Sistemi per il recupero delle informazioni

Lezione 1: Introduzione ai sistemi informativi e ai sistemi di recupero delle informazioni

Gabriele Pozzani

Corso di Laurea Magistrale in Editoria e Giornalismo
Anno Accademico 2011/2012
Università degli Studi di Verona, Italia

Indice

Versione 1.0 pubblicata online il 27/12/2011.

1 Introduzione ai sistemi informativi

Ogni organizzazione ha una ben precisa struttura interna (e.g., gerarchie, suddivisione in dipartimenti, centri di costo). In ogni caso ogni organizzazione ha un sistema informativo, eventualmente non esplicitato nella struttura, che spesso è di supporto ai vari sottosistemi dell'organizzazione stessa. Potendo essere sfruttato da diversi sottosistemi di un'organizzazione, un sistema informativo va studiato nel contesto in cui è effettivamente inserito.

Definizione 1 (Sistema Informativo)**.** *Il* Sistema Informativo *è la componente (sottosistema) di una organizzazione che gestisce le informazioni di interesse.*

Con gestione si intende:

- raccolta e acquisizione (recupero delle informazioni da fonti interne e/o esterne all'organizzazione stessa);
- archiviazione e conservazione (il mantenimento delle informazioni in appositi sistemi, e.g., databases);
- elaborazione, trasformazione;
- produzione (produzione di nuove informazioni a partire da informazioni preesistenti o la loro creazione "da zero");
- distribuzione, comunicazione e scambio (all'interno e/o all'esterno dell'organizzazione)

Definizione 2 (Informazione di interesse)**.** *Sono tutte quelle informazioni utilizzate dall'organizzazione per il conseguimento dei propri scopi.*

Un sistema informativo può, a sua volta, essere suddiviso in sottosistemi (in modo gerarchico o decentrato), più o meno fortemente integrati.
Un sistema informativo può essere scomposto in due parti principali:

sistema esterno (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. È compito del progettista "adattarsi" alle esigenze dell'organizzazione e al formato delle informazioni per progettare il sistema informativo richiesto.

→ TECNOLOGIA disponibile per il sistema.

sistema interno (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. Ad esempio, se si sceglie di memorizzare le informazioni su nastri magnetici la procedura di lettura dovrà inevitabilmente

essere ad accesso sequenziale[1]. D'altro canto l'uso dei CDROM permette un accesso "random"[2] alle informazioni.

Notiamo che finora si è parlato di sistema informativo in modo molto generale, senza fare alcun riferimento all'uso di termini legati all'informatica e limitare il campo all'uso di PC, come invece si potrebbe pensare di fare subito. Il pensare le informazioni collegate e gestite dall'attuale tecnologia dei computer è un errore "comune". Notiamo però che il concetto di "Sistema Informativo" è indipendente da qualsiasi automazione. Esistono infatti organizzazioni la cui ragion d'essere è la gestione delle informazioni (e.g., servizi anagrafici e banche) e che operano da secoli anche senza l'ausilio dei computer.

Definizione 3 (Sistema Informatico)**.** *È la porzione automatizzata del sistema informativo. In altre parole è la parte del sistema informativo che gestisce informazioni sfruttando tecnologie informatiche.*

Sistema informativo e informatico sono a loro volta parte del sistema organizzativo.

Definizione 4 (Sistema Organizzativo)**.** *È l'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.

Riassumendo, in ogni azienda può essere identificato il sistema organizzativo, una cui parte è adibita alla gestione delle informazioni, il sistema informativo. La parte automatizzata e basata sull'utilizzo di tecnologie informatiche del sistema informativo è detta sistema informatico. La struttura "gerarchica" dei diversi sistemi di un'azienda che abbiamo introdotto è esemplificata in Figura 1.

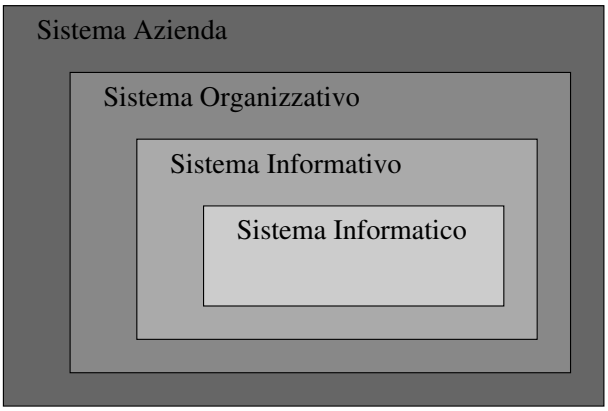

Figura 1: Organizzazione gerarchica dei sistemi di un'azienda

[1] http://it.wikipedia.org/wiki/Accesso_sequenziale
[2] http://it.wikipedia.org/wiki/Accesso_casuale

2 Dati e 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.

Data la generale tendenza a standardizzare le attività dei sistemi informativi complessi, sono state introdotte con il tempo forme di organizzazione e codifica delle informazioni. Ad esempio, nei servizi anagrafici si è iniziato con registrazioni discorsive delle informazioni anagrafiche per poi passare a registrazioni più schematiche (e.g., nome e cognome, data di nascita, codice fiscale).
A dispetto del nome, i sistemi informativi in realtà sono studiati per gestire dati. Ma quale è la differenza tra dato e informazione? Nel vocabolario della lingua italiana si trovano le seguenti definizioni.

