# GESTIONE CONOSCENZA DEL SOFTWARE

**Definizione di dato:**
Qualcosa che puo' essere rappresentato dal nostro dominio.
Sono alla base del processo di ragionamento.
L'informazione è il risultato del processamento.
Facendo una certa analisi su dei dati in output otteniamo informazione.
Se non interviene conoscenza né un processo vero e proprio il dato rimane dato.
L'informazione può essere vista perciò come una collezione di dati da cui una persona può ricavare certe conclusioni.
L'informazione la trasmettiamo facilmente...la conoscenza viene acquisita tramite il ragionamento.
Quindi vediamo l'informazione come: L'atto di organizzare e manipolare un insieme di dati in modo tale da variare la conoscenza della persona che la riceve.
Gerarchia DIKW(data, information, knowledge, wisdom): modello generico per rappresentare le relazioni funzionali tra dati, informazione, conoscenza e saggezza (wisdomà sistemi esperti)

**DATA MINING:**
Scopre nuova conoscenza attraverso l'analisi dei dati, in un database strutturato.
.La Statistica - il campo della matematica applicata connesso con l'analisi dei dati - può essere definita altrimenti come "estrazione di informazione utile da insiemi di dati".
Il concetto di **data mining** è analogo. L'unica differenza è che questa recente disciplina ha a che fare con cospicui insiemi di dati.
In sostanza il data mining è l'"analisi matematica eseguita su databasedi grandi dimensioni". Il termine *data mining* è diventato popolare nei tardi anni '90 come versione abbreviata della definizione appena esposta.
Oggi il **data mining** (letteralmente: *estrazione di dati*) ha una duplice valenza:

- Estrazione, con tecniche analitiche all'avanguardia, di informazione implicita, nascosta, da dati già strutturati, per renderla disponibile e direttamente utilizzabile;
- Esplorazione ed analisi, eseguita in modo automatico o semiautomatico, su grandi quantità di dati allo scopo di scoprire*pattern* (schemi) significativi.

In entrambi i casi i concetti di informazione e di significato sono legati strettamente al dominio applicativo in cui si esegue data mining, in altre parole un dato può essere interessante o trascurabile a seconda del tipo di applicazione in cui si vuole operare.
Questo tipo di attività è cruciale in molti ambiti della ricerca scientifica, ma anche in altri settori (per esempio in quello delle ricerche di mercato). Nel mondo professionale è utilizzata per risolvere problematiche diverse tra loro, che vanno dalla gestione delle relazioni con i clienti (CRM), all'individuazione di comportamenti fraudolenti per finire all'ottimizzazione di siti web.

## Esempi

Che cosa **non è** estrazione di dati?

- Cercare un numero di telefono nell'elenco;
- Fare una ricerca in Internet su "vacanze alle Maldive".

Che cosa **è** estrazione di dati?

- Scoprire che alcuni cognomi (Benetton, Troncon, Cavasin) sono molto comuni in specifiche aree dell'Italia;
- Fare una ricerca nel web su una parola chiave e classificare i documenti trovati secondo un criterio semantico (p. es. "corriere": nome di giornale, professione, ecc.)

**TEST MINING = DATA MINNG**(applicata al testo di dati)+linguaggio base.
Scopre della nuova conoscenza attraverso l'analisi del testo, in un database non strutturato. È una forma particolare di data mining dove i dati consistono in testi in lingua naturale, in altre parole, documenti "destrutturati". Il text mining unisce la tecnologia della lingua con gli algoritmi del data mining.
L'obiettivo è sempre lo stesso: l'estrazione di informazione implicita contenuta in un insieme di documenti.
Negli ultimi anni ha avuto un notevole sviluppo, a causa dei progressi delle tecniche di elaborazione del linguaggio naturale (NLP in inglese), della disponibilità di applicazioni complesse attraverso gli Application service provider (ASP) e dell'interesse verso le tecniche automatiche di gestione della lingua mostrato sia dagli accademici, sia dai produttori di software, sia dai gestori dei motori di ricerca.

Processo text mining:

1. TESTO
2. PROCESSO DI TESTO -analisi sintattica e semantica del testo
3. GENERAZIONE FUNZIONALITA' – insieme di parole
4. SELEZIONE FUNZIONI- semplice conteggio, statistiche
5. TEXT/DATA MINING -classificazione, apprendimento supervisionato
6. CLUSTERING -apprendimento non supervisionato
7. ANALISI DEI RISULTATI

---

Una Similarity measure è una funzione che calcola il grado di somiglianza tra due vettori. Utilizzando una misura di similarità tra la query e ogni documento:

- E 'possibile classificare i documenti recuperati in ordine di pertinenza presunta.
- E 'possibile applicare una certa soglia in modo che la dimensione del set recuperata può essere controllato.

L'ontologia è una rappresentazione formale, condivisa ed esplicita di una concettualizzazione di un dominio di interesse. Una ontologia è composta da una tupla(C,R,F,I,A) + impegno ontologico

- Concetti
- Relazioni e Funzioni si applicano entrambe ai concetti del mondo reale e rappresentano le relazioni tra le classi del dominio.
- Istanze sono i singoli oggetti contenuti in una classe-
- Assiomi modellano in maniera esplicita espressioni sempre vere
- Impegno ontologico accordo sul significato del vocabolario usato per condividere conoscenza.

**Tipologie di sistemi software**:
Information System: Sistemi software impiegati nei processi di memorizzazione e recupero di dati. Suddividiamo l'information system in alcuni sottoprocessi:
• ricerca
• recupero (quando ho una base di informazione statica e devo recuperare dati da questa base che abbiamo già costruito)
• filtraggio (per flussi di informazione)
I task non vengono quasi mai rappresentati, siamo consapevoli noi del nostro task, ma il motore di ricerca non ne sa nulla.
Knowledge Management System: Si focalizzano sul Know how (come operare) relativo a

processi più o meno articolati o attività ripetitive; esperienza acquisita mediante attività passate. Utilizzati per sostenere processi e task knowledge-intensive. Oltre a rappresentare informazione esplicita (documenti), i KMS costruiscono una base di conoscenza a partire dal comportamento degli individui, monitoraggio di colloqui chat, mail… I KMS monitorano canali di comunicazione e fanno inferenza per interpretare situazioni, attività, comportamenti e soluzioni.

**Obiettivi:**

• Catturare, creare e condividere i best practices (tecniche, metodi e processi ottimali per una certa attività/scopo)
• Essere di supporto a sistemi di Experience Management e Decision Support
• Creare tassonomie che organizzano e mettono in relazioni elementi del mondo
• Individuare skills e esperti di settore
• Creare communities o knowledge networks dove poter condividere conoscenza interazioni tra utenti

Otteniamo: performance, competitività, innovazione, condivisione della conoscenza.
Information Overload (sovraccarico di informazioni)
Fenomeno che si manifesta quando l'utente interagisce con l'informazione.
Il problema è nel modo di gestire le informazioni, nessuno va contro l'eccesso, il problema è allocare efficacemente l'attenzione e gestire l'informazione stessa.
Un eccesso di informazione crea una deficienza nell'attenzione. Per questo occorre allocare efficacemente l'attenzione rispetto alle sorgenti disponibili.
E' lo stato in cui un utente (o anche un sistema) non è in grado di processare o filtrare la mole di dati in input. Avviene tipicamente un fallimento nelle normali funzionalità.

7 Comportamenti che l'utente intraprende per affrontare l'Information Overload:
Disfunzionali:
• Omission: evitare di processare alcuni input.
• Error: processare alcuni input in modo non corretto
• Escaping: evitare di processare gli input ed interrompere le attività correnti
Potenzialmente disfunzionali:
• Queueing: ritardare il processamento di alcuni input
• Approximation: ridurre gli standard di precisione nel processare/categorizzare gli input.

**Information Seeking**

Attività che l'individuo intraprende per cercare informazioni.
Il processo di Information Seeking ha luogo quando si riconosce un gap nella conoscenza corrente che puo' motivare una persona ad acquisirne di nuova.
Esiste tipicamente un obiettivo da soddisfare:
• Apprendere conoscenza
• Eseguire un certo taskPiu' in generale e' una attivita' motivata di acquisizione di conoscenza da sorgenti informative selezionate opportunamente dall'individuo.

