Sistemi per il recupero delle informazioni

Gabriele Pozzani

A.A. 2012/2013

Corso di Laurea Magistrale in Editoria e Giornalismo

Introduzione al corso

Informazioni generali (I)

- Orario (a partire dall'8 novembre)
  - 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

Statistiche esiti			
Esiti Esami	**Esiti Percentuali**	**Media voti**	**Deviazione Standard**
Positivi	57.69%	26	3
Respinti	7.14%		
Assenti	19.23%		
Ritirati	14.83%		
Annullati	1.09%		

Distribuzione degli esiti positivi													
18	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**30 e Lode**
0.0%	3.8%	2.8%	0.9%	9.5%	5.7%	3.8%	6.6%	10.4%	8.5%	1.9%	12.3%	25.7%	7.6%

Valori relativi all'AA 2011/2012 calcolati su un campione di **182 iscritti**. I valori in percentuale sono arrotondati al numero intero più vicino.

- 182 iscrizioni da parte di 118 diversi studenti
- Tolti i 20% di esami annullati (a causa di errori nella compilazione del registro) o assenti
  - Il 72.41% (105 su 145) degli esami consegnati sono positivi
  - Di questi il 56% (59 su 105) prende almeno 27
- Il 59% dei ritirati (16 su 27) sono in realtà voti positivi rifiutati dagli studenti (che poi hanno confermato o migliorato il voto)

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.
    - Alcuni capitoli e slide originali 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://www.ims.uni-stuttgart.de/ir/
  - 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. 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
7. Presentazione di alcuni software e sistemi reali per il recupero dell'informazione

Perché questo corso?

- Perché per fare bisogna prima conoscere
  - E se si conosce di solito si fa/lavora meglio
- Perché il mondo (e Internet) è pieno di informazioni, ma bisogna saperle trovare
- Perché avere conoscenze approfondite in un ambito porta a poter introdurre nuove tecnologie, metodologie e strategie in altri settori con cui si può entrare in contatto successivamente
  - Aiutare ad innovare l'azienda in cui si lavorerà

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

Scomposizione di un SI

- Un SI è composto da due parti principali:
  - Sistema esterno (ectosystem)
  - Sistema interno (endosystem)

Scomposizione di un SI: ectosystem

Comprende tutti i fattori che non possono essere controllati dal progettista del sistema informativo:

- PERSONE
  - utente: persona che usa il sistema per memorizzare o recuperare informazioni
  - finanziatore: persona (o organizzazione) che sostiene i costi relativi al sistema
  - operatore: persona che presta servizi all'utente
- FORMATO in cui le informazioni sono disponibili. L'organizzazione deve gestire delle informazioni che possono avere formati diversi, e.g., cartaceo, diversi formati digitali.
- TECNOLOGIA disponibile per il sistema.

Scomposizione di un SI: endosystem

Comprende tutti i fattori che il progettista del sistema informativo può controllare e definire:

- il SUPPORTO usato per memorizzare le informazioni (e.g., stampe, mappe, nastri magnetici, CDROM).
- le PROCEDURE usate per processare le informazioni (e.g., scaffali, scanner,. . .).
- gli ALGORITMI usati per processare o recuperare le informazioni.
- le STRUTTURE DATI usate per organizzare le informazioni.

N.B.: in molti casi, il supporto vincola le procedure.

