chi è il responsabile dell’archivio e qual è il suo numero di telefono?” → *interpretazione* della risposta
  - Risposta: Mario Rossi, 10150 → *dato*
  - Domanda + risposta: informazione

- Nei sistemi informatici, le informazioni vengono rappresentate per mezzo di *dati* : necessità di un *contesto*

Dati e applicazioni

- I dati possono variare nel tempo
  - Le modalità con cui i dati sono rappresentati in un sistema sono di solito stabili
  - Le operazioni sui dati variano spesso
- Obiettivo:
  - Separare i dati dalle applicazioni che operano su essi

- Le **basi di dati** sono una collezione di dati per rappresentare informazioni di interesse
  - Caratteristiche
    - di *grandi dimensioni* → molti dati contemporaneamente
    - *condivise* → accessi concorrenti da parte di molti utenti
    - *persistenti* → il contenuto viene mantenuto nel tempo, anche nel caso di problemi hardware e software

Basi di Dati (DB) e DBMS

- **DBMS** = Data Base Management System
  - Sistema per la gestione di basi di dati

- Caratteristiche principali di un DBMS
  - **Affidabilità** = protezione dei dati, in caso di guasto HW o SW capacità di ripristinare i dati (almeno parzialmente)
  - **Sicurezza/privatezza** = abilitazioni diverse a seconda dell'utente
  - **Efficienza** = tempi di risposta e occupazione spazio accettabili (dipende dalla tecnica di memorizzazione dei dati)
  - **Efficacia** = facilitare l'attività di organizzazione

Pro e contro dei DBMS

- **Vantaggi**
  - I dati diventano una risorsa di un'organizzazione
    - Comune per utenti e applicazioni
  - Offrono un modello formale della realtà di interesse
    - Preciso, riutilizzabile
  - Consentono un controllo centralizzato dei dati
    - Riduzione di ridondanze e inconsistenze
  - Garantiscono l'indipendenza dei dati
    - Sviluppo di applicazioni flessibili e modificabili
- **Svantaggi**
  - Complessi, costosi, necessitano specifici SW e HW
  - Difficile separare i servizi utili da quelli inutili
  - Inadatti alla gestione di poche informazioni per pochi utenti

Utenti di una base di dati

- Si prevedono di solito alcune tipologie di utenti
  - *Progettista*
  - *Amministratore*
  - *Programmatore* di applicazioni
  - *Utente esperto*: usa la base di dati per il proprio lavoro, conosce procedure di interazione
  - *Utente generico*: consulta la base di dati saltuariamente

- Nelle piccole basi di dati queste figure spesso coincidono
  - In molti casi non c'è un vero progettista
    - Rischio di non rappresentare correttamente la realtà di interesse (minimondo)
    - Problemi di gestione nel lungo periodo

Il modello relazionale

- Si basa sul **concetto matematico di relazione**
  - Naturale rappresentazione per mezzo di tabelle
  - Una tabella è un elemento del database che può rappresentare
    - Una delle entità in gioco
    - Un legame (spesso chiamato anch‘esso relazione) tra due o più entità

- Una relazione matematica è un insieme di ennuple ordinate:
  - una relazione è un insieme → non c'è ordinamento fra le ennuple di una tabella
  - le ennuple sono distinte
  - ciascuna ennupla è ordinata → l’ i-esimo valore proviene dall’ i-esimo dominio

Tabelle e relazioni

- Una tabella **rappresenta** una relazione se
  - I valori di ogni colonna sono fra loro *omogenei*
  - Le righe sono *diverse* fra loro
  - Le *intestazioni* delle colonne sono diverse tra loro
- In una tabella che rappresenta una relazione
  - L'ordinamento tra le righe è irrilevante
  - L'ordinamento tra le colonne è irrilevante
- Ogni colonna è associata ad un particolare attributo della relazione
  - Le colonne costituiscono le *descrizioni* delle caratteristiche degli oggetti rappresentati
  - Le celle di una tabella rappresentano i *dati*
    - L'intestazione della colonna *contestualizza* i dati

Il concetto di dominio

- Per ogni attributo, il **dominio** è l'*insieme dei valori ammessi* dagli elementi da rappresentare
  - E' necessario uno studio approfondito sui domini dei dati
- I riferimenti fra dati in relazioni diverse sono rappresentati per mezzo di valori dei domini che compaiono nelle ennuple
- Vantaggi del modello basato sui valori
  - Indipendenza dalle strutture fisiche che possono cambiare dinamicamente
  - Si rappresenta solo ciò che è rilevante dal punto di vista dell'applicazione
  - L'utente finale vede gli stessi dati dei programmatori
  - I dati sono portabili più facilmente da un sistema ad un altro

Informazione incompleta

- Il modello relazionale impone ai dati una struttura rigida:
  - Solo alcuni formati di ennuple sono ammessi: quelli che corrispondono agli schemi di relazione
  - I dati possono non corrispondere al formato previsto
  - In particolare alcune informazioni possono *non essere presenti*, o *non avere senso* per alcuni oggetti

- Come rappresentare questa situazione?
  - Non conviene (ma spesso si fa) usare valori del dominio
    - Ad esempio 0, stringa nulla, 99, “ZZZ”
  - Tecnica rudimentale ma efficace:
    - **valore nullo**: denota l’assenza di un valore del dominio (e non è un valore del dominio)

Il concetto di chiave

- Ogni relazione (tabella) deve avere una **chiave**

- Definizione informale:
  - Insieme di attributi che identificano univocamente ogni singola ennupla di una relazione (riga di una tabella)

- Definizione formale
  - Un insieme K di attributi è *chiave* per una relazione se
    1) Non contiene due ennuple distinte
    2) Se togliamo un attributo da K, si possono avere ennuple uguali
- L'esistenza delle chiavi garantisce l'accessibilità, senza rischi di "confusione", a ciascun dato della base di dati
  - Si possono fare riferimenti incrociati a precise ennuple

Differenze con i fogli elettronici (Excel)

- Di una base di dati deve essere progettata prima la struttura e poi inseriti i dati
  - Una base di dati deve essere sempre consistente
  - Difficile fare delle operazioni di modifica della struttura una volta iniziato a inserire i dati
- Linguaggio standard di interrogazione per i database
  - Con un foglio elettronico non abbiamo la stessa di collegare più tabelle (se non con operazioni sui valori delle celle)
- Possibilità di informazione ripetuta
  - Lo stesso dato può essere inserito in diverse parti (ridondanza)
    - I valori possono essere inconsistenti
- Le modifiche ai valori dei dati sono propagate per mantenere la consistenza

Informazione non strutturata e multimediale

- Le basi di dati nascono per rappresentare informazione strutturata sotto forma di numeri, stringhe, date, valute
  - Nei DBMS più diffusi, tutto ciò che è al di fuori viene rappresentato come un BLOB
  - Binary Large Object
    - Non ci sono funzioni specifiche per gestire i BLOB
    - Spesso i BLOB sono su file esterni non gestiti dal DBMS
- Cosa finisce nei BLOB?
  - Documenti full text
  - Audio, immagini, video
- La ricerca si occupa di progettare strumenti per la gestione dell'informazione nei BLOB