Information seeking parte dall'utente...stiamo in uno stato un po' più avanzato rispetto allo stato primordiale.
Quello che ci interessa di più è l'information seeking e il bisogno informativo.

**Modello di information Seeking:**

Il modello di Ellis si focalizza invece sulle relazioni tra stage (o features) che l'individuo intraprende durante la ricerca sebbene gli stessi autori notino come tali relazioni dipendano fortemente dal contesto specifico.

Il modello include 8 stages non necessariamente lineari temporalmente:
**1) Starting**: consiste nelle attività' preliminari effettuate prima di avviare una ricerca di informazioni, ad esempio identificare il mezzo iniziale utile per la ricerca. Attività preliminari, selezionare una sorgente piuttosto che un'altra, per caratterizzare l'utente che tipo di sorgente viene utilizzata, se già conosciamo diverse sorgenti possiamo già utilizzare una selezione, considerazioni per selezionare una sorgente rispetto ad un altra...sarebbe opportuno modellarlo.
Entrano in gioco cose difficili da valutare:
La familiarità;
Autorevolezza
Qualità (interfaccia)
Accessibilità (se devo pagare per l'info...o se ci metto un'ora per ottenerla)
Quattro misure molto difficili da definirle chiaramente e da misurare.
Sono qualità individuali. Sono misure che dovrebbero essere molto contestuali riferite all'utente
(questi tre non hanno un ordine temporale: Chaining, Browsing, Monitoring)

**-2) Chaining: Chaining** attività molto importante, una delle poche che l'utente puo' fare all'interno del web.
due tipi di chaining:
(forward chaining)avanti:andare a vedere chi cita l'informazione/pagina/sito corrente, conosciamo questa informazione solo attraverso servizi esterni (base di dati in biblioteca.

(backward chaining)indietro:elemento della bibliografia, elementi informativi pubblicati prima
Informatico: avanti(forward chaining) perchè è interessato a cose più recenti, più affidabili rispetto al passato
Umanista: indietro, gli serve lo storico, le fonti, su cosa si basa...7

**3) Browsing**: seconda attività in parallelo:
Browsing: sembra l'attività vera e propria di seeking due tipologie:
across-document browsing: altri documenti
Within-document browsing: stesso documento lo stiamo facendo sullo stesso documento o su altri.
Stiamo ancora cercando informazione non stiamo apprendendo (quindi è browsing)-Monitoring: è un po' più diretto. Continuiamo a dare un'occhiata alle info per selezionare le sorgenti più importanti...è un'attività che si ripete, ad es siti di congressi o riviste online che fanno uscire continuamente nuove info e noi monitoriamo periodicamente le nuove info(andiamo a vedere periodicamente se sono uscite nuove info).

**4) Differentiating**: Differenziazione: a questo punto abbiamo collezionato l'informazione, ne abbiamo più di una(pagina web documento) dobbiamo fare quindi il filtraggio, quali sono le tipologie di interazione con l'informazione (le 7 fasi che abbiamo accennato)

**5) Extracting:** identifcare selettivamente materiale d'interesse da una sorgente informativa precedentemente individuata e approfondirlo.

**6) erifying-Ending**: controllare l'accuratezza dell'informazione e se sono presenti errori.
I processi di information seeking nel campo della computer science possono essere organizzati in base a due caratteristiche:

**7) Stabilita**’ temporale del bisogno informativo
• Stabile o dinamico

**8) Specifcita’** del bisogno informativo
• Specifco
• E.g., Qual e’ il candidato favorito nelle prossime elezioni?
• Ampio respiro
• E.g., Conoscere la vita di JFK
-Tipologia di dati prodotti dalla information source testuale
• Strutturata dove i dati sono conformi a un certo schema con una semantica chiara associata
ad ogni campo, e.g., record di una base di dati.
• Non strutturata dove l’informazione è tipicamente espressa in linguaggio naturale
Nel nostro contesto, per processo si intende una attivita’ condotta da umani, eventualmente con l’assistenza di un sistema informatico.
Un sistema invece indica un sistema software automatizzato e relativa unita’ elaborativa sviluppata a supporto degli utenti durante un processo.
Percio’ un sistema di Information Filtering e’ sviluppato per supportare gli utenti durante processi di Information Filtering.
Processo di Information Seeking suddiviso in due tipi: bisogno informativo e sorgente informativa.
Il bisogno informativo se è stabile ricadiamo in un processo di information filtering, se è dinamico facciamo difficoltà a metterlo in relazione l'uno con l'altro, (information Retrieval)
I bisogni informativi dell'utente non sempre sono dinamici, non sempre abbiamo bisogno di soddisfare un nuovo bisogno informativo possiamo aver bisogno di colmare la stessa lacuna.

## INFORMATION RETRIEVAL

Il processo di IR è fondamentale perché’ è alla base di molti sistemi di gestione dell’informazione come ad esempio i motori di ricerca, sebbene tale processo risulta incompatibile con il dominio. L'IR è un campo interdisciplinare che nasce dall'incrocio di discipline diverse. L'IR coinvolge la psicologia cognitiva, l'architettura informativa, la filosofia (vedi la voce ontologia), il design, il comportamento umano sull'informazione, la linguistica, la semiotica, la scienza dell'informazione e l'informatica. Molte università e biblioteche pubblicheutilizzano sistemi di IR per fornire accesso a pubblicazioni, libri ed altri documenti.
Motivi: I motori di ricerca per il Web devono trattare una sorgente informativa per niente statica; La sequenza di query formulate da un utente non sempre sono scorrelate tra loro e
variabili.
(architettura motore di ricerca)

Il motore di ricerca e’ composto principalmente da quattro moduli:

1. Un crawler segue i link presenti nelle pagine Web ed effettua il download
2. La copia viene immagazzinata in un repository locale
3. Un indexer analizza il testo delle pagine e lo memorizza con una opportuna rappresentazione in una base di conoscenza; questo permette di massimizzare la velocita’ nel retrieval
4. A partire da una query, un query engine costruisce una lista ordinata di risultati.

**Retrieval al contesto del Web.**

Due soluzioni vengono normalmente impiegate per adattare il processo di Information

**1.** Il motore di ricerca possiede una base di conoscenza dove memorizza con una opportuna rappresentazione copie delle pagine recuperate dal Web, in questo modo e' possibile velocizzare il processo di retrieval a partire da una query. Periodicamente si aggiorna questa base di conoscenza con le modiche apportate sulle pagine Web.
◦ Attenzione: la sorgente non è più considerata statica.
◦ Svantaggi: Tempo e risorse computazioni (network e storage) per mantenere aggiornata la base di conoscenza del motore di ricerca; Inoltre, a causa della rapidità' negli aggiornamenti del Web e dei limiti nella banda passante degli attuali network, la base di conoscenza sara' sempre non-sicronizzata rispetto all'informazione corrente sul Web.

**2.** L'insieme delle query formulate da un utente vengono considerate scorrelate l'una dall'altra
◦ Svantaggi: Ignorare eventuali correlazioni tra le query utente non permette di analizzare eventuali evoluzione o alterazioni dei bisogni informativi, perciò' si ignorano aspetti importanti delle esigenze dell'utente.
Sistemi di Information Retrieval basati sul modello a spazio vettoriale sono vantaggiosi perché' la loro computazione può' essere eseguita velocemente e produrre risultati in frazioni di secondo, ma mostrano alcuni grossi svantaggi a livello semantico:
L'**information retrieval (IR)** (lett: *recupero d'informazioni*) è l'insieme delle tecniche utilizzate per il recupero mirato dell'informazione in formato elettronico. Per "informazione" si intendono tutti i documenti, i metadati, i file presenti all'interno di banche dati o nel world wide web. Il termine è stato coniato da Calvin Mooers alla fine degli anni '40 del Novecento, ma oggi è usato quasi esclusivamente in ambito informatico.

Per recuperare l'informazione, i sistemi IR usano i linguaggi di interrogazione basati su comandi testuali. Due concetti sono di fondamentale importanza: query ed oggetto:

- Le **query** ("interrogazioni") sono stringhe di parole-chiavi rappresentanti l'informazione richiesta. Vengono digitate dall'utente in un sistema IR (per esempio, un motore di ricerca).
- Un **oggetto** è un'entità che mantiene o racchiude informazioni in una banca dati. Un documento di testo, per esempio, è un oggetto di dati.