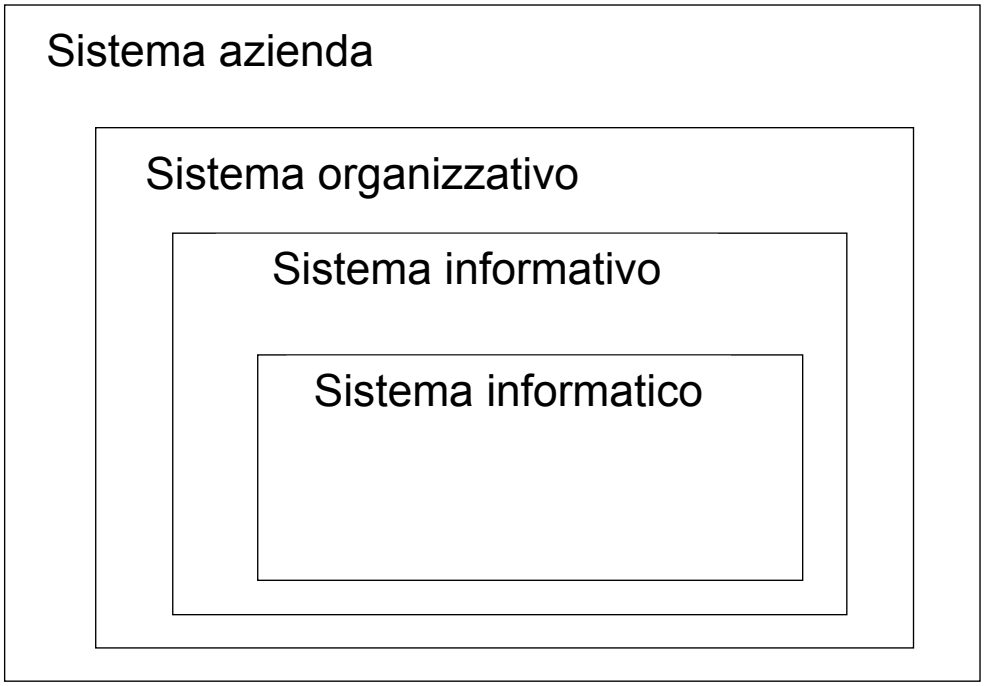
Sistemi informativi e automazione

- 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

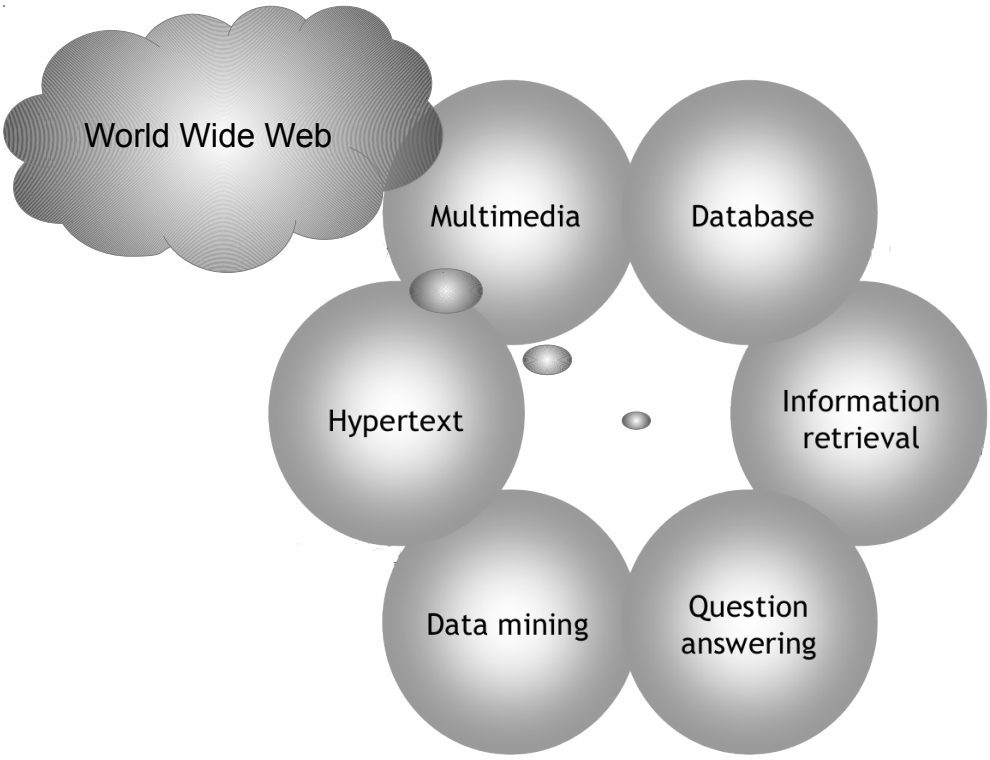
- che cosa significano questi numeri?
- cartelli stradali in Finlandia
- ma la differenza?
  - senza "interpretazione" il dato serve a ben poco

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
  - generalmente fuori linea
- 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

IR multimediale vs IR tradizionale

- L'informazione multimediale è strutturalmente semplice ma il significato è complesso e a volte sfumato
  - la struttura dipende dai media (immagini, video, audio)
  - la dimensione dei dati è importante
  - il significato è esterno ai dati
- Le interrogazioni dipendono da livelli di linguaggi "intermedi" che possono essere complessi
  - bassa (o nessuna) corrispondenza tra forma e significato di un oggetto
  - query by example
  - similarity search vs. corrispondenza esatta (booleana)
- Interazione utente
  - raffinamento, relevance feedback
  - browsing visuale (immagini)
  - sommario sintetico (video)
  - audio?

Modelli per dati semistrutturati

I dati gestiti all'interno di siti Web presentano una struttura irregolare

Si introduce la nozione di dato semistrutturato

- lo schema non è definito in anticipo ma implicito nella struttura dei dati
- lo schema è ampio e può cambiare nel descrivere i dati che provengono da fonti diverse
- lo schema è descrittivo (specifica i dati come sono) anziché prescrittivo (specifica i dati come dovrebbero essere)
- i dati non sono tipizzati in modo forte, e la tipizzazione può cambiare da istanza a istanza

I dati semistrutturati sono adatti per

- lo scambio di oggetti tra sorgenti eterogenee
- la gestione di collezioni di documenti
- la rappresentazione di proprietà specifiche (es. link interni/esterni)