Biblioteche e archivi digitali – 1

- Molte iniziative di associazioni, enti e centri di ricerca riguardano il settore delle digital libraries
- Definizione
  - Una biblioteca (archivio) digitale è una *collezione organizzata* di documenti e informazioni in formato digitale
- In generale si assume che una biblioteca (archivio) digitale abbia tre caratteristiche principali
  - Una collezione
  - Un mandato per il mantenimento e la diffusione delle informazioni
  - Una funzione e una serie di strumenti per la mediazione con l'utenza, sia generica che specialistica

Biblioteche e archivi digitali – 2

- *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

Information retrieval

Information retrieval – 1

- La *disponibilità* di informazioni, anche in formato digitale, non implica che gli utenti possano accedervi facilmente
  - 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
  - Problema classico di biblioteche e archivi
  - Esploso con la diffusione del Web
    - La soluzione è la creazione di cataloghi (indici) e applicare tecniche di *information retrieval*

Information retrieval – 2

- La catalogazione *descrive*, 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
- L'information retrieval è nato per trattare documenti testuali
  - L'informazione è contenuta nella *semantica delle parole* che compongono i documenti
  - L'estrazione delle parole da un documento è un'operazione abbastanza semplice per le lingue basate su di un alfabeto
    - Lavorare con documenti sonori è molto più complicato

Indicizzazione

- Consente di *descrivere il contenuto semantico* dei documenti
  - Normalmente si parla di documenti in senso lato, includendo anche media diversi dal testo
- I documenti sono rappresentati da *descrittori*, chiamati *indici*, che possono essere:
    - I termini che compongono un documento testuale
    - Gli spunti tematici, le successioni di accordi, la timbrica, le figurazioni ritmiche per i documenti musicali
- L'indicizzazione è svolta estraendo in modo automatico l'informazione *direttamente dal documento*
    - Possono essere utilizzate *altre fonti*, come dizionari o *metainformazioni* o essere fatta manualmente

Perché si indicizzano i documenti?

- 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

- Una volta indicizzati i documenti è possibile effettuare delle ricerche nei *soli indici* dei documenti
  - La ricerca negli indici è meno *onerosa computazionalmente*
  - Sono state sviluppate tecniche ad hoc per *accedere in modo efficiente* agli indici e velocizzare i tempi di ricerca

- L'operazione di indicizzazione è normalmente molto onerosa
  - E' fatta incrementalmente, nel caso di nuove acquisizione, e *prima* che gli utenti interroghino il sistema

Interrogazioni

- Gli utenti interagiscono con un sistema di IR formulando delle interrogazioni (query)

- In IR una query è una rappresentazione approssimata dell'*esigenza informativa* di un utente
  - **Nota**: Nel mondo dei database, con il termine query si intende un'esatta descrizione di una funzione nel dominio degli attributi e delle tabelle di un database

- L'utente può descrivere la propria esigenza informativa
  - Descrivendo le caratteristiche dei documenti potenzialmente interessanti, ad esempio usando *metadati*
  - Fornendo (estratti di) documenti simili a quelli cercati, secondo il paradigma *query-by-example*

Tipologie di utenti

- 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*
  - *Utente casuale*: *non conosce* esattamente cosa sta cercando, formula interrogazioni *generiche*, *si affida* al sistema di IR
- L'IR nasce per servire utenti esperti (bibliotecari)
  - I modelli semplici, interfacce complesse, pochi parametri noti e modificabili dagli utenti
- La diffusione dei Web search engine ha invertito la tendenza
  - Modelli complessi, interfacce semplici, molti parametri nascosti e impostati dal sistema

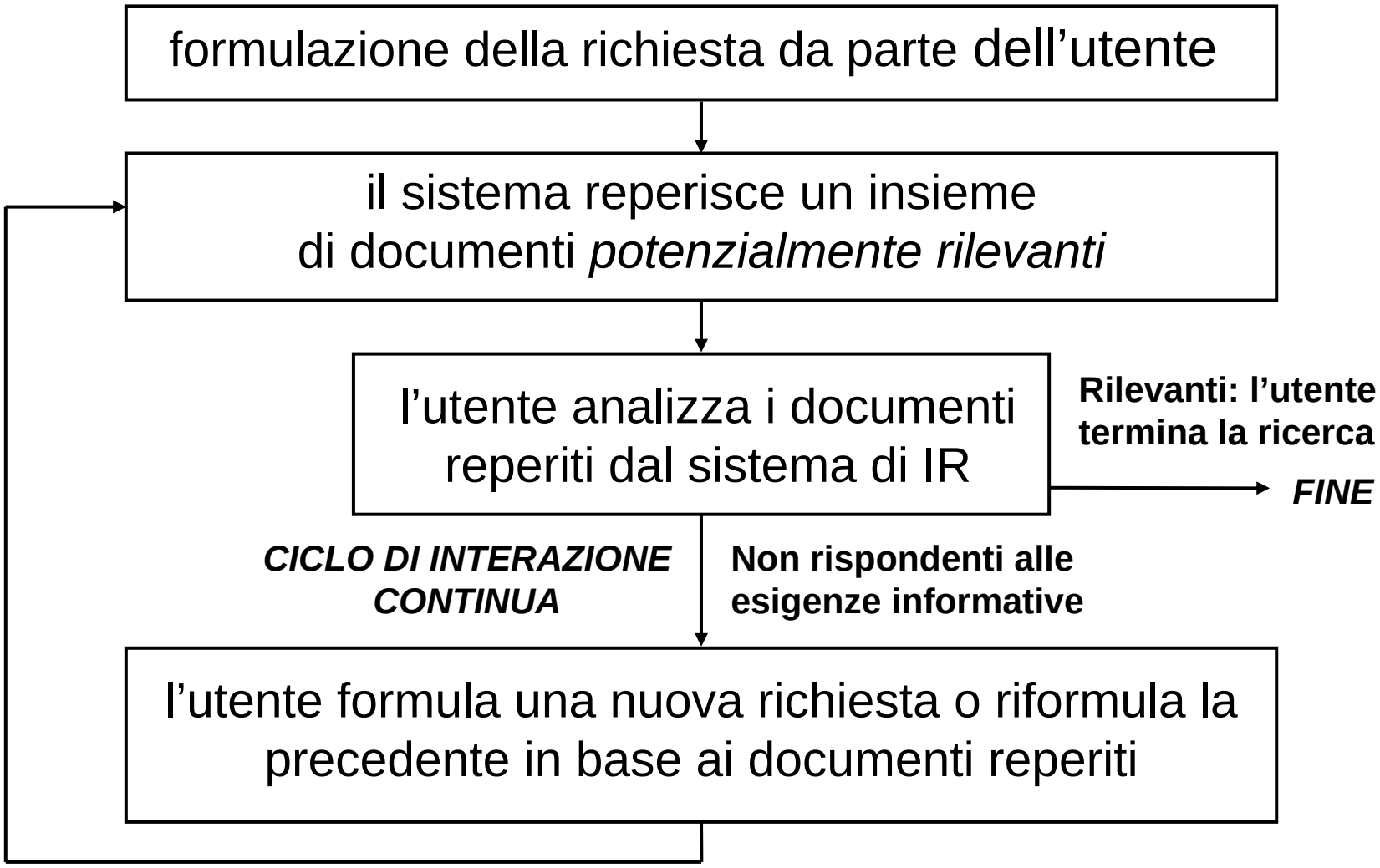
Il ruolo dell'utente

- L'utente ha un ruolo *fondamentale* nei sistemi di information retrieval
  - 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*
  - L'utente deve *interagire* con il sistema, valutandone le risposte, e *iterare* la propria richiesta variandone il contenuto
    - Raramente le ricerche vanno a buon fine al primo tentativo

