

Sistemi per il recupero delle informazioni

Gabriele Pozzani

A.A. 2014/2015

**Corso di Laurea Magistrale in
Editoria e Giornalismo**

Introduzione al corso

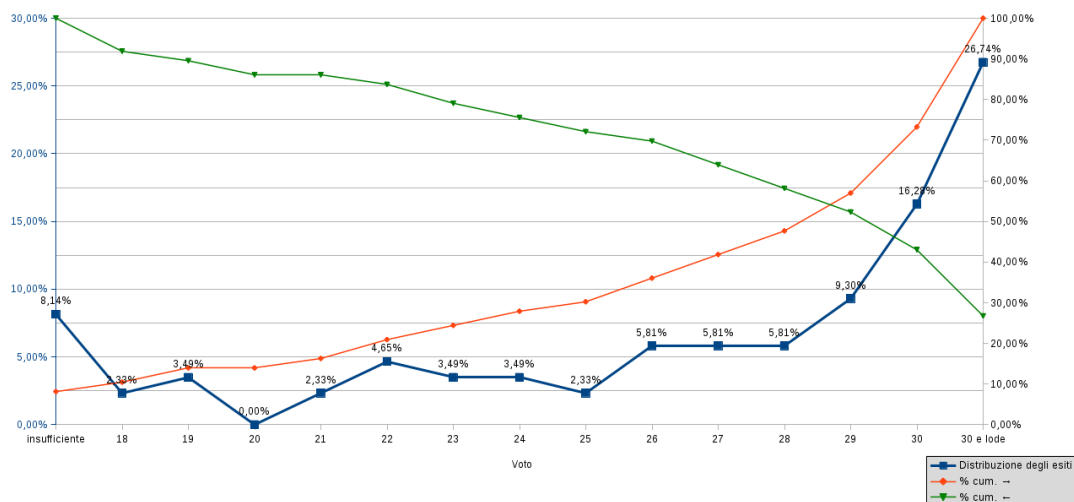
Informazioni generali (I)

- Orario
 - Giovedì dalle 8.30 alle 10.10
 - Venerdì dalle 8.30 alle 10.10
 - Vedere il calendario sul sito del corso
- Ricevimento
 - Giovedì dalle 14.30 alle 16.30 su appuntamento (da richiedere tramite email) a Lettere o a Scienze MM.FF.NN.

Informazioni generali (II)

- Esame
 - Scritto
 - È possibile presentarsi a qualunque appello e poter ripetere l'esame quante volte si vuole (entro l'AA di erogazione del corso).
 - Chi ha ottenuto una votazione positiva ad un appello può “congelare” il voto e ripetere l'esame presentandosi ad uno dei successivi appelli. Chi ha già un voto positivo e si ripresenta ad un successivo appello e consegna il compito perde senza possibilità di appello il voto precedente già acquisito (che il nuovo compito sia migliore o peggiore del precedente). Chi si ripresenta ma non consegna e si ritira invece mantiene la vecchia votazione già acquisita.

A proposito di esami



- Il 64% prende almeno 27

Informazioni generali (III)

- Materiale
 - Slide e dispense pubblicate sul sito del corso (sufficienti per l'esame)
- Per approfondire e come riferimento
 - Yates, Ricardo. *Modern information retrieval: the concepts and technology behind search*. New York: Addison Wesley, 2011.
 - Slide originali e alcuni capitoli disponibili gratuitamente online <http://www.mir2ed.org/>
 - Manning, Christopher. *Introduction to information retrieval*. New York: Cambridge University Press, 2008.
 - Libro e slide originali disponibili gratuitamente online <http://informationretrieval.org/>
 - Korfhage, Robert. *Information storage and retrieval*. New York: Wiley Computer Pub, 1997.
 - Abiteboul, Serge. *Data on the web: from relations to semistructured data and XML*. San Francisco: Morgan Kaufmann, 1999.