Definizione 5 (Informazione)**.** *Un'*informazione *è ogni notizia o elemento che consente di avere conoscenza più o meno esatta di fatti, situazioni, modi di essere.*

Definizione 6 (Dato)**.** *Un* dato *è ciò che è immediatamente presente alla conoscenza, prima di ogni elaborazione.*

In altre parole, l'informazione è conoscenza, mentre il dato è l'elemento, il valore che trasporta/rappresenta l'informazione. Si può quindi dire che l'informazione sia un dato più il suo significato, la sua interpretazione.
Ad esempio, se su un foglio di carta leggiamo `Mario 2075`, questi sono due dati che però non significano molto e non apportano quindi nessuna conoscenza. Se il foglio di carta è fornito in risposta alla domanda "A chi mi devo rivolgere per il problema X, e quale è il suo numero di telefono?" allora i dati possono essere interpretati e fornire così informazione e arricchire la conoscenza.
Nei sistemi informatici (e non solo) le informazioni vengono rappresentate in modo essenziale attraverso i dati. Questo perché la rappresentazione precisa di forme più ricche di informazione e conoscenza è difficile. I dati permettono una rappresentazione più concisa ed efficiente delle informazioni. L'interpretazione dei dati è invece resa implicita o memorizzata separatamente.

3 Tipi di sistemi informativi

Esistono diversi tipi di sistemi informativi:

- database
- ipertesti
- data mining
- question answering
- information retrieval

I diversi tipi di sistemi informativi possono essere caratterizzati in base

- alla natura e alle caratteristiche delle informazioni da gestire;
- al tipo e alla natura delle interrogazioni che possono essere eseguite sulle informazioni;
- a come avvengono gli aggiornamenti alle informazioni;
- ai problemi specifici che devono essere affrontati per la corretta gestione delle informazioni.

Caratterizzeremo ora alcuni tipi di sistemi informativi in base a questi quattro fattori.

3.1 Database

I sistemi di database si occupano della memorizzazione delle informazioni di interesse di un'organizzazione. I database sono studiati quindi per gestire informazioni

- semplici nella loro natura: i dati di interesse per un'azienda sono nomi, valori numerici, date, e altri dati simili la cui natura è di semplice definizione;
- potenzialmente molto complesse nella struttura: tra le informazioni vi sono interrelazioni e dipendenze che devono essere modellate perché le informazioni siano mantenute correttamente;
- di elevate dimensioni: un'azienda può produrre grandi quantità di dati ogni giorno (si pensi per esempio ad una banca e alle sole informazioni sulle transazioni) e deve poterlo fare per lunghi periodi di tempo, anni, decenni.

I dati memorizzati in un database sono poi usati dall'azienda per il suo quotidiano funzionamento. Questo richiede di poter recuperare i dati tramite query (interrogazioni) che normalmente sono:

- anche complesse: richiedono di incrociare dati e informazioni diverse e per farlo possono usare interrogazioni annidate e join (unione di tabelle diverse contenenti dati diversi)
- ripetitive: i lavori che un'azienda svolge quotidianamente sono infatti sempre gli stessi. Le query possono quindi essere classificate, precompilate e standardizzate, lasciando agli utenti solo la possibilità di scegliere il valore di pochi parametri.

Essendo i database alla base del funzionamento quotidiano dell'organizzazione, i dati che contiene vengono aggiornati frequentemente (ogni ora, ogni minuto o meno) e in modo casuale (irregolarmente) nel tempo. Questi aggiornamenti devono poi poter avvenire "in linea", cioè senza che questo comporti il blocco del lavoro in contemporanea di altre componenti dell'organizzazione che necessitino dei dati.

Tutte queste considerazioni e l'importanza dei dati per l'organizzazione, fanno sì che i database debbano essere progettati per gestire:

- la sicurezza dei dati: accesso controllato e limitato ai dati;
- l'integrità dei dati: accessi contemporanei (in lettura e/o scrittura) ai dati devono poter essere eseguiti senza che i dati vengano poi a trovarsi in uno stato inconsistente con la realtà e con ciò che gli utenti si aspettano dopo gli accessi;
- l'efficienza: una grande quantità di accessi ad una grande mole di dati richiede che, perché il lavoro dell'organizzazione si svolga bene, il database sia progettato al meglio e che tutte le operazioni su di esso siano eseguite nel modo più semplice ed efficiente possibile.