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 in base ad una misura di *similarità* con la query (omettendo quelli lontani)
  - L'utente *valuta* i documenti presentati dal sistema
    - Operazione *lunga e tediosa* nel caso di documenti sonori
    - Se questi non soddisfano la sua esigenza informativa l'utente deve formulare una *nuova query*

Il ciclo presentazione/valutazione – 2



Accesso a documenti sonori

- **Identificazione**: dato un documento sonoro, o una sua parte, riconoscere se è una copia di un documento dato
  - L'approccio deve essere robusto verso compressione lossy, trasformazioni D/A e A/D, aggiunta di rumore, filtraggi,…
  - Vengono utilizzate tecniche di *audio fingerprinting*
- **Reperimento**: trovare i documenti che più probabilmente hanno le caratteristiche richieste, in base al contenuto
  - *Semantico*, ovvero all'argomento potenzialmente trattato
  - *Acustico*, ovvero alla presenza di particolari fonemi
- **Match esatto**: trovare tutti e soli di documenti che sono descritti dai *metadati* forniti dall'utente
  - Approccio generale, condiviso con altri media

Efficacia dei sistemi di IR – 1

- Identificazione e reperimento sono processi *approssimati*
  - L'estrazione degli indici è soggetta ad errori
    - Presenza di rumore di fondo
    - Estrazione delle feature imprecisa
  - Gli indici sono surrogati dei documenti
    - Necessità di efficienza spesso a discapito dell'efficacia
  - La query descrive *parzialmente* l'esigenza informativa dell'utente
    - Conoscenza parziale di ciò che si sta cercando
    - Numero limitato di esempi a disposizione
  - La similarità tra gli indici può non corrispondere alla similarità soggettiva percepita dall'utente

Efficacia dei sistemi di IR – 2

- Per valutare in modo oggettivo l'efficacia di sistemi di IR vengono organizzate delle campagne di valutazione
  - La più nota è TREC (TExt Retrieval Conference) organizzata dal NIST a partire dal 1998
    - Suddivisa in diverse track, per alcuni anni anche una di Spoken Document Retrieval
  - In Europa vi è CLEF (Cross Language Evaluation Forum)
  - Per la musica MIREX (Music Information Retrieval Evaluation eXchange) per i notiziari TDT (Topic Detection and Tracking)

- I partecipanti valutano i loro sistemi utilizzando le stesse collezioni, in modo da confrontare i risultati
    - Dall'inizio delle campagne di valutazioni vi è stato un notevole incremento delle prestazioni dei sistemi

Campagne di valutazione per l'IR

- Viene generalmente utilizzato il modello Cranfield che prevede l'uso di una *collezione sperimentale*, composta da
  - Un insieme di documenti
    - Da alcune migliaia a miliardi di documenti
  - Un insieme di query
    - Decise da esperti del settore
    - Normalmente in numero molto ridotto rispetto al numero dei documenti (per TREC solamente 50 query per track)
  - Dei giudizi di rilevanza che associano ogni query ai documenti
    - Formulati da esperti del settore
    - Onerosi da ottenere

Efficacia dell'identificazione

- In linea di principio un sistema di identificazione è efficace se il primo documento elencato dal sistema è quello corretto
  - In un sistema di IR, può risultare interessante misurare la posizione all'interno della lista dei documenti restituiti

- Dati $n = \{1,\ldots,N\}$ esperimenti, nei quali il documento da identificare è stato restituito in posizione r_n

$$\text{Error Rate} = \frac{N - \left\| \sum_{n=1}^{N} r_n = 1 \right\|}{N}$$

$$\text{Mean Reciprocal Rank} = \frac{\sum_{n=1}^{N} \frac{1}{r_n}}{N}$$

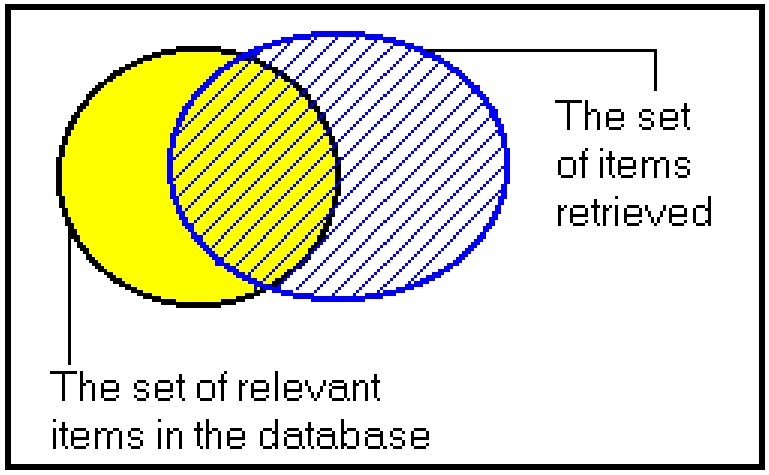
Efficacia del reperimento – 1

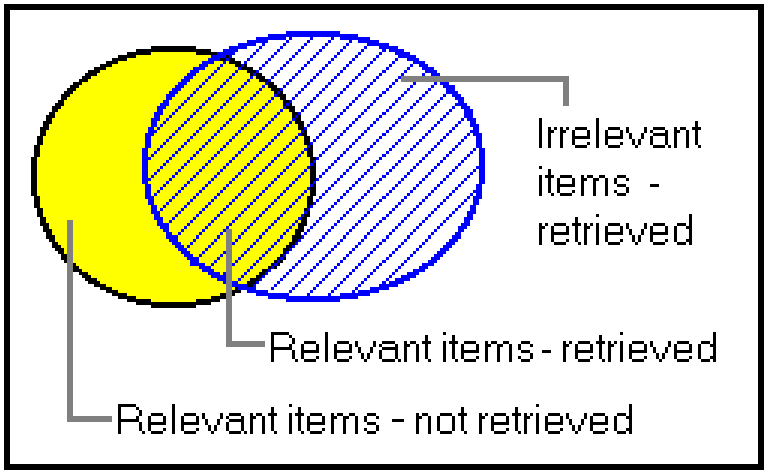
- Un sistema di IR fornisce una lista ordinata di documenti potenzialmente rilevanti per l'esigenza informativa dell'utente
  - L'effettiva rilevanza viene determinata da un insieme di esperti tramite *giudizi binari* (vero o falso)