Una tipica ricerca di IR ha come input un comando dell'utente. Poi la sua query viene messa in relazione con gli oggetti presenti nella banca dati. In risposta, il sistema fornisce un insieme di record che soddisfano le condizioni richieste.
Spesso i documenti stessi non sono mantenuti o immagazzinati direttamente nel sistema IR, ma vengono rappresentati da loro surrogati. I motori di ricerca del Web come Google e Yahoo sono le applicazioni più note ed ovvie delle teorie di Information Retrieval.

**COMPONENTI DEL SISTEMA IR**

1. **Operazioni sul testo**, forma gli indici per le parole(rimuovere le stoppword,rimuovere le parole alla radice, rimuovere prefissi e suffissi. Queste operazioni possono essere finalizzate non solo alla compressione(eliminazione di articoli, congiounzioni..)ma anche alla generalizzazione(non rappresentare termini ambigui ma concetti). Quando la rappresentazione del documento comprende l'intero insieme delle parole che lo compongono, si parla di full text logical view.
2. **Rappresentazione e indicizzazione dei documenti:** Nei sistemi IR ogni documento viene rappresentato mediante un insieme di parole chiave o termini indice. Un termine indice è una parola ritenuta utile per rappresentare il contenuto del documento. Gli indici vengono

utilizzati per generare **strutture di puntamento** ai documenti della collezione, facilitandone il recupero a fronte di una query.

3. **Ricerca** : ricerca documenti che contengono un dato token di una query dall'ordine inverso.
4. **Ranking**: un ranking è un ordinamento che dovrebbe riflettere gli interessi dell''utente. E' basato su:

identificazione di gruppi di termini comuni

condivisione dei termini pesati

probabilità di rilevanza.

## INFORMATION FILTERING

In un processo di Information Filtering si assume che gli utenti mostrino bisogni informativi e interessi a lungo termine e le sorgenti contengano informazioni usufruibili direttamente dall'utente.Tipicamente i sistemi software basati sull'Information Filtering devono gestire un grosso volume di informazioni, generate dinamicamente fornendo all'utente quelle che piu' verosimilmente soddisfano i suoi bisogni informativi.
Obiettivo: migliorare le capacita' dell'utente durante la ricerca di informazioni. Abbiamo un grosso volume di info generate dinamicamente...non ci aspettiamo che si ripetano spesso nel tempo (vd. le news). Vogliamo filtrare tra questa grossa mole le informazioni di interesse per l'utente.
La distinzione principale del processo di Information Filtering con quello dell'Information Retrieval (alla base degli attuali motori di ricerca Web) e' la velocita' di aggiornamento dei bisogni informativi e la velocita' di aggiornamento (e produzione) della sorgente informativa.
Nel Filtering si suppone che il bisogno informativo sia più' stabile mentre la sorgente produca informazioni con un tasso piu' rapido.
Un processo ideale dovrebbe poter affrontare sorgenti che si aggiornano spesso e bisogni informativi che possono variare in ogni istante.

Nel processo di Information Filtering ci sono tre fasi:

**1. Collection**: selezione a priori in cui non entra in gioco l'utente. Esistono due metodologie per collezionare le informazioni:
◦ Attiva: esiste un modulo software che si occupa di ricercare l'informazione da analizzare
◦ Passiva: l'informazione viene fornite per mezzo di uno stream, e.g., news feed.

**2. Detection**: I software basati su IF tipicamente includono una componente di modellazione degli interessi dell'utente (user model o profle) il cui obiettivo e' individuare e rappresentare gli interessi a lungo termine dell'utente. E' una fase in cui entra in gioco l'utente e il bisogno informativo. (matching) argomento: bisogno informativo e documento ed una misura di similarità (1:max attinenza; 0:minima attinenza)non solo i bisogni informativi ma anche altre informazioni, prevedere i bisogni informativi, su un ambito di ricerca che non si conosce...predire. Anche preferenze, doc. lungo o corto, la lingua.

**3. Display**: come visualizzare i documenti rappr. grafica o testuale, grafico a torta...nulla vieta all'inf system di analizzare l'info e proporre un suo elaborato, ci sono già dei servizi on-line che costruiscono pagine dinamiche con l'informazione raggruppata. L'informazione data non è quella originale, ma un suo processamento.
Costruzione di un documento ex-novo a partire dalla query.

**4. Considerzione**: quando interrompere l'utente?Quando suggerire l'informazione, momenti in cui si è interrompibili, altri no, dipende anche dal tipo di informazione (se di particolare interesse o no).

Un **sistema di filtraggio dell'informazione** è un sistema che rimuove ridondanti o indesiderati informazioni da un flusso di informazioni utilizzando (semi) metodi automatici o informatizzati, prima della presentazione di un utente umano. Il suo obiettivo principale è la gestione del sovraccarico di informazioni e l'incremento della semantico-rumore rapporto segnale. Per fare questo il profilo dell'utente viene confrontato con alcune caratteristiche di riferimento.
.I sistemi IF eseguono il filtraggio  sulla base dei profili utente, quindi è necessario un modello strutturato degli interessi degli utenti. Il maggio problema dei sistemi IF, è che le Keywords, da sole, non sono del tutto appropriate per la rappresentazione dei contenuti per via della polisemia, sinonimia e concetti multi-parola.
**POLISEMIA**:in semantica indica la proprietà che una parola ha di esprimere più significati.
**SINONIMIA**:in semantica indica la relazione che c'è tra due lessemi che hanno lo stesso significato.
**MULTI.PAROLA**:parole che insieme identificano un concetto diverso da quelli che la formano, esempio :"intelligenza artificiale"

## INFORMATION FILTERING VS INFORMATION RETRIEVAL

IF e IR sono concettualmente dissimili:
- IF rimuove dati da uno stream dinamico di informazioni
    - Ad esempio: junk emails, news feeds fltering
- IR accede e recupera informazioni da collezioni statiche
    - Ad esempio: librerie digitali
    - Ma nella pratica spesso vengono riadattati per vari scopi,

La presenza di user profles nel processo di IF implica rappresentazioni più complesse dei bisogni informativi rispetto al processo di IR.
Nel IR si studiano tecniche per ridurre il tempo di recupero mentre nel'IF si preferisce aumentare la precisione nei risultati utilizzando tecniche di matching tra bisogno informativo e informazione più' sofisticate, che possono risolvere almeno in parte il problema dei false match. Inizialmente i ricercatori hanno proposto molti approcci basati sul processo di IF, attualmente pero' si tende ad adattare sistemi di IR tradizionali incorporando modelli utente tipici del IF. I
vantaggi di tale soluzione sono:
- Mantenere le attuali infrastrutture dei motori di ricerca ampiamente collaudate aumentando la precisione nei risultati includendo gli interesse dell'utente
- Si mantengono le stesse interfacce utente dei motori di ricerca, facili da comprendere per un utente
- Si sfruttano le funzionalità' offerte dai moduli del motore di ricerca all'interno del processo di filtraggio, e.g.,:
• Accedere alla storia delle query e dei risultati visionati dall'utente
• Accedere alla collezione di pagine memorizzate all'interno della base di conoscenza
• L'unico vincolo e' riuscire a mantenere i tempi di risposta nella produzione dei risultati al di sotto di un limite prefssato (e.g., 0,5 secondi)

Esempio
Google Personalized tiene traccia di un profilo utente allo scopo di posizionare le pagine di maggiore interesse in cima.
Ad esempio, per un utente che ha visitato spesso articoli scientifici riguardo tecniche di IR e IF, si ottengono due liste di risultati distinti a partire dalla query "information retrieval".

1. Quando IR si occupa della raccolta e organizzazione dei testi, IF si occupa della distribuzione dei testi a gruppi o individui
2. Quando IR tipicamente si occupa della selezione dei testi da una banca dati relativamente statici, IF si occupa invece della selezione o eliminazione dei testi da un flusso di dati dinamici.
3. Quando IR si occupa di rispondere alle interazione dell'utente con testi all'interno di una singola ricerca, IF si occupa a lungo temine della ricerca su una serie di informazioni-episodio.