3.2 Ipertesti

I sistemi ipertestuali (di cui sono un esempio le classiche pagine Web) rispondono ad esigenze e a compiti diversi da quelli dei database. Le informazioni mantenute sono in generale:

- complesse, multiformi e non codificate nel loro significato;
- con livelli di strutturazione molto variabili: le relazioni tra le informazioni possono essere diverse, di diversa natura, con significati diversi e assolutamente non gerarchiche.

Nei sistemi ipertestuali non si hanno delle vere interrogazioni. L'accesso ai dati avviene tramite la navigazione e l'esplorazione libera da parte dell'utente delle informazioni. Per questo le "interrogazioni" non sono prevedibili.
Nei sistemi ipertestuali i dati vengono però, in generale, aggiornati poco frequentemente (sicuramente meno di quanto accade in media in un database). Inoltre molti aggiornamenti avvengono fuori linea, cioè mentre i dati non sono accessibili ad altri utenti.
Problemi specifici da tenere presente nella progettazione di un sistema ipertestuale sono quindi:

- l'interazione degli utenti con il sistema;
- l'usabilità (la facilità d'uso) da parte degli utenti del sistema;
- la portabilità, cioè la possibilità che il sistema sia indipendente dalla piattaforma da cui viene acceduto (sistema operativo, browser).

3.3 Sistemi di information retrieval per il testo

I sistemi di information retrieval possono essere sviluppati sia per gestire informazioni testuali che informazioni multimediali (immagini, audio, video).
Nel caso dei sistemi per la gestione di testi le informazioni da gestire sono in generale:

- di natura semplice (ad es.: autori, argomenti, riferimenti);
- poco strutturate: i testi hanno una strutturazione semplice (capitoli, sezioni) e limitata nella profondità; inoltre possono contenere tabelle, la cui struttura è però facilmente comprensibile ed analizzabile;
- molto numerose: il sistema deve poter gestire centinaia o migliaia di testi (si pensi per esempio al numero di testi in una classica libreria o biblioteca), ognuno con una dimensione non trascurabile.

Sui testi gli utenti possono eseguire query:

- anche complesse: più clausole (richieste) collegate tra loro;
- basate su specifiche o informazioni parziali (ricerche approssimate);
- tramite iterazioni successive: l'utente può eseguire una query sul risultato di una query precedente, raffinando quindi la ricerca.

Gli aggiornamenti ai dati testuali (inserimento di nuovi testi o modifica di dati su testi già presenti) sono, in generale:

- periodici e a bassa frequenza;
- fuori linea

Al fine di poter recuperare informazioni testuali tra una grande quantità di documenti, i sistemi di information retrieval testuali devono tenere conto e fare uso delle seguenti "idee":

- uso di dizionari per riconoscere i termini;
- uso di ontologie e thesauri per collegare termini "appartenenti" allo stesso argomento o comunque in relazione tra loro.

3.4 Sistemi di information retrieval multimediali

Nei sistemi di information retrieval multimediali le informazioni da gestire e ricercare sono in forme diverse da quella testuale: audio, immagini, video. La diversa natura delle informazioni rende necessario l'uso di tecnologie e idee diverse dal caso testuale e porta con sé problemi diversi.
Le informazioni multimediali hanno:

- natura semplice (una sequenza di pixel o frame) ma di significato complesso (una singola immagine può apportare informazioni e conoscenze che vanno oltre al semplice soggetto raffigurato);

- struttura variabile a seconda del tipo di informazione: ad esempio, un'immagine è strutturata come una sequenza di pixel ognuno con la sua struttura interna per definirne il colore, un video è, nel caso più semplice, una sequenza di frame (immagini) ognuno strutturato come un'immagine;
- elevate dimensioni: audio, video e immagini di buona o alta qualità hanno dimensioni considerevoli, che possono arrivare a centinaia di Megabyte o Gigabyte.

La natura complessa delle informazioni multimediali fa sì che le interrogazioni:

- abbiano una scarsa corrispondenza tra forma (come l'utente formula l'interrogazione) e significato dell'interrogazione stessa;
- possano essere eseguite tramite specifiche "sintattiche" o tramite una ricerca per somiglianza (ad esempio, cercare immagini che per colore, dettagli, soggetto, "forma" del soggetto siano simili ad una data);
- possano essere eseguite tramite iterazioni successive in cui l'utente raffina i risultati della ricerca eseguendo un'analisi di rilevanza di quanto è stato recuperato.

Gli aggiornamenti alle informazioni multimediali sono, in generale:

- non frequenti
- fuori linea

Ancora la natura complessa delle informazioni multimediali porta con sé problemi specifici che devono essere considerati in un sistema di information retrieval multimediale:

- la rappresentazione codificata (sequenza di pixel, sequenza di immagini, sequenza di byte) non contiene direttamente il significato del dato;
- la codifica e la rappresentazione influiscono sui sistemi di gestione.