- 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

Efficacia del reperimento – 2

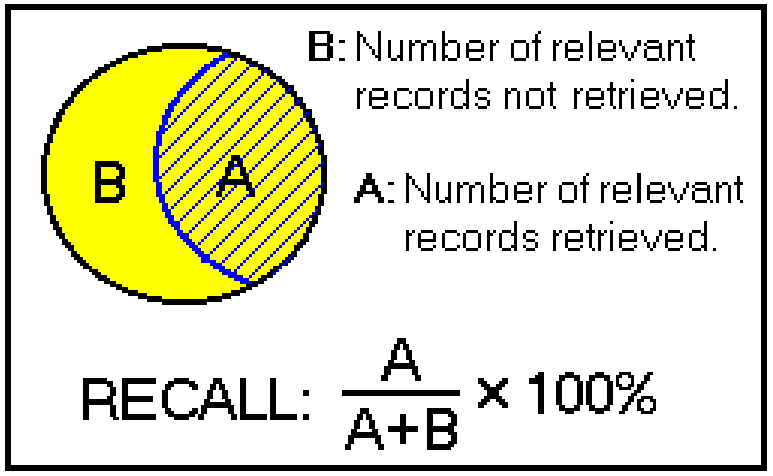
- I due parametri più utilizzati sono *precisione* e *richiamo*
- Data un'esigenza informativa e una query che la rappresenta, la collezione di documenti può essere partizionata
  - In base alla loro effettiva rilevanza dei documenti
  - In base al fatto che i documenti siano stati reperiti



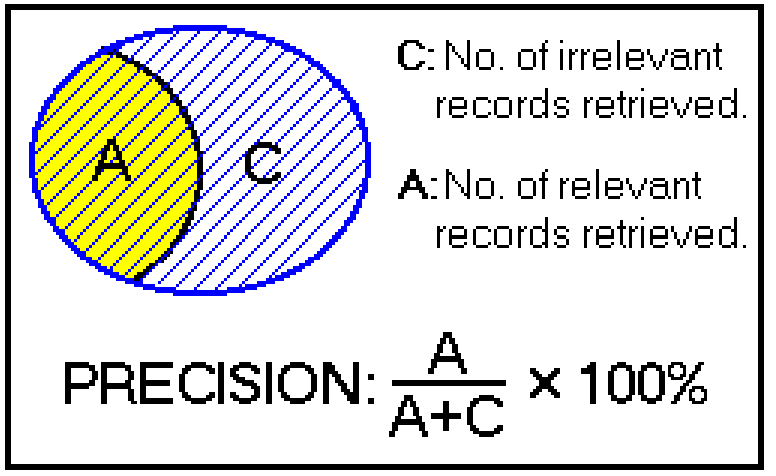


Efficacia del reperimento – 3

- **Richiamo** (recall) è il rapporto tra il numero di documenti rilevanti reperiti e il totale dei documenti rilevanti
  - 1 = tutta la verità



- **Precisione** (precision) è il rapporto tra il numero di documenti rilevanti reperiti e il totale dei documenti reperiti
  - 1 = nient'altro che la verità



Efficacia del reperimento – 4

- Richiamo e precisione sono in relazione di proporzionalità inversa
  - Aumentare il richiamo significa perdere in precisione e viceversa
- Dato che i documenti vengono riportati in liste ordinate, è anche importante l'ordine in cui vengono presentati i documenti
  - Vengono calcolate per i primi K documenti
  - La precisione viene calcolata a diversi livelli di richiamo
- Misure ad un solo valore
  - *Average precision*: la media della precisione calcolata ogni volta che un documento rilevante è osservato nella lista ordinata
  - *F-measure*: la media armonica di precisione e richiamo
  - *R-precision*: la precisione dei primi R documenti reperiti, dove R è il numero di documenti rilevanti

Il problema della rilevanza

- Nonostante la diffusione delle campagne di valutazione, esistono dei problemi irrisolti nella valutazione della rilevanza dei documenti

- La rilevanza:
  - E' soggettiva, in base alle competenze di chi valuta e alla sua interpretazione dell'esigenza informativa
  - Varia nel tempo, anche per lo stesso soggetto
  - Il giudizio su un documento influisce sui giudizi successivi
  - Dato che è impossibile conoscere la rilevanza di milioni di documenti, ci si avvale di strumenti automatici per reperire un pool di documenti da valutare

- In alternativa, vengono condotti esperimenti di reperimento in laboratorio, ovviamente molto costosi

Identificazione:
Audio Fingerprinting

Audio fingerprinting

- Un audio fingerprint è una descrizione compatta di un file audio che mantiene alcune caratteristiche percettivamente significative

- Un sistema di audio fingerprinting ha lo scopo di identificare duplicati di file audio anche in presenza di
  - Compressione lossy, anche a bassi bit rate
  - Presenza di disturbi, sia del supporto analogico iniziale che aggiunti durante la registrazione
  - Conversione digitale/analogica/digitale

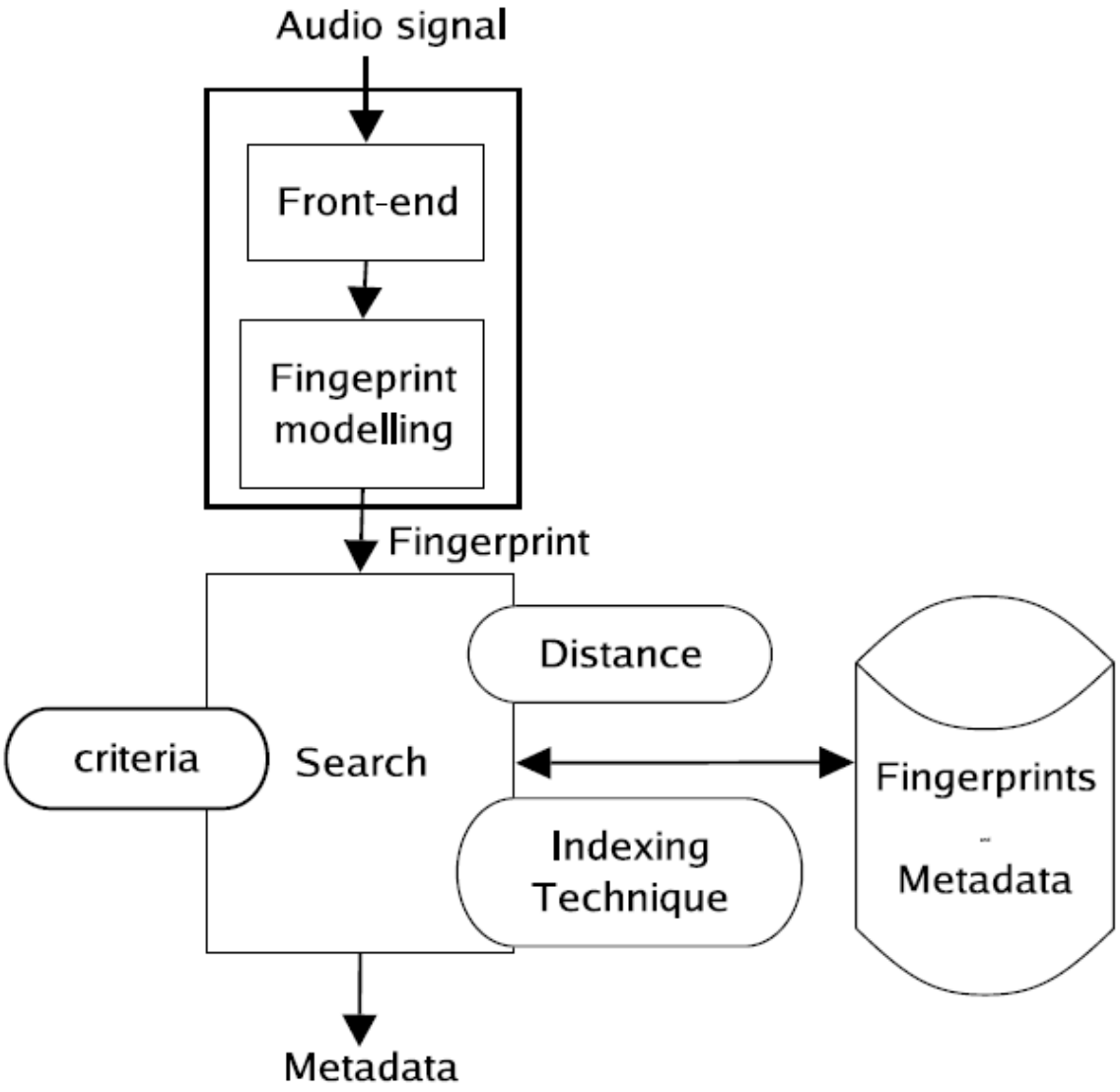
- Caratteristiche principali
  - Robustezza e affidabilità
  - Dimensione del fingerprint
  - Efficienza computazionale