Obiettivo del corso

- Studio degli aspetti fondamentali dei sistemi per il recupero dell'informazione
 - Funzionalità e caratteristiche comuni
 - Sistemi sul WEB: motori di ricerca
- Introduzione ai dati semistrutturati e all'uso di XML per l'editoria elettronica

Programma

1. Sistemi per il recupero delle informazioni
2. Analisi di testi
3. Sistemi per il recupero delle informazioni multimediali: cenni
4. Sistemi per il recupero delle informazioni sul Web
5. Le interfacce utente per il recupero delle informazioni: cenni
6. Documenti semistrutturati e recupero dell'informazione: structured text retrieval

Cosa vedremo?

- È vero che i migliori risultati sono i primi?
- Perché i motori di ricerca sbagliano?
- Perché a volte possiamo ricercare su tutto il testo e a volte solo su alcune informazioni?
- Se due persone eseguono la stessa ricerca ottengono gli stessi risultati?
- I motori di ricerca sono imparziali?
- Come funziona la ricerca di immagini in Internet?
- Come fanno i motori di ricerca a trovare le pagine dei siti?
- I siti e i loro contenuti sono tutti uguali agli occhi dei motori di ricerca?

Introduzione ai Sistemi Informativi e ai Sistemi per il Recupero delle Informazioni

Sistema Organizzativo

- Insieme delle risorse e regole utilizzate per lo svolgimento coordinato delle attività (processi) necessarie al perseguimento degli scopi dell'organizzazione
- Le risorse di un'organizzazione (azienda, ente, amministrazione) sono:
 - Persone
 - Denaro
 - Materiali
 - Informazioni

Sistema informativo (I)

- Definizione
 - “è l'insieme delle attività umane e dei dispositivi di memorizzazione ed elaborazione che organizza e gestisce l'informazione di interesse di un'organizzazione di dimensioni qualsiasi”

Sistema informativo (II)

- Componente di una organizzazione che **gestisce le informazioni di interesse** (cioè utilizzate per il perseguimento degli scopi dell'organizzazione)
- Ogni organizzazione ha un sistema informativo, eventualmente non esplicitato nella struttura
- Il sistema informativo è di supporto ad altri sottosistemi, e va quindi studiato nel contesto in cui è inserito

Gestione delle informazioni

- Raccolta, acquisizione
- Archiviazione, conservazione
- Elaborazione, trasformazione, produzione
- Distribuzione, comunicazione, scambio

Informazione di interesse

- Informazioni utilizzate per il perseguimento degli scopi dell'organizzazione

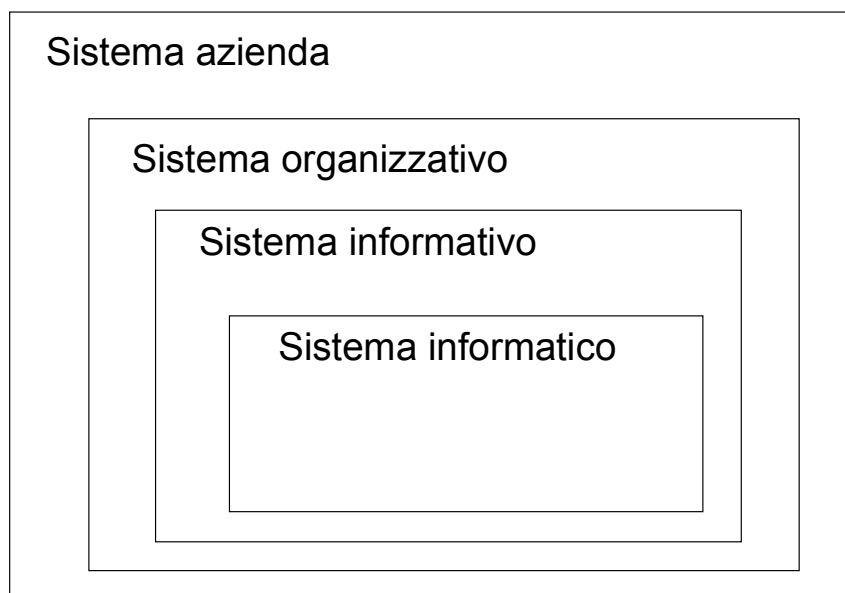
Sistemi informativi e automazione

- Il concetto di “sistema informativo” è indipendente da qualsiasi automatizzazione:
 - esistono organizzazioni la cui ragion d’essere è la gestione di informazioni (per esempio servizi anagrafici e banche) e che operano da secoli (senza computer e dispositivi elettronici)