4 Information retrieval (IR)

I sistemi di Information Retrieval (in italiano: sistemi per il recupero delle informazioni) sono sistemi software che supportano la rappresentazione, la memorizzazione, l'organizzazione e il reperimento di informazioni non strutturate in archivi di grandi dimensioni, sulla base di criteri di classificazione e ricerca esatti, basati sull'identificazione del contenuto informativo attraverso l'utilizzo controllato del linguaggio naturale. Applicazioni tipiche dei sistemi di IR sono:

- motori di ricerca;
- ricerche bibliografiche;
- ricerca documentaria;
- consultazione di archivi giuridici e normativi;
- catalogazione di oggetti eterogenei;
- archiviazione di documenti in prosa;
- analisi letteraria e linguistica.

Si possono individuare due tipi principali di ricerche in un sistema di IR:

querying: ricerca mirata in funzione del contenuto basata su

- classificazione argomentale dei documenti;
- strutturazione del contenuto;
- indicizzazione;
- dizionari dei sinonimi;
- lemmatizzazione.

browsing: ricerca esplorativa basata sulla navigazione "libera" dei documenti tramite ricerche incrementali, tendenti ad ottenere approssimazioni successive dell'insieme di documenti da recuperare.

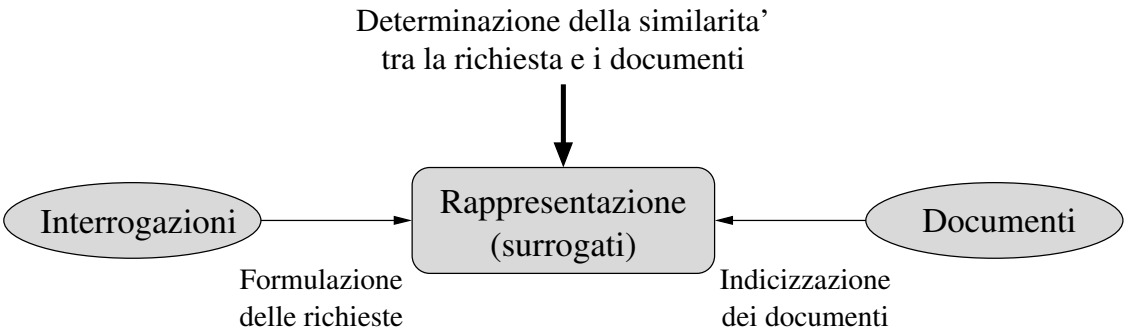

Figura 2: Architettura funzionale di un sistema di IR

L'oggetto base su cui i sistemi di IR lavorano sono i "documenti". Un documento è una qualsiasi collezione di informazioni rintracciabili in base alla descrizione del suo contenuto. Tipici esempi sono: testi in prosa, dati numerici e tabelle, immagini e disegni, suoni e voci.
I sistemi di IR commerciali operano prevalentemente sul testo, identificando le altre informazioni tramite didascalie e note. Diversamente, 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 complessità e di incertezza.

4.1 Architettura dei sistemi di IR

Per motivi di efficienza e complessità, un sistema di IR può non operare, in generale, però direttamente sui documenti:

- un documento viene elaborato al fine di ottenere una sua rappresentazione "limitata", detta *surrogato* (ad esempio, l'insieme dei termini che li identificano);
- le interrogazioni esprimono condizioni sui termini attraverso cui si vogliono ricercare i documenti;
- la ricerca avviene sui surrogati dei documenti stessi, e la qualità del risultato dipende dall'accuratezza dei surrogati.

L'architettura funzionale di un sistema di IR è rappresentata in Figura 2.

4.2 Documenti

L'oggetto base di un sistema di IR è un documento. La maggior parte dei documenti disponibili sono testi scritti in linguaggio naturale.
Nel contesto dell'IR un documento è formato da un insieme di dati memorizzati in qualche forma. Sono quindi documenti:

- libri;
- articoli stampati;
- comunicazioni informali (lettere, messaggi);
- informazioni codificate (file, e-mail, documenti digitali, immagini, suoni, ipertesti).

Come abbiamo in parte già visto, le tecniche di recupero dell'informazione dipendono dal tipo di documento che si ha a disposizione.

4.2.1 Documenti su carta

I documenti cartacei vengono solitamente digitalizzati e memorizzati in formato digitale per mezzo di scanner. Ogni pezzo di documento viene opportunamente identificato e isolato: testo (ana-

lizzato tramite OCR, Optical Character Recognition[3]), figure (memorizzate come immagini), didascalie (associate alle figure che accompagnano).