Applicazioni del fingerprinting

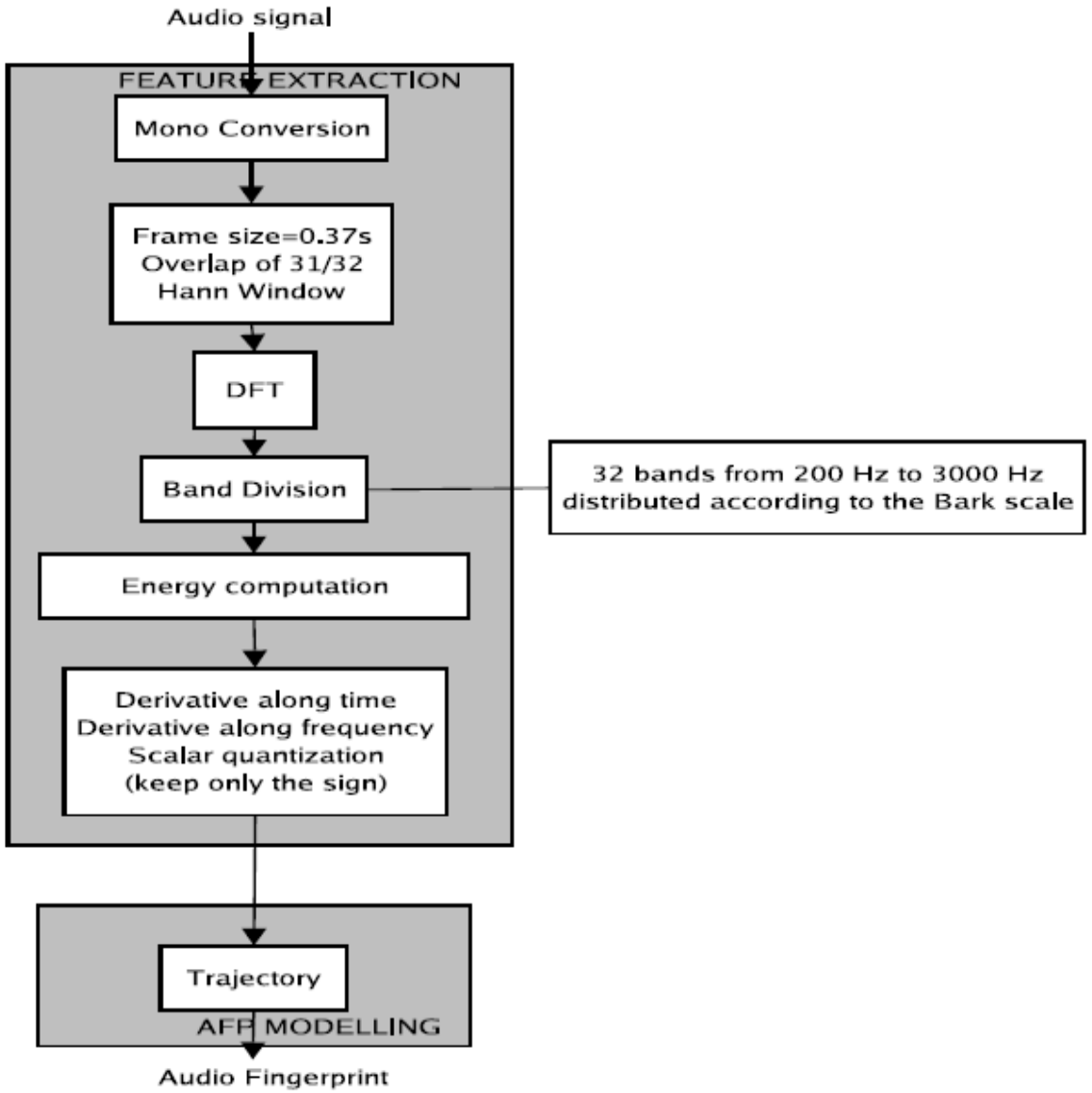
- L’audio fingerprinting è normalmente applicato
  - Alla tutela del diritto d’autore
    - Diffusione di copie illegali (file sharing, siti Web)
  - Al reperimento automatico di metadati
    - Servizio per gli utenti, accompagnato alla vendita
  - Localizzazione temporale di un estratto all’interno di un file completo
    - Tracciamento di spot pubblicitari e promo

- Nel caso degli archivi sonori, si può
  - Controllare la diffusione di materiale in possesso dell’archivio
  - Fornire strumenti avanzati di ricerca
  - Controllare la presenza di duplicati (interi o parti)