## WORD SENSE DISAMBIGUATION

Processo di decisione del senso di una parola usata in uno specifico contesto. I differenti significati delle parole polisemiche sono chiamati sensi. Il contesto determina il corretto senso.. Il processo di disambiguazione presenta due approcci: il primo basato sulla conoscenza basata su dizionari, il secondo basato sul corpus e usa una collezione di materiale scritto.

## COLLABORATIVE/SOCIAL FILTERING

Fa uso di un database di preferenze utente allo scopo di:
a) trovare utenti con interessi e comportamenti simili
b)predire se le informazioni su un elemento osservato siano di interesse dell'utente sulla base della valutazione di altri utente sullo stesso elemento. Il processo di raccomandazione, prendendo in input una grande gamma di articoli e la descrizione del bisogno informativo degli utenti, ha l'obiettivo di presentare una piccola gamma di elementi adatti alla richiesta dell'utente. I sistemi che utilizzano questo tipo di filtraggio sono:

- USER-TO-USER :ogni utente è rappresentato da un vettore n dimensionale di elementi e la raccomandazione è basata su pochi utenti,ma simile all'utente attivo. Questi utenti vengono definiti neighbors(vicini di casa).
- ITEM-TO-ITEM: utilizza una tabella di elementi e simili che più utenti tendono ad acquistare. Questo tipo di approccio è utilizzato da amazon.com. L'algoritmo trova oggetti simili a ciascuno degli acquisti e dei voti dell'utente attivo e aggrega a questi elementi, raccomandando il più popolare o gli elementi correlati.

## SVANTAGGI

**1. Gray sheep**: L'utente, dopo alcuni voti (Stellette di amazon ad esempio) si annoia a giudicare tutti gli item che vede e inizia a dare giudizi intermedi (Esempio su 5 stelle, l'utente sceglie sempre 3). Facendo così il sistema di IF lavora male perche non riesce più a disambiguare gli item che sono piaciuti da quelli che non sono piaciuti all'utente.

**2. Cold Start**: Avvio freddo, si verifica quando c'è un nuovo utente o un nuovo item. Il sistema di raccomandazione non riesce a raccomandare degli item nuovi perchè nessuno li ha ancora valutati o acquistati, nel caso del nuovo item. Nel caso del nuovo utente, il sistema non riesce a fornire delle raccomandazioni ad utenti nuovi perchè non può calcolare la similarità tra utenti nuovi e utenti esistenti, visto che essendo nuovi non hanno ancora valutato/acquistato un item.

**3. Matrice** Sparsa: Le due tecniche di raccomandazione User-to-user e Item-to-item sono realizzate con delle matrici. Siccome è improbabile che tutti gli utenti valutino/acquistino tutti gli item, la matrice avrà molte celle vuote (matrice sparsa).

**4. Cross Recommendation**: I sistemi di IF devono evitare di consigliare agli utenti degli item che sono differenti dagli item dalla categoria di item che solitamente acquista. Vedi l'esempio sulle slide del libro del Pastore che viene raccomandato ad un utente che compra libri di informatica..

**5. Canto delle sirene**: Utenti falsi che si registrano al sistema e giudicano ottimi degli Item di cui sono autori o che vogliono che siano frequentemente raccomandati. In questo modo c'è un risultato falsato delle raccomandazioni.

## CONTENT-BASED FILTERING

ogni utente opera in maniera indipendente e gli oggetti sono rappresentati mediante alcune caratteristiche(es film—attori, regia, staff, tecnici..). Il filtraggio è basato sul confronto tra il contenuto delle voci dell'articolo e le preferenze utente definite nel profilo dello stesso.

INDICAZIONI DI RICERCA:
Intelligent Information Access =

1. Accesso personalizzato da profili utente +
2. Accesso Semantico da identificazione di concetto nel documento.

| Parameters | Information Retrieval | Information Filtering |
|---|---|---|
| Representation of Information Needs | queries | profiles |
| Goal | selection of relevant items for query | filtering out irrelevant items or collecting items |
| Frequency of use | ad hoc use<br>one time user | repetitive use<br>long term users |
| Type of Users | Not known to the system | "Profiled" |
| Database | (relatively) static | very large<br>dynamic |

### Misure di prestazione

Ci sono molti modi per misurare quanto bene l'informazione intesa si associa all'informazione recuperata.

**Precisione**

La precisione (in inglese *precision*) è la proporzione di documenti pertinenti fra quelli recuperati:

$P$ = (numero di documenti pertinenti recuperati) / (numero di documenti recuperati)

Nella classificazione binaria la precisione è analoga al valore positivo di previsione. La precisione può anche essere valutata a rispetto a un certo valore soglia, indicato con $P@n$, piuttosto che relativamente a tutti i documenti recuperati: in questo modo, si può valutare quanti fra i primi*n* documenti recuperati sono rilevanti per la query.

Si noti che il significato e l'uso del termine "precisione" nel campo dell'IR differiscono dalla definizione di accuratezza e precisione tipiche di altre discipline scientifiche e tecnologiche.

## Recupero

Il recupero, o richiamo, (in inglese *recall*) è la proporzione fra il numero di documenti rilevanti recuperati e il numero di tutti i documenti rilevanti disponibili nella collezione considerata:

$R$ = (numero di documenti rilevanti recuperati) / (numero di documenti rilevanti)

$$recall = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ relevant\ documents}$$

$$precision = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved}$$

Nella classificazione binaria, questo valore è chiamato sensitività.

## Misura F

La misura F (in inglese *F-measure*) è la media armonica pesata fra precisione e recupero. La versione tradizionale, detta anche *bilanciata*, è data da:In generale, la formula è:

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}}$$

Altre due formule comuni sono $F0.5$, che assegna alla precisione un peso doppio rispetto al recupero, e la $F2$, che al contrario pesa il recupero al doppio della precisione.

| n | doc # | relevant |
|---|---|---|
| 1 | 588 | x |
| 2 | 589 | x |
| 3 | 576 | |
| 4 | 590 | x |
| 5 | 986 | |
| 6 | 592 | x |
| 7 | 984 | |
| 8 | 988 | |
| 9 | 578 | |
| 10 | 985 | |
| 11 | 103 | |
| 12 | 591 | |
| 13 | 772 | x |
| 14 | 990 | |

Let total # of relevant docs = 6
Check each new recall point:

R=1/6=0.167; P=1/1=1

R=2/6=0.333; P=2/2=1

R=3/6=0.5; P=3/4=0.75

R=4/6=0.667; P=4/6=0.667

Missing one relevant document. Never reach 100% recall

R=5/6=0.833; P=5/13=0.38

Information Retrieval (IR) si occupa della rappresentazione, la conservazione, organizzazione e accesso alle informazioni voci "

- l'utente deve prima tradurre questo bisogno di informazioni in una query che possono essere trasformati con un motore di ricerca (IR o sistema) ".

- "Data la query dell'utente, l'obiettivo chiave di un sistema di IR è di recuperare informazioni che potrebbero essere utili e pertinenti per l'utente. Il l'accento è sul reperimento delle informazioni in contrapposizione alla il recupero dei dati.