Sistema Informatico

- Porzione automatizzata del sistema informativo:
 - la parte del sistema informativo che gestisce informazioni con tecnologia informatica

Gerarchia dei sistemi



Gestione delle informazioni

- Nelle attività umane, le informazioni vengono gestite in forme diverse:
 - idee informali
 - linguaggio naturale (scritto o parlato, formale o colloquiale, in varie lingue)
 - disegni, grafici, schemi
 - numeri e codici
- e su vari supporti
 - mente umana, carta, dispositivi elettronici

Informazioni e dati

- Nei sistemi informatici (e non solo), le **informazioni** vengono rappresentate in modo essenziale, spartano: attraverso i **dati**

Informazioni e dati (I)

(definizioni dal Vocabolario della lingua italiana 1987)

informazione: notizia, dato o elemento che consente di avere conoscenza più o meno esatta di fatti, situazioni, modi di essere.

dato: ciò che è immediatamente presente alla conoscenza, prima di ogni elaborazione; (in informatica) elementi di informazione costituiti da simboli che debbono essere elaborati

Informazioni e dati (II)

In altre parole:

- Dato:
 - Elemento di conoscenza di base costituito da simboli che devono essere elaborati.
- Informazione:
 - Interpretazione dei dati che permette di ottenere conoscenza più o meno esatta di fatti e situazioni.

Dati e informazioni



- che cosa significano questi numeri?
- cartelli stradali in Finlandia
- ma la differenza?
 - senza "interpretazione" il dato serve a ben poco

Dati e informazioni



Lun-Ven



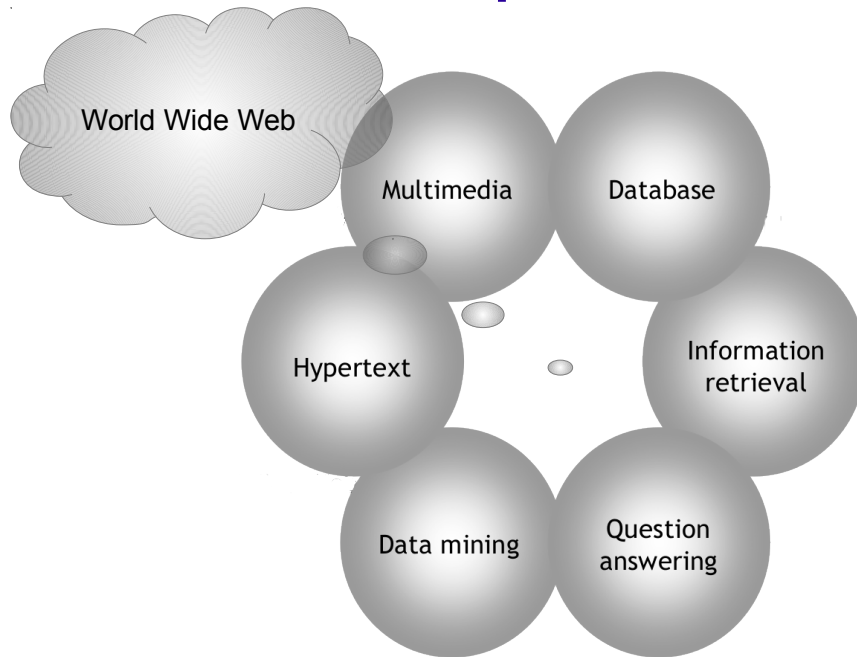
Sabato



Festivo

- che cosa significano questi numeri?
- cartelli stradali in Finlandia
- ma la differenza?
 - senza "interpretazione" il dato serve a ben poco

Tipi di SI



Sistemi di basi di dati

- Informazioni
 - semplici nella natura
 - potenzialmente molto complesse nella struttura
 - di elevate dimensioni
- Interrogazioni
 - anche complesse: annidamenti, join
 - ripetitive, quindi standard, classificate e precompilate
- Aggiornamenti
 - frequenti e casuali
 - in linea
- Problemi specifici
 - sicurezza
 - integrità
 - efficienza