Schema generale per il fingerprinting



Un approccio al fingerprinting – 1

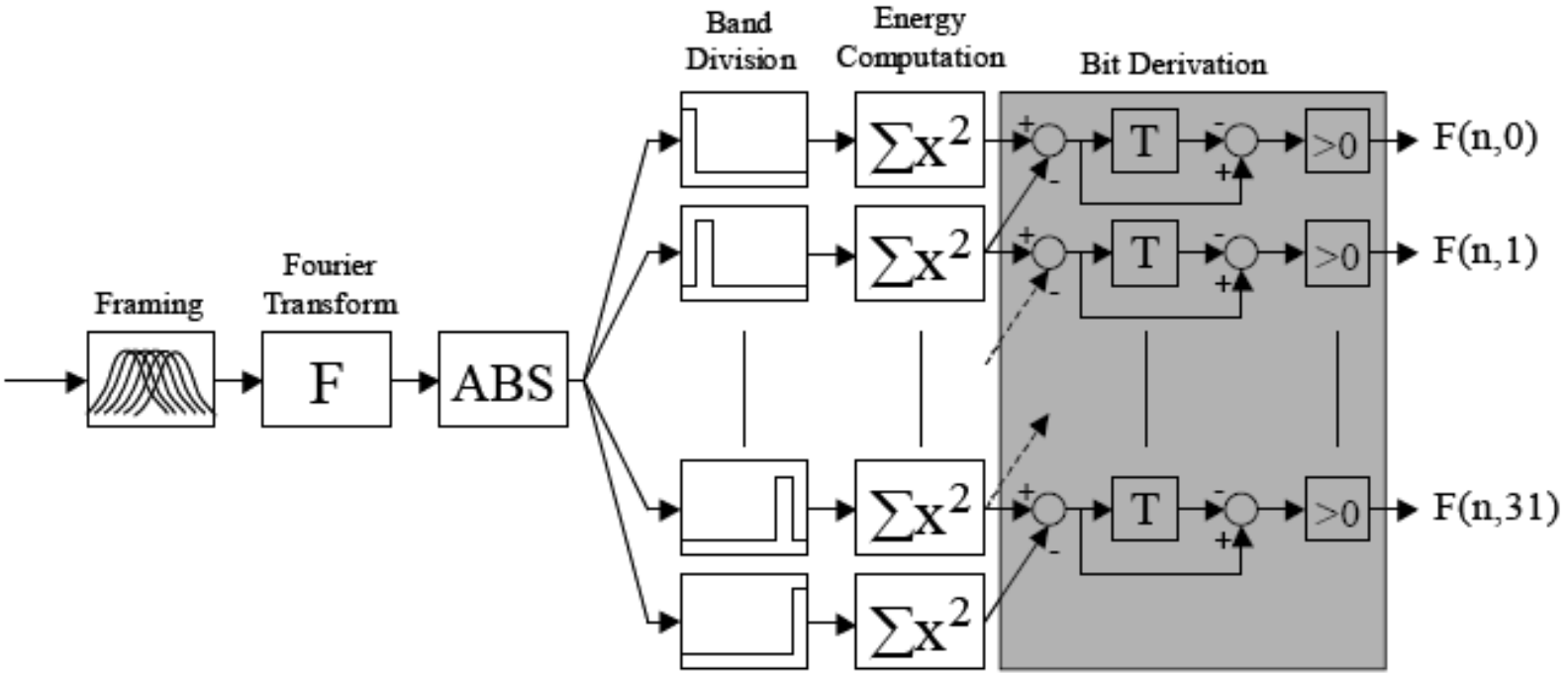


Un approccio al fingerprinting – 2

- L’audio viene suddiviso in frame che si sovrappongono in gran parte
  - Rappresentazione molto ridondante temporalmente
    - Utile perché in linea di principio i due file da confrontare non sono allineati
- Viene utilizzata la rappresentazione in frequenza
  - Percettivamente più rilevante della rappresentazione nel tempo
    - La fase del segnale è ignorata
- Lo spettro è suddiviso in bande contigue, spaziate logaritmicamente
  - Ispirato al concetto di banda critica (anche se le bande non corrispondono esattamente alla scala Bark)
- Ogni banda viene rappresentata con un singolo bit
  - Valore basato sul valore dell’energia e il confronto con una banda contigua

Diagramma di funzionamento

- Schema di trasformazione del segnale audio in un vettore binario
  - Si usano 32 bit per efficienza nell'occupazione di memoria



Calcolo effettivo dei fingerprint

- La funzione che a partire dall'energia nelle diverse bande calcola il valore del fingerprint è stata trovata in modo euristico
  - Semplicemente funziona meglio di altre
    - Il confronto con le bande vicine e con il valore precedente rende il fingerprint più stabile

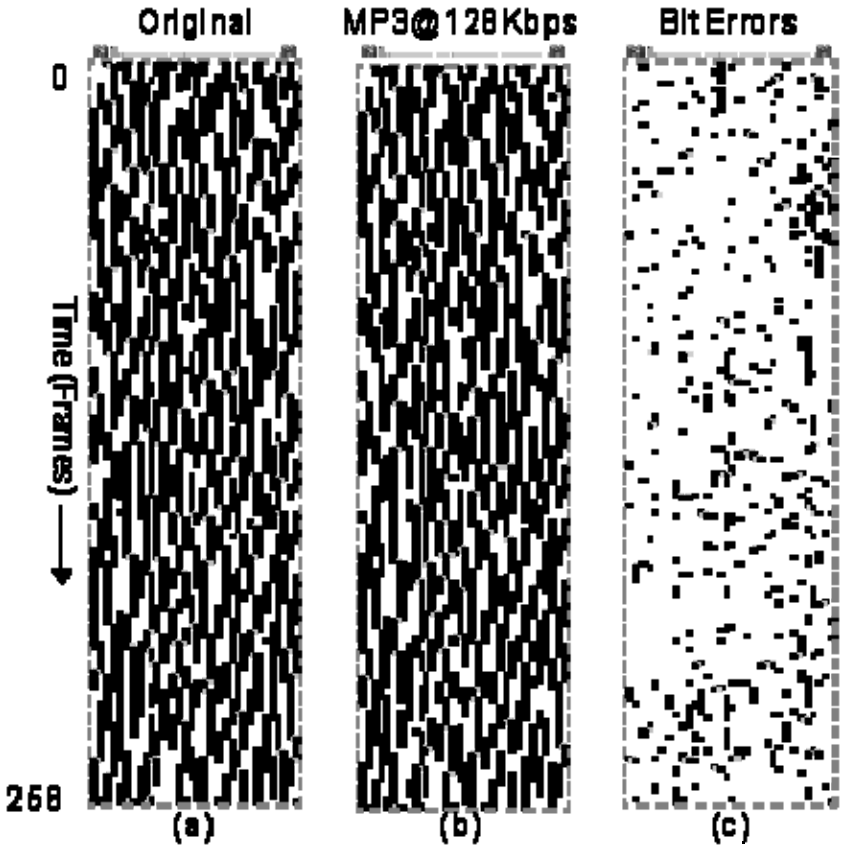
- Il valore *F(n,m)* del fingerprint al frame *n* per la banda *m*, corrispondente al bit *m*, si calcola a partire dall'energia del segnale

$$E(n,m) = \sum_{x=b_m}^{b_{m+1}-1} x^2 \qquad \text{dove } b_m \text{ è l'estremo sinistro della banda m - esima}$$

$$F(n,m) = \begin{cases} 1 & \text{if } E(n,m) - E(n,m+1) - (E(n-1,m) - E(n-1,m+1)) \geq 0 \\ 0 & \text{if } E(n,m) - E(n,m+1) - (E(n-1,m) - E(n-1,m+1)) < 0 \end{cases}$$

Rappresentazione grafica

- Esempio della differenza tra un segnale di partenza e la sua versione compressa lossy
- Ogni frame è un vettore di 32 bit (unsigned int), rappresentato in b/n
- La differenza tra i due non è altro che il risultato dell'operazione di XOR