4.2.2 Documenti di testo

La rappresentazione base di questi testi è costituita da stringhe di caratteri che includono simboli di punteggiatura, spazi e altri simboli utili a dare più struttura al testo.
Un testo è solitamente composto di capitoli, paragrafi e sezioni e tale scomposizione aiuta la comprensione del documento. La scomposizione non è comunque definita a priori e non è detto che sia presente.
Un modo per fornire una struttura ai documenti di testo è l'uso di linguaggi di marcatura.
I marcatori consentono di definire la semantica del contenuto del documento. Ad esempio:

TESTO	TESTO CON MARCATURA
Sistemi per il recupero delle informazioni Gabriele Pozzani	<corso> Sistemi per il recupero delle informazioni </corso> <docente> Gabriele Pozzani </docente>

Tipici linguaggi di marcatura sono XML ed HTML. In particolare HTML (HyperText Markup Language) è utilizzato per costruire pagine Web. In questo caso i marcatori servono a definire lo stile di presentazione delle varie parti del testo. Ad esempio `<b>Gabriele</b>` indica che la parola `Gabriele` dovrà essere visualizzata in grassetto (bold).

4.3 I surrogati

Un surrogato è una rappresentazione limitata di un documento intero. L'uso dei surrogati quindi implica una conoscenza incompleta del documento. Dovendo poi il sistema di IR basarsi sui surrogati per valutare le interrogazioni sottoposte dall'utente, 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 giudicare il contenuto del documento dal solo titolo? Può al più essere possibile valutare se il documento è "interessante" o meno, sempre che il titolo non sia fuorviante.
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.

4.3.1 Identificativo del documento

In genere un documento viene associato ad un identificativo (codice). Questo può essere semplice e poco significativo per l'utente. Si consideri ad esempio, l'identificativo assegnato ad ogni libro da una biblioteca.

[3] http://it.wikipedia.org/wiki/Riconoscimento_ottico_dei_caratteri

In altri casi può essere un identificativo più elaborato che permette di "inserire" (e riconoscere) il documento all'interno di una certa struttura. Ad esempio, una biblioteca può assegnare 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. Per questo, spesso l'identificativo è accompagnato da altre informazioni utili (e.g., titolo, autore, editore, ...).

4.3.2 Chiave

Una chiave è insieme di parole (o frasi), scelte dall'autore o dall'editore, che permettono di rappresentare sinteticamente il contenuto di un documento.
Ad esempio, 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.

4.3.3 Sommario

Il sommario (o abstract) è una brevissima descrizione (normalmente scritta dall'autore stesso) del contenuto di un documento. È ad esempio usato per articoli di ricerca e tesi.
Un sommario 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.

4.3.4 Estratto

L'estratto consiste in frasi prese dal documento. Vi sono vari metodi per la sua costruzione. Comunque si deve notare che l'estratto è creato da qualcuno diverso dall'autore del documento. Inoltre, l'estratto sarà più o meno significativo a seconda delle frasi scelte.

4.3.5 Revisione

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

4.4 Formulazione delle interrogazioni

L'IR è storicamente nata per soddisfare le richieste effettuate dagli utenti interessati a reperire informazioni rilevanti rispetto alle loro esigenze. Queste richieste sono formulate tramite interrogazioni (query) al sistema di IR.
Per rispondere ad un'interrogazione, ogni sistema di IR si basa su una tecnica di recupero delle informazioni, ovvero su un meccanismo che permette di confrontare la richiesta, espressa in uno specifico linguaggio (si veda la parte del corso riguardante i tipi di query), con le rappresentazioni dei documenti (surrogati). La tecnica utilizzata dipende dal tipo di documenti e dal tipo di query permesso dal sistema di IR (si veda la parte del corso riguardante i tipi di matching).
A seguito di un'interrogazione, il sistema segnala i documenti ritrovati e il loro numero. L'utente può riformulare, specializzare o generalizzare l'interrogazione fino ad ottenere un insieme soddisfacente di documenti ritrovati.
L'esame dei documenti si può avvalere di due funzionalità:

- ranking: i documenti sono presentati all'utente in ordine decrescente di rilevanza secondo i pesi assegnati ai termini. Questo è il metodo usato solitamente nei motori di ricerca sul Web.
- browsing: i documenti sono raggruppati in classi di somiglianza, permettendo all'utente di "sfogliarli" secondo un ordine logico.

5 Criteri di Efficacia

In un sistema di IR è importante considerare non solo l'*efficienza*, ovvero come il sistema si comporta in termini di tempi di risposta, uso della memoria, ecc..., ma anche l'*efficacia* (effectiveness), ovvero quanto il sistema è in grado di soddisfare l'utente fornendogli le (sole) informazioni rilevanti e semplificando il più possibile la sua attività di indagine conoscitiva.
Il problema di valutare l'efficacia di un sistema di IR non è di facile soluzione, in quanto include diversi aspetti soggettivi. Ad esempio, due utenti con diversi livelli di conoscenza a priori, formulando la stessa richiesta, possono fornire diverse valutazioni sull'insieme di documenti reperiti dal sistema.
Al fine di misurare l'efficacia dei sistemi di IR si introducono degli indici di valutazione. I due principali indici sono il *richiamo* (recall) e la *precisione* (precision). Entrambi si basano sull'idea di partizionare lo spazio dei documenti in archivio (cioè l'insieme dei documenti su cui eseguire le interrogazioni).