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`

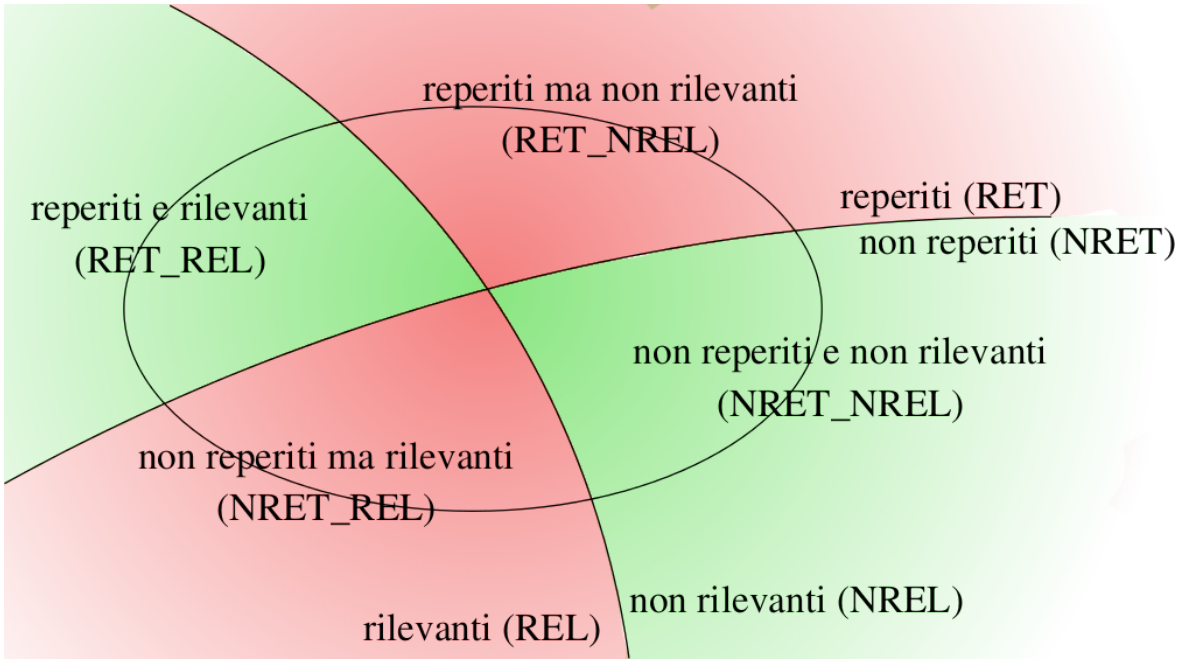
Il problema dell'IR (2)

- 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

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 a seconda del grado di rilevanza rispetto all'interrogazione dell'utente
  - Recuperare tutti i documenti rilevanti scartando il maggior numero possibile di documenti non rilevanti

Lo spazio dei documenti



Le possibili ricerche utente

- 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
  - Comunicazioni informali
    - Lettere
    - Messaggi
  - Informazioni codificate
    - File
    - e-mail
    - dati numerici e tabelle
    - immagini e disegni
    - suoni e voci

Documenti e IR (2)

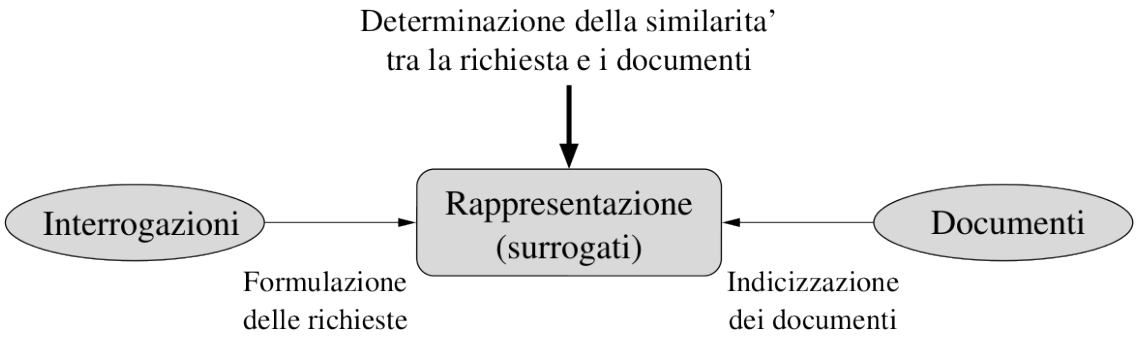
- Le tecniche di IR dipendono dal tipo di documenti
  - I sistemi commerciali operano prevalentemente sul testo, identificando le altre informazioni attraverso didascalie e note
  - I sistemi che operano sul contenuto di immagini, disegni geometrici e sequenze video ne esplorano la struttura e le proprietà 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)

Architettura funzionale di un sistema di IR

- Un sistema di IR non opera direttamente sui documenti
  - i documenti sono rappresentati dall'insieme di termini che lo identificano
  - le interrogazioni esprimono condizioni sui termini attraverso cui si vogliono ricercare i documenti
  - la ricerca può avvenire su surrogati dei documenti stessi, e la qualità del risultato dipende dall'accuratezza dei 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 identif cativo (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 identif cativi 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 f ne 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