**Classificazione matematica dei modelli**

- ***Modelli Set-theoretic*** rappresentano i documenti mediante insiemi. Le somiglianze derivano in genere da operazioni teoriche su questi insiemi. I modelli più comuni sono:
  - Modello Booleano Standard
  - Modello Booleano Esteso
  - Recupero fuzzy
- ***Modelli Algebrici*** rappresentano i documenti e le query con vettori, matrici o tuple, che, utilizzando un numero finito di operazioni algebriche, vengono trasformati in una misura numerica, la quale esprime il grado di somiglianza dei documenti con la query.
  - Modello a Spazio Vettoriale
  - Modello a Spazio Vettoriale Generalizzato
  - Topic-based vector space model (literature: [1], [2])
  - Modello Booleano Esteso
  - Enhanced topic-based vector space model (literature: [3],[4])
  - Latent semantic indexing aka latent semantic analysis
- ***Modelli Probabilistici*** trattano il processo di recupero dei documenti come un esperimento aleatorio multi-livello. Le somiglianze sono quindi rappresentate come probabilità. I teoremi probabilistici come ilteorema di Bayes sono spesso usati in questi modelli.
  - Binary independence retrieval
  - Uncertain inference
  - Language models
  - Divergence from randomness models

**Classificazione in base alle proprietà dei modelli**

- *Modelli senza interdipendenza dei termini* trattano diversi termini/parole come non interdipendenti. Ciò viene rappresentato spesso nei modelli a spazi vettoriali affermando che i vettori dei termini siano ortogonali, o nei modelli probabilistici affermando che le variabili dei termini siano indipendenti.
- *Modelli con interdipendenza dei termini intrinseca* consentono una rappresentazione diretta delle interdipendenze tra termini. Comunque il grado di interdipendenza tra due termini è definito dal modello stesso. In genere, esso è direttamente o indirettamente derivato (vedi per es. dimensional reduction) dalla co-occorrenza di questi termini nell'intero insieme di documenti.
- *Modelli con interdipendenza dei termini trascendente* consentono una rappresentazione diretta delle interdipendenze tra termini, ma essi non riportano come l'interdipendenza tra due termini sia definita. Si riferiscono ad una fonte esterna per stabilire il grado di interdipendenza tra due termini (ad esempio un umano o degli algoritmi sofisticati).

**MODELLO DI RITROVAMENTO**: un modello di ritrovamento specifica dettagliatamente la rappresentazione dei documenti, la rappresentazione delle query

e le funzioni di ritrovamento. Determina anche una nozione di rilevanza che puo essere binaria o continua. Ci sono due classi di modelli di ritrovamento: modello booleano e vector space model (o modello probabilistico).

**MODELLO BOOLEANO**: un documento e rappresentato come un insieme di parole chiave. Le query sono espressioni booleane di parole chiave, connesse da AND, OR e NOT, incluso l'uso delle parentesi. L'output del modello e l'affermazione se un documento e rilevante o no, non ci sono ranking o risultati intermedi. Il modello booleano e popolare perche e facile da capire per query semplici ed e libero da formalismi. Il modello booleano puo essere esteso con l'inclusione del ranking. Pero e troppo rigido, infatti AND significa "tutti", OR significa "nessuno". Difficile esprimere query complesse, difficile controllare in numero di documenti ritrovati, difficile ordinare l'output ed e difficile gestire un feedback di rilevanza.

**MODELLO PROBABILISTICO**: un documento e tipicamente rappresentato da una bag of words, e l'utente specifica un insieme di termini desiderati con un peso (opzionale). Il ritrovamento con un modello probabilistico e basato sulla similarita tra la query ed i documenti. L'output e presentato dai documenti ordinati tramite un rank in base alla similarita con la query. Questo modello supporta il feedback di rilevanza.

**VECTOR-SPACE MODEL**: Rappresenta i documenti come insieme di pesi di keyword. La pesatura e fatta con il tf-idf (formula), metodo di peso ottimo per quantita non enormi di documenti, e si calcola la similarita tra i documenti e la query, ordinando i risultati. Il vector space model utilizza un semplice approccio basato sulla matematica, considera sia le occorrenze dei termini locali ad un documento che le occorrenze dei termini globali. Fornisce corrispondenza parziale e classifica dei risultati. Lavora bene anche con grandi moli di documenti. Non prende in considerazione informazioni semantiche e sintattiche. Assume che ogni termine sia indipendente.

## ONTOLOGIA

• Un’ontologia è una *descrizione formale esplicita di un dominio* di interesse.
• **Descrizione**: una forma di rappresentazione della conoscenza
• **Formale**: simbolica e meccanizzabile
• **Esplicita**: elenchi estensionali di frammenti di conoscenza
• **Dominio**: ristretta ad un determinato sottoinsieme dello scibile, affrontato da un certo punto di vista.

Una ontologia descrive le parole comuni e i concetti (significati) usati per descrivere e rappresentare un’area di conoscenza (dominio).
Una ontologia può essere usata da persone, applicazioni, database etc. per condividere una conoscenza comune riguardo ad un certo dominio (educazione, medicina, riparazione di automobili etc.). L’ontologia include le definizioni dei concetti del dominio e delle loro relazioni in un modo usabile dal computer (ma anche comprensibile agli umani).

Più nel dettaglio si tratta di una teoria assiomatica del primo ordine esprimibile in una logica descrittiva. Il termine ontologia è entrato in uso nel campo dell'intelligenza artificiale e della rappresentazione della conoscenza, per descrivere il modo in cui i diversi schemi vengono combinati in una struttura dati contenente tutte le entità rilevanti e le loro relazioni in un dominio.

Le componenti si una ontologia sono:

- **Concetti** :set degli oggetti di cui vogliamo parlare, generalmente organizzati in classificazioni e vengono usati per descrivere concetti.
- **Relazioni e funzioni**:si applicano entrambe a concetti del mondo reale e rappresentano le relazioni esistenti tra le classi del dominio.
- **Istanze**:singoli oggetti contenuti in una classe.
- **Assiomi:**modellano in maniera esplicita espressioni sempre vere. Usati per a)definire il significato dei vari componenti dell'ontologia, b) definire le relazioni complesse, c)verificare la correttezza di un'informazione e/o dedurne una nuova.
- **Impegno ontologic**o:accordo sul significato del vocabolario usato per condividere conoscenza.

**ONTOLOGIA ={C,R,F,I,A} + impegno ontologico.**

Le ontologie si classificano in:

- top-level, concetti molto generali o comune senso di conoscenza. Sono indipendenti dal dominio.
- Domain ontologies: vocabolario relativo a un generico dominio. Esempio: fisica, medicina..
- Task ontologies:vocabolario relativo a una generica attività o compito; esempio diagnostica, vendite..
- Application ontologies:conoscenza proveniente da domain ontologies e task ontologies. È in genere la specializzazione di domain e task ontologies.

Una ontologia riflette la struttura del mondo reale, riguarda la struttura dei concetti e le rappresentazioni fisiche non sono un problema. Una struttura di classe OBJECT ORIENTED, riflette le strutture del codice e dei dati, di solito riguarda il comportamento dei metodi e descrive la rappresentazione fisica dei dati.

## WORD NET

E' una ontologia linguistica TOP Level che rappresenta in maniere esplicita e formale la conoscenza linguistica umana. L'idea nasce nel 1985 ed è cosi pensata:

- Obiettivo: ricerca concettuale nei dizionari
- Risultato : definizione di un database lessicale
- Linea di ricerca: memoria lessicale umana.

Gli elementi del lessico (sostantivi, verbi, aggettivi e avverbi) sono raggruppati in insiemi detti **Synset.** Ogni synset corrisponde a un significato; l'idea è che i termini che fanno parte di uno stesso synset siano sinonimi.

Oltre alle relazioni tassonomiche , cioè iperonimia e iponimia, in word net sono presenti altre relazioni semantiche. Queste variano in funzione del tipo di parola e includono:

per i sostantivi

- iperonimia, iponimia, coordinazione, olonimia, meronimia

per i verbi:

- iperonimia, troponimia, implicazione, coordinazione.

Gli aggettivi sono classificati come:

- nomi relativi
- simile a,
- partecipi dei verbi

gli avverbi seguono la classificazione dell'aggettivo da cui derivano.

**IPERONIMIA:** rapporto di superiorità di un vocabolo che, rispetto a un altro, abbia un significato più esteso e comprensivo. Il termine che, riferito a un altro, ha senso, ovvero una comprensione logica, incluso nel senso di questo. Ad esempio conica è iperonimo di ellisse, parabola e iperbole, oppure fiore è iperonimo di gisglio.

**IPONIMIA**: relazione tra due termini, il primo dei quali ha un senso, ovvero una comprensione logica, che include il senso del secondo. In altre parole il primo è iponimo del secondo. Ad esempio si ha iponimia tra numero razionale e numero reale.

**OLONIMIA**: L'olonimia è una relazione semantica, l'olonimia indica la relazione tra un temine che indica l'intero e un termine che ne rappresenta una parte o un membro. Ad esempio albero è una olonimia di corteccia  o tronco.