Sistemi ipertestuali

- Informazioni
 - complessità variabile (testo → multimedia)
 - codifica non uniforme (...HTML...), significato non formalizzato (...XML...)
 - con livelli di strutturazione molto variabili
- Interrogazioni
 - navigazione e esplorazione libera
 - non prevedibili / non classificabili (~alcuni siti WWW)
- Aggiornamenti
 - dipendono dall'applicazione (generalmente poco frequenti)
 - in linea / fuori linea (dipende dall'applicazione)
- Problemi specifici
 - interazione, usabilità
 - orientamento, aiuto
 - portabilità, indipendenza dalla piattaforma e dal dispositivo (multicanalità)

Sistemi di information retrieval (testo)

- Informazioni
 - di natura semplice: autori, argomenti, riferimenti
 - poco strutturate: testi in prosa, tabelle
 - molto numerose
- Interrogazioni
 - struttura semplice, molte clausole, indicatori di importanza (pesi)
 - specifiche parziali, ricerca approssimata
 - iterazioni successive
- Aggiornamenti
 - periodici a bassa frequenza, programmabili
- Problemi specifici
 - dizionari, thesauri
 - ontologie

Sistemi di information retrieval (multimedia)

- Informazioni
 - di natura semplice ma di significato complesso
 - il concetto di struttura è variabile (immagine, filmato, audio)
 - elevate dimensioni per ogni esemplare di dato
- Interrogazioni
 - scarsa corrispondenza tra forma e significato
 - specifiche "sintattiche", ricerca per somiglianza
 - iterazioni successive, analisi di rilevanza
- Aggiornamenti
 - non frequenti, generalmente fuori linea
- Problemi specifici
 - la rappresentazione codificata non contiene il significato del dato
 - la codifica e la rappresentazione influiscono sui sistemi di gestione

Information retrieval

- I sistemi di Information Retrieval
 - Sono correlati alla rappresentazione, memorizzazione, organizzazione ed accesso ad informazioni in archivi di grandi dimensioni
- Principali obiettivi sono l'indicizzazione, classificazione e ricerca dei documenti
 - basati sull'identificazione del contenuto informativo attraverso un utilizzo controllato del linguaggio naturale
- Possibili applicazioni
 - ricerche bibliografiche
 - ricerca documentaria
 - consultazione di archivi giuridici e normativi
 - catalogazione di oggetti eterogenei
 - archiviazione di documenti in prosa
 - analisi letteraria e linguistica

Gli inizi (1)

- L'umanità organizza le informazioni per una successiva ricerca da quasi 5000 anni
 - Tali informazioni erano contenute in appositi edifici: le biblioteche
 - La prima biblioteca conosciuta si trovava a Ebla (attuale Tell Mardīkh, Siria) nel 2500 AC circa

Gli inizi (2)

- Con l'aumento dei documenti si sono rese necessarie apposite strutture dati per la ricerca veloce: gli indici
 - Per secoli sono stati creati manualmente
 - Insiemi di categorie ed etichette
 - L'informatizzazione ha permesso la costruzione automatica di grandi indici

Gli sviluppi meno recenti

- Le biblioteche sono state tra le prime istituzioni ad adottare sistemi di IR
 - Inizialmente con l'automatizzazione della ricerca nei cataloghi
 - Poi introducendo la catalogazione tramite etichette, parole chiave, sommari e permettendo la ricerca tramite interrogazioni “complesse”

Gli sviluppi più recenti

- Molto è cambiato con la nascita del Web
 - Internet è diventato il più grande repository di conoscenza della storia umana
 - Data la sua dimensione, l'accesso alle informazioni in Internet passa sempre per una ricerca
 - Motori di ricerca
 - La ricerca su Internet è completamente basata sulle tecniche e tecnologie di IR

Il problema dell'IR (1)

- Gli utenti dei sistemi di IR (come i motori di ricerca) necessitano di informazioni di complessità variabile
- Esempio

Trovare tutti i documenti che riguardano il finanziamento da parte del Governo di progetti per lo sviluppo ferroviario

Il problema dell'IR (2)