Ricerca efficiente – 1

- Una *ricerca lineare*, dove il fingerprint dell'audio da riconoscere viene confrontato con tutti i fingerprint nel database non è fattibile
  - Anche per un database di piccole dimensioni, 1000 file da 3 minuti, un solo riconoscimento impiegherebbe *giorni*

- E' necessario utilizzare un indice, ma:
  - *Problema 1:* il fingerprint è una rappresentazione troppo semplificata (soli 32 bit per millisecondi di audio)
    - Alta probabilità di falsi positivi
  - *Problema 2*: il fingerprint non può essere robusto a tutti i disturbi (un rumore impulsivo in una banda può modificare uno o più bit di un dato frame)
    - Alta probabilità di falsi negativi

Ricerca efficiente – 2

- La similarità deve essere calcolata su di una finestra temporale, centrata sul fingerprint trovato nell'indice, che coinvolga un numero sufficientemente elevato di fingerprint
  - Viene calcolato uno XOR tra i bit delle due finestre temporali
  - Si contano i bit=1 e si confrontano con una data soglia
    - Soglia da determinarsi sperimentalmente

- Obiettivi
  - Compensare la presenza di falsi positivi
    - I pochi falsi positivi saranno probabilmente circondati da valori distanti, abbassando il valore globale della similarità
  - Avere almeno un valore che non sia un falso negativo
    - In questo modo è possibile allineare i due frame e calcolarne la distanza con la funzione di XOR

Problematiche

- L’assunzione che almeno un indice sia corretto non è realistica
  - La presenza di disturbi sovrapposti può alterare tutti i frame del segnale (che potrebbero tutti avere almeno un bit diverso)
  - E’ necessario sostituire il match esatto tra i fingerprint, che soggiace al concetto di indicizzazione, con una funzione distanza
    - La distanza di Hamming si presta naturalmente ad essere utilizzata
    - La ricerca nell’indice dei fingerprint può essere estesa a quelli che hanno distanza maggiore di 0 (tipicamente 1 o 2)
- L’assunzione che due frame si allineino perfettamente non vale per passaggi su supporti analogici
  - La funzione di XOR può essere raffinata applicando tecniche di allineamento automatico (Dynamic Time Warping)

Prestazioni del fingerprint

- Collezione di 200 mila file (MP3 di durata media 3 minuti)
- Task di riconoscimento di brani su 24 ore di emesso televisivo o radiofonico
  - Nota: il sistema utilizzato usa un calcolo del fingerprint diverso,
- Tempi di riconoscimento: 8 ore per una trascrizione completa
- Error Rate = 5%, Mean Reciprocal Rank = 97.3
- Errore medio nella rilevazione delle durate: 1.2 secondi
- Percentuale falsi positivi = 0.4%

Textual and Spoken Document Retrieval

Indicizzazione automatica di testi – 1

- L'indicizzazione automatica di un documento contenente testo 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* presenti nell'intera collezione di documenti
    - 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 *accelera* la ricerca (esempio, *indice analitico*)

Indicizzazione automatica di testi – 2

- 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

Indicizzazione di documenti sonori

- L'ovvia estensione ai documenti sonori vocali riguarda l'estrazione automatica dei termini tramite tecniche di speech recognition
  - Operazione analoga all'analisi lessicale per i testi

- Problematiche
  - Alta percentuale di errori in fase di riconoscimento
    - La lingua parlata è molto più ridondante della lingua scritta, gli errori possono essere compensati
  - Affidabilità legata alla presenza di dizionari
    - Potrebbero non essere disponibili per alcuni corpora
  - Necessità di segmentare documenti lunghi
    - Necessità di tecniche di speaker identification
    - Applicazione di sistemi di topic detection and tracking

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 dei termini

- 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 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

- 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
- Ad esempio, nel Web, che contiene documenti su *qualsiasi argomento*, le stop-words sono le parole *molto frequenti* nelle 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*
- 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 "dentifricio" 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

parole \ **documenti**	D_1	D_2	D_3
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
....		...	

La fase di reperimento

- La fase di indicizzazione estrae degli indici dai documenti testuali
  - Gli indici sono delle *parole*, che esprimono in modo sintetico il contenuto informativo dei documenti

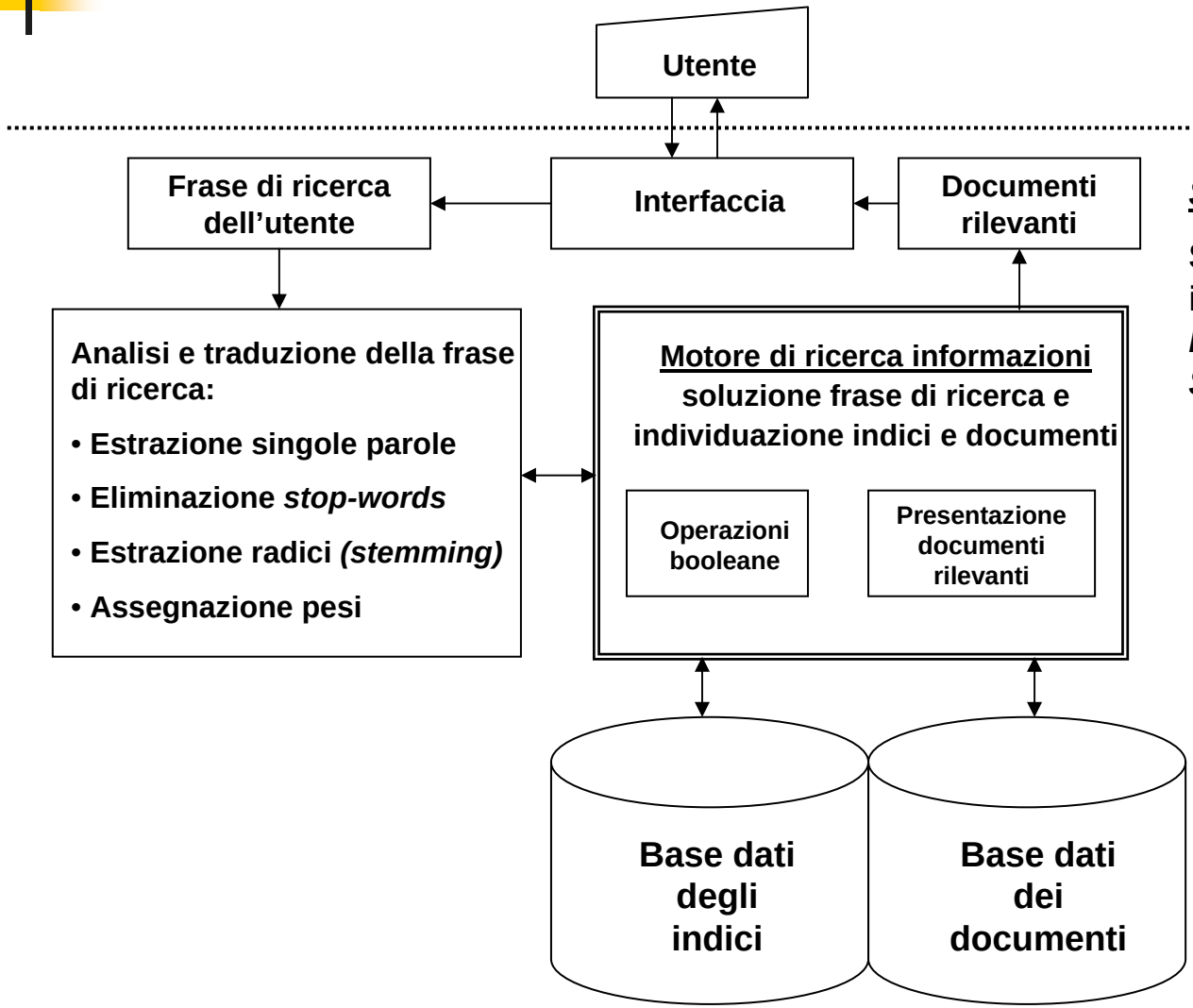
- La fase di ricerca si basa anch'essa sull'uso di *parole* che sintetizzano l'esigenza informativa
  - L'utente formula la sua query utilizzando alcune parole, spesso indicate con il termine di *parole chiave* o *key-words*
    - Il sistema *indicizza* la query, così come ha fatto per i documenti, e 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 nel settore è molto attiva