**MERONIMA**:la meronimia è una relazione semantica utilizzata in linguistica. Un meronimo indica un costituente o membro di qualcosa. Ad esempio dito e meronimo di mano in quanto dito parte della mano, similmente ruota e meronimo di automobile.
**TROPONIMIA**: un troponimo descrive un particola modo di fare qualcosa (riferito solo ai verbi). Esempio: camminare ha un significano più specifico di andare, recitare ha un significato più specifico di parlare...

Word net viene usato per  Sistemi Information Retrieval e Text Categorigation  per aggiungere semantica al processo di ritrovamento/categorizzazione.
La memoria lessicale umana si suddivide in 4 parti dedicate rispettivamente a **nomi, verbi aggettivi e avverbi**; gli ideatori di word net ispirandosi a tale teoria hanno suddiviso la conoscenza lessicale in maniera simile.

PAROLA :  un'associazione tra una word form e una word meaning, cioè tra una stringa, insieme di lettere che la costituiscono e  il concetto lessicale che tale stringa vuole esprimente, cioè il suo significato. Il mappaggio tra la stringa e il significato viene realizzato con  la matrice lessicale.
Lo scopo principale di word net è quello di riuscire a traferire a un computer tutta la conoscenza linguistica, cioè la word form, le word meaning e il mapping tra le due.

## XML, DTD, RDF

XML è un linguaggio di markup che definisce un meccanismo un meccanismo sintattico che consente di estendere e controllare il significato di altri linguaggi marcatori. XML è utilizzabile dalla definizione della struttura dei documenti alla allo scambio di informazioni tra sistemi eterogenei, dalla rappresentazioni di immagini alla definizione di formati dati.. un documento xml deve essere ben formato(avere un prologo, un elemento radice unico e tutti i tag devono essere bilanciati) e valido(deve essere ben formato ed essere conforme ai requisiti strutturali definiti nella dtd. In XML la logica dei dati è separata dalla logica di rappresentazione dello stesso documento. Inoltre XML fornisce un meccanismo generico per rappresentare i dati e per il trasporto degli stessi.

**DTD** è linguaggio schema, ossia un linguaggio formale per esprimere schemi. Uno schema non è altro che una definizione formale della sintassi di un linguaggio XML. DTD specifica le regole a cui devono attenersi i documenti XML. Fornisce anche meta-info sui contenuti del documento, nomi di elementi validi,nomi e valori di attributi validi, come gli elementi possono essere annidatin con gli altri. Tipicamente la DTD è localizzata in un documento separato dall'XML. Non fornisce nessuna informazione sulla semantica del documento. La DTD è opzionale.

XML SCHEMA:Un documento che descrive cosa un altro documento valido può contenere. Contiene le specifiche sintattiche di un documento XML che descrive i contenuti consentiti del doc XML. XML SCHEMA è l'unico linguaggio di descrizione del contenuto di un fule XML, ben formato e valido. Lo scopo è quello di delineare quali elementi sono permessi, quali tipo di dati sono ad essi associati e quale relazione gerarchica hanno fra loro gli elementi contenuti nel file

XML.

IL Linguaggio DTD a differenza di XMLS non permette:

- di definire alcuna restrizione sul contenuto del testo
- di avere un controllo del contenuto misto
- di avere alcun controllo circa l'ordinamento degli elementi.
- Di creare tipi di dato definiti dall'utente.
- Di ereditare definizioni di elementi , attributi e tipi di dato
- supportare l'evoluzione degli schemi
- rendere possibile l'integrazione della documentazione negli stessi schemi.

## Il semantic web:alcune tecnologie di base

### I metadati e RDF

Il semantic web si basa sull'ipotesi che le macchine possano accedere a un insieme strutturato di informazioni e a un insieme di regole di inferenza da utilizzare per il ragionamento automatico. Il temine web semantico è stato associato all'idea di un web nel quale agiscano agenti intelligenti: applicazioni in grado di comprendere il significato dei testi presenti sulla rete e perciò in grado e perciò guidare l'utente verso l'informazione ricercata.. La sfida del semantic web è fornire un linguaggio per esprimere dati e regole per ragionare sui dati, che consenta l'esportazione sul web delle regole da qualunque sistema di rappresentazione della conoscenza.. Uno standard legato a questi temi è RDF, un linguaggio in sintassi XML per definire e esprimere ontologie, Le informazioni sulla risorsa,la quale è identificata univocamente da URI, vengono dette anche metadati. L'uso efficace di metadati richiede che vengano stabilite delle convenzioni per la semantica, la sintassi e la struttura.

Resource Description Framework è lo strumento base proposto dalla W3C per la codifica, lo scambio e il riutilizzo di metadati strutturati e consente l'interoperabilità tra applicazioni che si scambiano informazioni sul web. E' costituito da due componenti:

- **RDF model and Syntax**: espone la struttura del modello RDF e descrive una possibile sintassi.
- **RDF Schema**: espone la sintassi per definire schemi e vocabolari per i metadati. Ogni predicato è in relazione con altri predicati e permette di dichiarare l'esistenza di proprietà di un concetto. RDF schema permette, inoltre, la definizione di nuovi tipi di classi e le gerarchie ad esse associate. In RDF si possono rappresentare le risorse come istanze di classi e definire sottoclassi e tipi.
- **L'RDF Data model** si basa su tre principi chiave:

1. qualunque cosa può essere identificata da un Universal Resource Identifer(URI)
2. The least Power: utilizzare il linguaggio meno espressivo per definire qualunque cosa.
3. Qualunque cosa può dire qualunque cosa su qualunque cosa.

Il data RDF model consente di rappresentare statement RDF in modo sintatticamente neutro ed è basato su tre tipi di oggetti :

**Resources**: qualunque cosa identificata da URI è detta Risorsa. Le risorse sono sempre individuate da URI che sono identificatori univoci di risorse che possono essere URL(Universal Resource Locator) o URN (Universal Resource Name).

**Properties**:una property è un aspetto specifico, un attributo, una caratteristica o una relazione per descrivere una risorsa. Le proprietà associate alle risorse sono identificate da un *nome* e assumono dei *valori*

**Statements**: una risorsa con una proprietà distinta da un nome e un valore della proprietà per la

specifica risorsa, costituisce un RDF statement. Uno statement è una tupla composta da un oggetto(risorsa), predicato(proprietà) e da un oggetto(valore). L'oggetto di uno statement, cioè il property value, può essere una espressione oppure un altra risorsa.

**LIMITI RDF:** RDF ammette solo relazioni binarie, poche caratteristiche di proprietà, restrizione locale dei range, complicata descrizione dei concetti, restrizione di cardinalità e assiomi disgiunto. Una possibile soluzione è l'utilizzo di OWL che prevede classi , proprietà, individui e valori che sono immagazzinati come documenti web semantici. Ci sono classi, gerarchie di classi,più nodi per specificare l'organizzazione delle classi e classi predefinite.
L'approccio tradizionale al reperimento di informazioni è essenzialmente sintattico, e non è in grado di soddisfare adeguatamente le esigenze degli utenti che sono interessati all'aspetto semantico dell'informazione. Le ontologie permettono di rappresentare la conoscenza, e di renderla disponibile in modo interoperabile grazie alle tecnologie del Semantic Web viste.
Se l'XML si rivolge alla descrizione dei documenti a livello sintattico, RDF è particolarmente indicato per rappresentare dati, fornendo un metodo potenzialmente capace di gestire contenuti, ad esempio il reperimento delle informazioni in funzione delle specifiche esigenze dell'utente.

**RDFS VS OO**:

L'RDFS è centrato sulle proprietà mentre l'OO è centrato sulle classi.
RDFS definisce una proprietà in termini di quale classe può avere quella proprietà e quali classi possono essere oggetto di quella proprietà. L'OO definisce una classe in termini di quale proprietà possiede e di che tipo sono i valori della stessa. Tutti possono riusare e referenziarsi a una classe o proprietà definita da altri usando gli URI.

NAMESPACE. :Un namespace XML è una collezione di nomi che possono essere usati come elementi o nomi di attributo di un documento XML. Il namespace qualifica i nomi degli elementi univocamente sul web per evitare conflitti tra elementi con lo stesso nome. Il namespace è identificato da un URI, ed è usato per definire un dominio.