- Questa descrizione completa non è in generale una buona interrogazione per un sistema di IR
 - L'utente deve tradurre questa richiesta in un'interrogazione
 - La traduzione produce un insieme di parole chiave, o termini d'indicizzazione, che riassumono la richiesta iniziale
 - Esempio:

finanziamento governo sviluppo
ferroviario

Il problema dell'IR (3)

- Sulla base dell'interrogazione, l'obiettivo del **sistema di IR** è recuperare tutti i documenti rilevanti e utili all'**utente**
 - Cioè il sistema di IR deve ordinare i documenti, che (lui) conosce, a seconda di quello che (lui) ritiene essere il grado di rilevanza rispetto all'interrogazione dell'utente
 - Recuperare tutti i documenti rilevanti scartando il maggior numero possibile di documenti non rilevanti per l'utente

Il problema dell'IR (4)

- Come sempre, quando si è in due possono nascere fraintendimenti e incomprensioni
 - Cosa è rilevante per l'utente e come può essere valutato?
 - Cosa è rilevante per il SRI e come può essere valutato?
 - Cosa intende con un certo termine l'utente e come lo intende invece il SRI?

Il problema dell'IR (5)

- Esempio:
 - finanziamento governo sviluppo
ferroviario
 - Quale governo?
 - Italiano? UE? Qualunque governo nel mondo?
 - Quali finanziamenti?
 - Diretti o indiretti? Attuali o passati?
 - Ferroviario?
 - Merci o persone? Nazionale o internazionale?

Il problema dell'IR (6)

- Concludendo
 - Bisogna capire e ricordarsi innanzitutto che quando si interagisce con un SRI (specialmente nel WEB) si ha di fronte un'entità diversa da noi e pensata da persone (molto) diverse da noi
 - Noi facciamo una domanda avendo in mente un certo bisogno
 - Il SRI “interpreta” la nostra domanda secondo i suoi mezzi e risponde secondo quello che LUI ha capito, conosce e pensa importante
 - Noi leggiamo la risposta considerando che necessita di essere reinterpretata secondo i nostri bisogni

Le possibili ricerche utente (I)

- Esistono diversi tipi di bisogni informativi e quindi diversi tipi di ricerche
 - Bisogno di informazioni puntuali
 - Settimo presidente della repubblica italiana
 - Numero di respinti all'esame di “sistemi per il recupero delle informazioni”
 - Bisogno informativo di costruzione di un contesto
 - Cosa è, cosa si intende per, e su quali principi si basano le scienze della comunicazione?
 - Bisogno informativo soggettivo
 - Migliori film commedie

Le possibili ricerche utente (II)

- Ricerca mirata in funzione del contenuto (querying)
 - classificazione argomentale
 - strutturazione del contenuto
 - indicizzazione
 - dizionari dei sinonimi
 - lemmatizzazione
- Ricerca esplorativa (browsing)
 - navigazione
 - approssimazioni successive
 - storia e orientamento
 - ricerca incrementale

Formulazione delle interrogazioni

- A seguito di una interrogazione, il sistema segnala il numero di documenti ritrovati
 - l'utente può riformulare, specializzare o generalizzare l'interrogazione fino a che il numero di documenti ritrovati appare soddisfacente
- L'esame dei documenti si avvale di due funzionalità
 - ranking: i documenti sono presentati all'utente in ordine decrescente di rilevanza, secondo i pesi assegnati ai termini
 - browsing: i documenti sono raggruppati in classi di somiglianza, permettendo all'utente di "sfogliarli" secondo un ordine logico

Documenti e IR (1)

- Un documento è una qualsiasi collezione di informazioni rintracciabile in base alla descrizione del suo contenuto
 - Testi
 - Libri
 - Articoli
 - e-mail
 - Comunicazioni informali
 - Lettere
 - Messaggi
 - Informazioni codificate
 - File
 - Dati numerici e tabelle
 - Immagini e disegni
 - Suoni e voci

Documenti e IR (2)

- Le tecniche di IR dipendono dal tipo di documenti
 - I sistemi commerciali operano prevalentemente sul testo, identificando le altre informazioni attraverso didascalie e note
 - I sistemi che operano sul contenuto di immagini, disegni geometrici e sequenze video ne esplorano la struttura e le proprietà visive
 - Video e audio richiedono algoritmi di pattern matching che si estendono nel tempo e presentano un elevato grado di incertezza