La fase di reperimento

- La fase di indicizzazione 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 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 nel settore è molto attiva

Il processo completo di reperimento



***Software* di gestione:**

Sistema di reperimento informazioni o *Information Retrieval System (IRS)*

Il modello booleano – 1

- Un modello molto diffuso per i linguaggi di interrogazione è il modello *booleano*, che si applica alla pesatura binaria
  - 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 è il seguente:
  - **AND** (binario): *entrambi* i termini devono essere presenti
  - **OR** (binario): *almeno uno* dei termini deve essere presente
  - **NOT** (unario): 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*

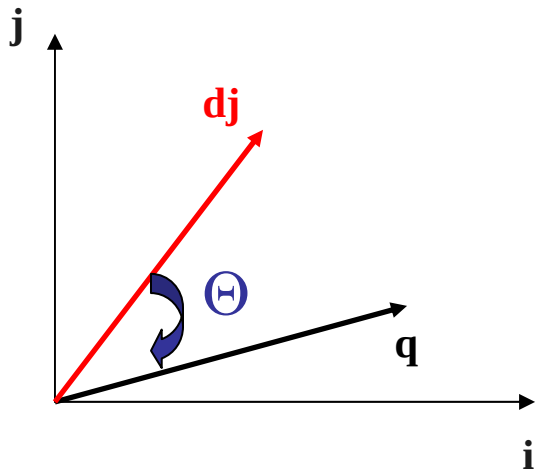
Il modello booleano – 2

- Gli operatori booleani possono essere *combinati* tra loro
  - (Mozart OR Beethoven) AND (sonata OR concerto) AND (NOT (piano OR clavicembalo OR organo))

- Il modello booleano ha alcune caratteristiche:
  - *Vantaggi*:
    - Implementazione software *intuitiva*
    - 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

Il modello vettoriale – 1

- E’ una estensione del modello booleano che si basa sul calcolo della *similarità* tra i documenti e la query
  - Documenti e query sono rappresentati come vettori di pesi
    - w_{ij} = peso (>0) con cui il termine *i* descrive il documento *dj*
- Un documento è un vettore in uno spazio t-dimensionale
  - t è il numero complessivo dei termini indice



Il modello vettoriale – 2

- La similarità è definita come il coseno dell'angolo formato dai due vettori nello spazio t-dimensionale

$$sim(d_j, q) = cos(\Theta) = \frac{\sum_{i=1}^{t} w_{ij} * w_{iq}}{|d_j| * |q|}$$

$$\text{con} \quad 0 \leq sim(d_j, q) \leq 1$$

- Il modello consente quindi di reperire anche documenti che soddisfano solo parzialmente l'interrogazione

- Il prodotto può essere calcolato in modo efficiente, considerando che i due vettori sono sparsi

Scelta dei pesi

- L'efficacia del retrieval dipende in larga misura dalla scelta dei pesi
  - Un termine che appare spesso in un documento è potenzialmente un buon descrittore (tf = term frequency)
  - Un termine che appare in molti documenti non è discriminante (idf = inverse document frequency)
    - Risultati noti nel settore della linguistica
- Dati:

  N = #documenti

  n_i = #(documenti che contengono i)

  freq(i,j) = #(occorrenze di i in d_j)

$$w_{ij} = tf * idf = \frac{freq(i,j)}{max(\ freq(i,j))} * log\left(\frac{N}{n_i}\right)$$

Estensione ai documenti vocali

- La scelta di reperire le trascrizioni dei documenti ha alcuni svantaggi
  - Perde l'informazione sulla confidenza del sistema sulla scelta di una particolare parola
  - Non considera trascrizioni alternative
    - Dipende quindi solo dalla qualità della trascrizione
- E' possibile estendere il modello vettoriale
  - Il concetto di appartenenza di un termine ad un documento è sostituito da una funzione di probabilità
  - La frequenza relativa è pesata da questa probabilità
    - Sono già disponibili soluzioni efficienti
  - Lo schema *tf*idf* viene calcolato e pesato in base alle probabilità in uscita dal modulo di trascrizione

Fonemi al posto di parole

- Uno dei problemi dello spoken IR è dato dalle parole al di fuori del vocabolario di riferimento
  - Nomi stranieri, toponimi, forme gergali
- Un approccio (ad esempio IBM) è di usare i fonemi con indici
  - La maggior segmentazione della trascrizione di fonemi riduce l'effetto degli errori di trascrizione
  - Non è necessario disambiguare
    - Ad esempio, i termini inglesi "C", "sea" e "see"
- Problema aperto
  - Che schema di pesatura ha senso utilizzare?
    - Booleano
    - Variazione del classico tf*idf

Reperimento basato su altri parametri

- In linea di principio, il reperimento può essere basato su qualsiasi parametro audio percettivamente rilevante
  - Intonazione
  - MFCC
  - Posizione dei formanti
- E' necessario definire una funzione di similarità nello spazio dei parametri
  - Possibilità di utilizzare più parametri contemporaneamente
    - Tecniche di data fusion per valutare la potenziale rilevanza
- Il reperimento basato su una funzione di similarità calcolata su più parametri non è efficiente
  - Ricerca lineare nello spazio dei parametri?

Efficienza della ricerca

- Locality Sensitive Hashing
  - Frame simili vengono mappati nello stesso valore da una opportuna funzione (detta di hash)
    - Si sfruttano le collisioni per trovare frame simili
    - Per sicurezza si usano diverse funzioni alternative
- Riduzione della dimensionalità
  - Si sfrutta la correlazione nello spazio dei parametri per calcolare la similarità in uno spazio di dimensioni minori
    - Principal Component Analysis
- Uso di spazi metrici
  - Si evidenziano dei cluster tra i frame, e si conduce la ricerca partendo dai centroidi dei cluster