## ORGANIZZAZIONE E INFORMATICA

l'informazione è un insieme di più dati, messi in relazione e interpretati, in modo da avere significato e rappresentare un evento, I dati possono essere archiviati mentre le informazioni vengono prodotte per essere utilizzate. **La conoscenza** è il modo in cui utilizziamo e mettiamo in relazione le informazioni. Il valore dell'informazione dipende da tra fattori:tempo, luogo e struttura. La conoscenza è indipendente dai canoni oggettivi perché siamo noi a definire la metodologia con cui deve essere acquisita.

L'azienda è un insieme di:

- Risorse : che il sistema azienda scambia con l'ambiente
- Processi: attraverso i quali le risorse vengono scambiate e gestite
- Unità organizzative: struttura azienda per svolgere i processi in modo efficiente e efficace.

Le linee guida tracciano a loro volta le idee quotidiane che rappresentano il modo di vivere e si stare in azienda (lavorare in qualità, competitività, essere innovativi, eleganza, rispetto per il cliente...).
Un modo per organizzare in maniera sistematica e logica dati ed informazioni,punti di partenza ed arrivo, sorgenti e destinatari, flussi di trasferimento, mezzi, processi e risorse impegnate è quello che possiamo definire :SISTEMA INFORMATIVO.

**Il SISTEMA INFORMATIVO** consente di :

1. Acquisire
2. archiviare----------> INFORMAZIONI
3. elaborare
4. comunicare

Una rappresentazione fisica del Sistema informatico e che è inclusa nel sistema Informativo, viene definito SISTEMA INFORMATICO.

**L'AZIENDA ha cinque fasi di sviluppo....**

1. creatività – capacita innovativa, trovare nuovi prodotti, nuovi mercati. La situazione genera una crisi di comando dovuta alla crescente difficoltà di orientare le azioni e le decisioni dei singoli verso obiettivi comuni.
2. Autorità - uomo forte alla guida dell'azienda, ne può scaturire crisi di autonomia.
3. delega - decentramento decisionale ma questo se nasce come reazione a un accentramento può provocare eccessi opposti.
4. coordinamento- distribuzione obiettivi e responsabilità, crescita ruoli dello staff. Fase che può portare a una crisi di burocrazia.
5. Partecipazione- intensa e paritetica collaborazione interpersonale per degenerazione burocratica,

…...**e quattro gruppi direzionali aziendali:**

1. stile di direzione
2. organizzazione
3. sistema si pianificazione
4. sistema informativo

Una organizzazione può avere finalità di lucro o di utilità sociale. Inoltre un altro aspetto è la sua dimensione valutata secondo:

- parametri economici
- paramenti occupazionali (numero di dipendenti, collaboratori..)
- parametri diversi legati al settore di attività.

Il sistema informativo deve essere costruito tenendo in forte conto i processi aziendali e le caratteristiche organizzative.

Una struttura organizzativa comprende le unita organizzative, con l'assegnazione a ciascuna di esse di obbiettivi, poteri e compiti. Essa comprende funzionigrammi e organigrammi. I funzionigrammi sono schemi che descirvono i compiti senza l'indicazione delle persone che ricoprono i ruoli. L'organigramma è la rappresentazione fondamentale della struttura organizzativa; indica quali sono i rapporti formali di dipendenza tra le entità aziendali, Non rappresenta, invece,gli effettivi cmpiti degli enti.

I flussi informativi, ossia i percorsi di trasmissione e di elaborazione attraverso l'azienda possono essere:

- **dal basso verso l'alto**, percorsi che aggiornano i livelli alti dell''organizzazione sull'andamento delle operazioni.
- **dall'alto verso il bass**o, contengono ordini, direttive, disposizioni emanate dai vertici allo scopo di orientare l'azione dei livelli operativi
- **orizzontale,** porta individui appartenenti a enti aziendali diversi a condividere informazioni e a collaborare alla risoluzione dei problemi.

I BUSINESS PROCESS sono attività collegate tra loro nel tempo e nello spazio, svolta da uomini e e mezzi(risorse) di una azienda. Il BP può essere definito come una tupla concettuale BP=(A,I,O,C)

- A = **A**ttività (operazioni su oggetti fisici o informativi o su decisioni) svolte da vari attori
- I = **I**nput (materie prime e risorse aziendali come uomini e mezzi)

- O = **O**utput (oggetti fisici, beni immateriali o servizi)
- C = **C**lienti (destinatari dell'output di processo)

Prestazioni e configurazioni dei *business process* dipendono da una serie di elementi tra loro interdipendenti (*leve*[1]):
- il flusso di attività (che forma il ciclo completo del processo)
- la struttura organizzativa (che definisce la divisione delle responsabilità e delle attività)
- le competenze delle risorse umane che intervengono nel processo
- la tecnologia utilizzata per eseguire il processo (tipicamente TI[2])
- il sistema di misurazione e controllo delle prestazioni di costi, tempo, qualità e servizio.

È proprio attraverso queste *leve* che la metodologia di innovazione BPR agisce sull'organizzazione aziendale. (vedi appendice)

La conoscenza aziendale è data da :
- **conoscenza esplicita o tangibile**, formata da informazioni strutturate, tipo dati, brevetti, regole,procedure, cioè tutto ciò che è trasmissibile e conservabile.
- **Conoscenza tacita e intangibil**e rugiada le informazioni non esprimibili come le competenze, le esperienze individuali,la capacità, le conoscenze.

**KNNOWLWDGE MANAGEMENT è** un sistema di gestione della conoscenza, cioè una disciplina che studia le conoscenze aziendali attraverso: Metodi e strumenti basati su :
innovazione culturale
innovazione organizzativa
innovazione tecnologica
**Finalizzata a** Sviluppare capacità e competenze in grado di aumentare la Competitività dell'impresa.
Per facilitare l'apprendimento della conoscenza si identificano tre figure:
1. **Knowledge Worker** -lavoratore della conoscenza, facilita l'apprendimento e la condivisione della stessa
2. **Knowledge Manager**- responsabile del processo KM
3. Chief Knowledge Officer- responsabile processo di Knowòledge Officer.

Le fasi di un processo di KM sono:
- Acquisire o creare nuova conoscenza.
- Codificare per poterla riconoscere rispetto alle altre conoscenze dello stesso tipo
- Archiviare la conoscenza informatica in archivi informatici.
- Ordinare, Indicizzare al fine di rintracciare la conoscenza negli archivi. Essa deve essere suddivisa in categoria.
- Gestire la conoscenza
- Distribuire la conoscenza per renderla disponibile all'utenza.
- Condividere.
- Collaborare aiutando gli altri ad acquisire o produrre nuova conoscenza.4
- Usare la conoscenza, cioè raggiungere gli obiettivi e fare Business.
- Creare, produrre nuovi elementi di conoscenza.
- Alimentare la conoscenza, cioè inserire i nuovi elementi creati rendendoli disponibili.

**CLASSIFICATORE LINEARE**

In generale il classificatore Rocchio guarda la vicinanza del documento da classificare al centroide degli esempi positivi e la lontananza dal centroide degli esempi negativi.
SVANTAGGI: Il classificatore Rocchio come tutti i Classificatori Lineari ha lo svantaggio che divide lo spazio dei documenti linearmente. Questo comporta gravi perdite di efficacia: la media è solo parzialmente rappresentativa dell'intero insieme

**Algoritmo di Rocchio (apprendimento)**
Sia l'insieme delle categorie {c1, c2,…cn}
For i from 1 to n let pi = <0, 0,…,0>
(inizializzione)
For each esempio di training <x, c(x)> ∈ D
Let d = vettore TF/IDF per il doc x
Let i = j: (cj = c(x))
(somma di tutti i vettori in ci per ottenere pi)
Let pi = pi + d

**Algoritmo di Rocchio (test)Dato un documento di test x**
Let d = vettore TF/IDF per x
Let m = –2 (inizializzazione)
For i from 1 to n:
(calcola la similarità col vettore prototipo)
Let s = cosSim(d, pi)
if s > m
let m = s
let r = ci (aggiorna il più simile)
Return class r

**CLASSIFICATORE BASATO SU ESEMPI**
IDEA: Non si costruisce una rappresentazione della categoria, ma si confida sui documenti del training set che sono più vicini al documento che vogliamo classificare.