Documenti e IR (3)

- I documenti che esistono su carta vengono solitamente digitalizzati tramite scanner per poter essere memorizzati in formato digitale
 - Ogni pezzo/parte di un documento viene opportunamente identificato e isolato
 - Testo
 - Figure
 - Didascalie
 - Perché la parte testuale sia elaborabile come tale è necessario processare il testo digitalizzato tramite software OCR (Optical Character Recognition)

Surrogati

- Alcuni SRI ci permettono di ricercare nel loro contenuto testuale, altri solo nei titoli, altri in titolo e abstract, ecc...
- Questo perché di uno stesso documento (e.g., libro) si può considerare una diversa rappresentazione “derivata”
- Queste rappresentazioni derivate sono dette surrogati

I surrogati (1)

- Un surrogato è una rappresentazione limitata di un documento intero.
 - L'uso dei surrogati implica una conoscenza incompleta del documento.
- Se i sistemi di IR sono basati sui surrogati, nasce il problema di dare valutazioni sulla base di informazioni incomplete.
 - Ad esempio, se il surrogato in questione è il titolo del documento, come è possibile giudicarne il contenuto?
- I principali surrogati sono:
 - identificativo del documento;
 - chiavi: parole chiave, frasi chiave;
 - Sommario;
 - Estratto;
 - revisione.
- A seconda del surrogato utilizzato, si avrà una conoscenza più o meno completa del documento e quindi valutazioni diverse delle interrogazioni.

Identificativo del documento

- In genere un documento viene associato ad un identificativo (codice).
 - Può essere semplice e poco significativo per l'utente.
 - gli identificativi assegnati dalle biblioteche
 - Può essere più elaborato e permettere di “inserire” (e riconoscere) il documento all'interno di una certa struttura.
 - un identificativo composto da (abbreviazioni di) autore, collezione, armadio in cui è collocato.
- In generale comunque gli identificativi forniscono poca (o nessuna) informazione sul documento.
 - spesso l'identificativo è accompagnato da altre informazioni utili (e.g., titolo, autore, editore, ...).

Chiave

- Insieme di parole (o frasi), scelte dall'autore o dall'editore, che permettono di rappresentare sinteticamente il contenuto di un documento.
 - le parole chiave sono spesso usate negli articoli di ricerca al fine di catalogare velocemente l'articolo nel suo ambito di ricerca e le principali idee che contiene.

Sommario

- Il sommario (o abstract) è una brevissima descrizione (normalmente scritta dall'autore stesso) del contenuto di un documento.
 - Usato per articoli di ricerca e tesi.
 - Se ben scritto permette all'utente di capire se un certo documento può essere o meno interessante.
 - Se il sommario riprende “solo” l'inizio del documento allora potrebbe non dare sufficienti informazioni.

Estratto

- Consiste in frasi prese dal documento
 - Vi sono vari metodi per la sua costruzione
 - sarà più o meno significativo a seconda delle frasi scelte
- E' creato da qualcuno diverso dall'autore del documento.

Revisione

- Simile ad un sommario (anche se in genere può essere più lungo), ma viene scritto da qualcuno diverso dall'autore del documento
 - il sommario è solo descrittivo
 - una revisione contiene anche dei commenti (critiche o giudizi) sul contenuto del documento

Full text

- L'ultimo “surrogato” è poi il documento originale stesso in tutte le sue parti
 - Compreso tutto il suo contenuto
- Un SRI che lavora su tutto il documento originale e permette la ricerca all'interno di tutto il suo contenuto si dice essere “full text”

Architettura funzionale di un sistema di IR

- Mettendo tutto insieme:
 - la ricerca può avvenire su surrogati dei documenti stessi, e la qualità del risultato dipende dall'accuratezza dei surrogati
 - i documenti/surrogati sono rappresentati dall'insieme di termini che li identificano
 - le interrogazioni esprimono condizioni sui termini attraverso cui si vogliono ricercare i documenti/surrogati