5.1 Partizionamento dello spazio dei documenti

Supponiamo di aver fissato il sistema di IR e di avere una query. Allora, l'insieme dei documenti si può partizionare in quattro sottoinsiemi, dove la classificazione di un documento considera se lo stesso è rilevante o meno (rispetto alla query data) e se è stato reperito o meno dal sistema di IR. In altre parole l'insieme dei documenti può essere diviso tramite due classificazioni indipendenti e ortogonali:

1. divisione tra documenti rilevanti (REL) e non rilevanti (NREL), rispetto alla query. Attenzione che stiamo ora considerando la rilevanza "teorica" dei documenti, cioè non la divisione tra i documenti che il sistema di IR considera rilevanti o meno, ma la divisione tra documenti rilevanti o meno da un punto di vista "esterno" e di più alto livello di quello del sistema di IR (questo punto di vista potrebbe essere considerato quello dell'utente);
2. divisione tra documenti reperiti (RET) e non reperiti (NRET) dal sistema di IR.

Essendo le due classificazioni indipendente e ortogonali, si ottengono quattro possibili sottoinsiemi disgiunti. Questo partizionamento e il significato dei quattro sottoinsiemi è rappresentato in Figura 3.

- i documenti in RET_REL sono i documenti che sono rilevanti per l'utente e vengono correttamente identificati come tali e ritornati dal sistema di IR;
- i documenti in RET_NREL sono i documenti che sono ritornati dal sistema di IR nonostante non siano rilevanti. Essi costituiscono del "rumore" che ogni sistema di IR dovrebbe cercare di ridurre al minimo. Tali documenti sono anche detti *false drops*, *false alarms* o *false hits*;
- i documenti in NRET_REL sono i documenti che pur essendo rilevanti non vengono riconosciuti come tali dal sistema di IR e quindi non vengono ritornati. In altre parole, sono i documenti per cui il sistema è "silenzioso". Essendo documenti rilevanti che però non vengono restituiti all'utente, anche questi dovrebbero essere ridotti al minimo. Tali documenti si chiamano anche *false dismissals*;

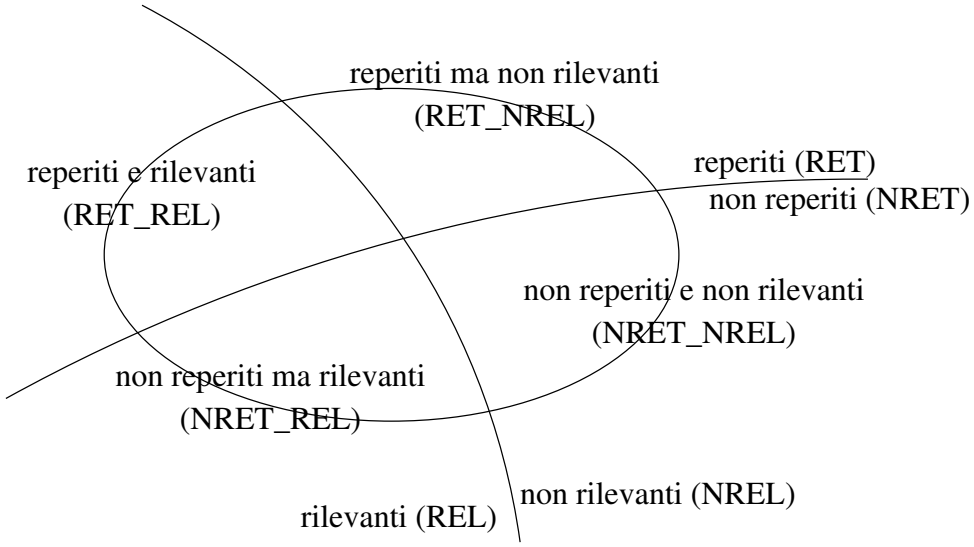

Figura 3: Partizionamento dello spazio dei documenti.

- i documenti in NRET_NREL sono i documenti che non sono rilevanti, che il sistema identifica correttamente come tali e quindi non vengono ritornati. In altre parole, sono i documenti giustamente da omettere nel risultato.

La situazione dal punto di vista dell'utente che fornisce la richiesta e che osserva la risposta fornita dal sistema di IR è rappresentata nella seguente tabella.

	Doc. rilevanti (REL)	Doc. non rilevanti (NREL)
Doc. reperiti (RET)	Corretti	Inesatti
Doc. non reperiti (NRET)	Omessi	Da omettere

In un sistema di IR *ideale* gli errori dovrebbero essere completamente assenti, cioè non si dovrebbero avere né false hits né false dismissals.