K-NN(K nearest neighbours)

VANTAGGI: k-NN, diversamente dai classificatori lineari non suddivide lo spazio dei documenti linearmente. Quindi risulta essere piu' "locale".
SVANTAGGI: Inefficienza a tempo di classificazione: k-NN deve calcolare la similarità di tutti i documenti del training set con il documento da classificare.
E' conveniente utilizzarlo per document-pivoted categorization: calcolare la somiglianza dei training document può essere fatto una volta per tutte le categorie.

Algoritmo di apprendimento Nearest-Neighbour
L'apprendimento si riduce al modo di immagazzinare le rappresentazioni degli esempi di training in D.
• Test dell'istanza x:
- Elabora la similarità tra x e tutti gli esempi in D.
- Assegna ad x la categoria del più simile in D.• Non si calcolano esplicitamente i prototipi delle categorie.

• Conosciuto anche sotto il nome di:
- Case-based
- Memory-based
  - Lazy learning

Nearest neighbor si basa su una metrica di similarità (o distanza)
- La più semplice per uno spazio contiunuo è la distanza euclidea.
- La più semplice per spazi d'istanza m-dimensionali binari è la distanza di Hamming
- Per i testi, la similarità basata sul coseno, per i vettori costruiti mediante indicizzazione TFIDF, è tipicamente la più efficiente.

**Training:**
For each each esempio di training <x, c(x)> ∈ D
Calcola il corrispondente vettore TF-IDF, dx, per il doc x

**Test dell'istanza y:**
Calcola il vettore TF-IDF d per il doc y
For each <x, c(x)> ∈ D
Let sx = cosSim(d, dx)
Ordina gli esempi, x, in D al decrescere di sx
Let N = I primi k esempi in D.
(ottiene così i vicini più simili)
Return la classe con più esempi in N

**SOCIETA' DELL'INFORMAZIONE**: diffusione dei dati e banche dati. Diffusione di internet. Comunicazioni senza fili. Convergenza delle tecnologie tra Informatica e Telefonia, Televisione, Fotografia, Cinematografia, ma anche Problematiche bancarie, identificazione personale. Più varietà di scelta, Più facili comunicazioni tra persone, Disbrigo di pratiche senza code, Reperibilità continua, Lavoro senza limite di tempo e di spazio, Più responsabilità, più competitività, Opportunità di emergere per chi ha talento. Necessità di tenersi aggiornati per non rischiare l'emarginazione(Info-ricchi e info-poveri), Rischio di violazione della sfera privata e della riservatezza, commercio concorrenza sconfinata, ingerenze, diffusione di epidemie (mucca pazza, sars, aviaria), Meno frontiere, più clandestinità, comunicazione, virus elettronici, inquinamento elettromagnetico. ICT: Tecnologie della comunicazione e informazione. Possono aiutare il processo di apprendimento,iniziando proprio dalla loro stessa conoscenza. Le ICT giocano un ruolo rilevante soprattutto nella formazione degli adulti, rappresentando una buona opportunità per favorire lo sviluppo delle competenze e delle esperienze, richieste dalla Società della Conoscenza. D'altronde a queste nuove Tecnologie è riconosciuta la potenzialità e la facilità di accesso per promuovere un apprendimento continuo e duraturo per tutta la vita. (LifeLong Learning) Tutto sta a fare in modo che l'adulto stesso apprezzi in prima persona l'utilità e la necessità di utilizzo di questi nuovi strumenti, superando così la diffidenza e le prime difficoltà. Vincoli legati ai docenti, le scuole, sistema scolastico. **DOC DIGITALI**: produzione, Immagazzinamento, Ricerca, Fruizione: queste funzioni che negli scritti tradizionali

sono separate, in quelli digitali sono integrate; 2. Multimedialità: I documenti digitali hanno formati diversi e completi; 3. Ipermedialità: Ogni documento può richiamare qualsiasi altro documento senza vincoli e confini;4. Riproducibilità: I documenti digitali o parti di essi sono riproducibili quasi a costo zero;5. Trasmissibilità e accessibilità: a livello globale senza limiti di spazio e di tempo;6.
Dinamicità,modificabilità e processabilità: attraverso il web possono essere cambiati, elaborati, condivisi, generati, eliminati;7. Computabilità e interattività: un documento digitale è a tutti gli effetti un programma con 0 e 1 e quindi può essere usato come modo di calcolo ed inoltre dà la possibilità di interagire agli utenti. Dato: Rappresentazione originaria e non interpretata di un fenomeno.
Informazione: Insieme di uno o più dati, messi in relazione ed interpretati, in modo da avere un significato e rappresentare un evento. L'Informazione non è altro che uno o più dati, sottoposti ad un processo che diventa significativo per il destinatario ed importante per il suo processo decisionale presente o futuro. I dati possono essere archiviati, le informazioni vengono prodotte per essere utilizzate. La conoscenza è poi proprio il modo in cui utilizziamo e mettiamo in relazione le informazioni, arricchendo ciò che sappiamo.
**ENERGIA INFORMATICA**: la conseguenza di questa evoluzione è il passaggio da Modelli di tipo Tayloristico (svolgimento di funzioni secondo leggi di tipo Meccanico) a Modelli Sistemici (in cui l'azienda è vista come un insieme integrato di uomini e strutture in continuo rapporto di osmosi tra loro).
**VALUTAZIONE INFORMAZIONE**:
l'Informazione, in quanto misura dell'ordine di un sistema, si presenta come opposta all'ENTROPIA (misura del disordine di un sistema fisico). La quantità di Informazione finale è minore di quella introdotta inizialmente. Si ha quindi una perdita, ovvero un aumento di entropia. Il valore dell'Informazione decresce ( V(I) ->0) se l'informazione non è fornita in tempo utile, se non viene comunicata nel luogo opportuno, infine è nulla se essa non è comprensibile. La Valutazione V(I) = F(t, I, s) dell'Informazione si può dare come :1. Misura del valore aggiunto;2. Misura sulla comparazione dei risultati che produce;3. Misura del grado di incertezza che riesce a ridurre.
**EQUAZIONI DI SHANNON**: I = -p log2 p dove I indica quanti bit sono necessari per avere certe informazioni, p la probabilità che l'evento si verifichi. C = W log2 (1 + S/N), dove C è la quantità di informazione che può essere trasmessa In bit al secondo dipende da W, ampiezza di banda e S/N rapporto Segnale-rumore. La prima equazione ci dice che più sono i messaggi a noi non noti, maggiori sono i bit di informazione necessari, ma Shannon ci dice anche esattamente quanti ne servono. Questa equazione ci dice che c'è un limite al numero di bit per secondo che possono essere trasmessi in un mezzo, un limite che è fissato dalla larghezza della banda e dal rumore del segnale. Il modo che permette di sfruttare questo limite con la max economia è attraverso la codificazione digitale.
**SISTEMA INFORMATIVO**: Un insieme di risorse che gestisce il minimo numero di info necessarie e realizza la massima efficacia comunicativa.
**AZIENDA**: L'azienda è un INSIEME di risorse (che il sistema azienda scambia con l'ambiente) e processi (attraverso i quali le risorse vengono scambiate e gestite). UNITA' ORGANIZZATIVE: in cui l'azienda è strutturata per svolgere i processi in modo efficace ed efficiente.
**PROCESSI**: quelli logistico-produttivi possono spaziare da una logica produttiva a processo continuo (settore petrolchimico) ad una produzione per piccole serie o singoli pezzi (officina meccanica), fino a una produzione di tipo cantieristico (edilizia) I processi di pianificazione e controllo devono essere disegnati in sintonia con le complessità e le caratteristiche dei sistemi fisici, tipo strategie di P. e C. per modelli di produzione per magazzino, assemblaggio o intera produzione per commessa, sino alla progettazione per commessa. E' opportuno notare che il grado d'adozione di tecnologie informatiche influenza estremamente i processi e può essere condizionato dal comportamento dei concorrenti. L'uso crescente di Internet costringe le aziende a ridefinire alcuni modelli gestionali, passando a modelli di azienda basati

sull’informazione ed alla nascita di nuove aziende digitali.