5.2 Precisione e richiamo

Basandosi sul partizionamento dello spazio dei documenti introdotto nella sezione precedente è possibile definire le due più comuni misure (cioè indici numerici) per quantificare l'efficacia dei sistemi di IR: richiamo (recall) e precisione (precision)[4].

$$\text{richiamo:} \quad R = \frac{\#RET\_REL}{\#REL} \qquad\qquad \text{precisione:} \quad P = \frac{\#RET\_REL}{\#RET}$$

Il richiamo è la proporzione dei documenti rilevanti che sono effettivamente recuperati. Esso misura la capacità del sistema di trovare i documenti rilevanti, cioè la capacità di ridurre il numero di false dismissals.
La precisione è invece la proporzione di documenti recuperati che sono rilevanti. Essa quindi misura la capacità del sistema di rigettare i documenti non rilevanti, cioè, in altri termini, la capacità di ridurre il numero di false hits.

[4]# indica il numero di documenti nell'insieme. Ad esempio: $\#REL$ è il numero di documenti in REL, cioè il numero di documenti rilevanti.

Si noti che mentre la precisione si può calcolare esattamente a partire dal risultato di una query (infatti in quel caso si conoscono i valori esatti sia di $\#RET\_REL$ che di $\#RET$), così non è per il richiamo, che richiede di conoscere quanti sono i documenti rilevanti ($\#REL$) in *tutta* la collezione. Quest'ultima informazione è difficilmente calcolabile, specialmente su collezioni di documenti molto grandi (ad esempio, tutte le pagine Web considerate da un motore di ricerca) sia perché è "teorica" e di "alto livello", sia perché, come abbiamo visto e vedremo ancora più avanti nel corso, può dipendere dall'utente. La precisione è quindi una misura "esatta", mentre il richiamo è una misura "inesatta", calcolata approssimativamente.
Entrambe le misure hanno valori sempre compresi tra 0 e 1, dove lo 0 rappresenta il caso pessimo, mentre l'1 il caso ottimo. Infatti se il richiamo vale 1 allora il sistema è stato in grado di recuperare tutti i documenti rilevanti (e quindi non vi sono false dismissals). D'altro canto quando vale 0 (o un valore molto vicino allo 0), significa che il sistema ha recuperato pochi dei documenti rilevanti (i.e., vi sono molti false dismissals). Un ragionamento simile avviene per la precisione. Quando essa vale 1, ancora, i documenti ritornati sono tutti e solo quelli rilevanti (i.e., non vi sono false hits), mentre quando vale 0 (o un valore molto vicino allo 0) solo una piccola porzione dei documenti ritornati è rilevante (i.e., vi sono molti false hits).
Le due formule di richiamo e precisione sono collegate avendo lo stesso numeratore. Tramite alcuni semplici passaggi algebrici, è possibile ottenere la formula che lega precisione e richiamo:

$$P = R \times \frac{\#REL}{\#RET}$$

Oltre che a legare le due misure, questa formula permette anche di valutare l'efficacia di un sistema di IR misurando la precisione a diversi "livelli" di richiamo, ovvero:
quanti documenti bisogna reperire ($\#RET$) affinché il risultato contenga una frazione pari a R (il richiamo) dei documenti rilevanti presenti nella collezione ($\#REL$), supponendo di conoscere tale valore?
In altre parole è possibile decidere il numero (threshold) di documenti da recuperare per ottenere almeno l'$R\%$ dei documenti rilevanti. All'aumentare di questo threshold la precisione diminuisce (in quanto per essere sicuri di ottenere una certa percentuale di documenti rilevanti includeremo anche una maggiore quantità di documenti non rilevanti, false hits).

Precisione e richiamo sono utilizzate per misurare l'efficacia di diversi sistemi di IR e poterli confrontare. Per fare ciò, per ogni sistema di IR vengono eseguite un insieme prestabilito di interrogazioni e con il risultato di ciascuna si calcolano precisione e richiamo. Confrontando i valori delle due misure su queste interrogazioni "campione" di due sistemi di IR è possibile, ma non sempre, osservare se uno dei due sistemi è "migliore" dell'altro. Ad esempio, si considerino i due sistemi di IR, IRS_1 e IRS_2, su cui vengono eseguite le stesse query "campione" Q_1, Q_2, Q_3, Q_4, sempre ovviamente sulla stessa collezione di documenti. Si supponga di ottenere le seguenti coppie di valori (precisione, richiamo):

	Q_1	Q_2	Q_3	Q_4
IRS_1	(0.9 , 0.4)	(0.86 , 0.34)	(0.65 , 0.6)	(0.78 , 0.67)
IRS_2	(0.79 , 0.22)	(0.45 , 0.3)	(0.65 , 0.53)	(0.67 , 0.5)

Si può notare che i valori di precisione e richiamo di IRS_1 sono sempre più alti (e quindi migliori) di quelli di IRS_2. Si può quindi dire che "statisticamente" (se le query "campione" sono state scelte con criterio e in numero sufficiente) IRS_1 è migliore di IRS_2. Se i valori non fossero stati così facilmente confrontabili, cioè se per alcune query IRS_1 avesse valori migliori di IRS_2 mentre

per altre query valesse il viceversa, allora non si sarebbe potuto concludere nulla su quale dei due sistemi fosse il migliore.

Si osserva che vi è una legge (non un teorema, cioè qualcosa che non è dimostrabile essere sempre vero, ma che è vero in media, nella maggior parte dei casi) *inversa* tra precisione e richiamo. Legge inversa significa che all'aumentare della precisione cala il richiamo e viceversa. Tale legge può essere visualizzata per diversi sistemi nel cosiddetto *precision-recall graph*, dove sull'asse orizzontale (delle ascisse) vengono rappresentati i valori del richiamo mentre sull'asse verticale (delle ordinate) vengono rappresentati i valori della precisione. Un esempio di precision-recall graph per tre sistemi di IR è raffigurato in Figura 4.

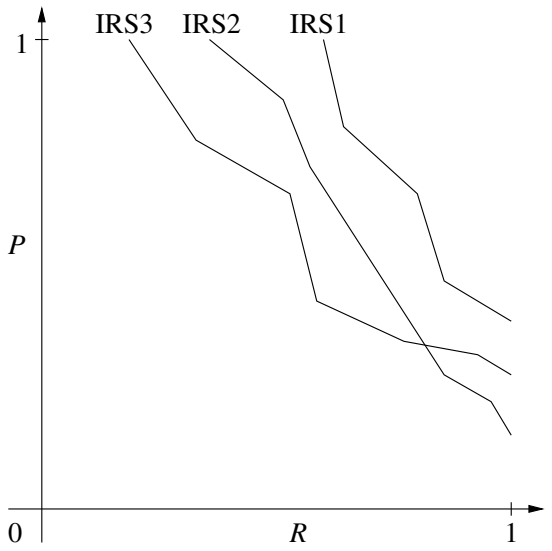

Figura 4: Esempio di precision-recall graph

Nel grafo i due sistemi IRS_1 e IRS_2 potrebbero essere i due precedentemente presi in considerazione nella tabella comparativa riportata sopra. Anche visivamente si vede che IRS_1 "sovrasta" (cioè è sempre al di sopra di) IRS_2 e può quindi essere considerato migliore. D'altro canto tra IRS_2 e IRS_3 non è possibile decidere in assoluto quale sia il migliore perché, come si vede, vi sono casi in cui IRS_2 è al di sopra di IRS_3, mentre in altri casi (i.e., altre interrogazioni) vale il viceversa.
Per capire intuitivamente perché tale legge vale si supponga di eseguire un sistema di IR in risposta ad una query. Poniamo che il sistema ritorni esattamente un solo documento tra tutti quelli della collezione e che questo documento sia rilevante (se non lo fosse si potrebbe dire che il sistema è proprio "da buttare" non essendo in grado di trovare nemmeno un documento rilevante). Allora in tal caso la precisione vale 1, dato che, dalla formula data per essa, si ha che $\#RET\_REL$ (il numero di documenti ritornati e rilevanti) è 1 così come $\#RET$ (il numero di documenti ritornati). D'altro canto il richiamo, anche se non possiamo calcolarlo esattamente, possiamo dire che sia basso, potendo ben supporre che i documenti rilevanti siano (molti) più di 1 soltanto. Quindi applicando la formula del richiamo si avrebbe che $\#RET\_REL$ è 1 mentre $\#REL$ è presumibilmente un numero maggiore di 1, ottenendo così una frazione piccola.
Considerando l'altro caso estremo in cui il sistema di IR ritorni tutti i documenti della collezione, il richiamo varrà 1 in quanto tra tutti i documenti tornati (cioè tutti i documenti della collezione) vi saranno anche tutti i documenti rilevanti. D'altro canto la precisione sarà bassa in quanto tra tutti i documenti tornati vi saranno anche tutti quelli non rilevanti.

Dato che, come si può immaginare, il sistema di IR ideale (cioè senza false hits e senza false dismissals) in generale non esiste, sorge la domanda se è da preferire un sistema che assicuri una maggiore precisione o uno che privilegi il richiamo. In altre parole, è più importante la precisione o il richiamo? Come si può ulteriormente immaginare, non esiste una risposta certa a questa domanda essendo una questione soggettiva. Eseguendo studi sociologici, Cleverdon[5] nel 1991 ha suggerito che gli utenti preferiscano una alta precisione ad un alto richiamo. Su[6] nel 1994 ha invece suggerito che nessuna delle due misure sia significativa per l'utente.

[5]Cyril W. Cleverdon. 1991. The significance of the Cranfield tests on index languages. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '91). ACM, New York, NY, USA, 3-12. DOI=10.1145/122860.122861

[6]L. T. Su. 1994. The relevance of recall and precision in user evaluation. *Journal of the American Society for Information Science*, 45: 207–217. DOI=10.1002/(SICI)1097-4571(199404)45:3<207::AID-ASI10>3.0.CO;2-1