Text & Data Mining

A cura di Raffaele Costantino

Concetti chiave

- **Knowledge Discovery in Database (KDD)**
  - *"scoperta di conoscenza da databases "*: il processo d'estrazione di informazioni implicite, precedentemente sconosciute e potenzialmente utili da database (Frawley 1991)
- **Data Warehousing** (immagazzinamento di dati)
  - un Data Warehouse è un magazzino centrale di dati che sono stati estratti da dati operazionali (informazioni orientate al soggetto, non volatili e di tipo storico)
  - grosse quantità di dati (es. cartelle contenute nel computer di un ospedale, documenti nell'archivio di un comune)

Informazione e Produttività

- Il proliferare di dati e la capacità di immagazzinarli in grossi databases ci obbliga a adattare le nostre strategie e a sviluppare metodi meccanici per *filtrare*, *selezionare* e *interpretare* i dati.

- Le organizzazioni che eccelleranno in questo avranno una migliore chance di sopravvivenza e, proprio per questo, l'informazione stessa diventerà un fattore di produzione di grande importanza.

- La combinazione di Data Warehousing e Data Mining indica un approccio nuovo e totalmente diverso al management d'informazioni

Differenti modi di utilizzo

- **A partire da un Data Warehouse l'utente può desiderare sapere:**
  - Dove si trovano i dati;
  - Quali dati ci sono;
  - In che formato essi esistono;
  - Come questi sono in relazione con altri dati provenienti da altri databases;
  - Da dove arrivano e a chi appartengono.
- **È necessario avere un altro database che contiene meta-dati che descrivono la struttura dei contenuti del database.**

Minatori in rete

- **Analogia tra estrazione mineraria e reperimento informazioni in Internet**
  - Così come è necessario rimuovere un'enorme quantità di rifiuti prima che i diamanti o l'oro possano essere trovati, allo stesso modo con il computer e gli strumenti di data mining, possiamo automaticamente trovare l'unica informazione-diamante tra le tonnellate di dati-rifiuti nel database.

Distinzioni

- **Il processo di KDD si divide in**
  - **DM** (Data Mining)
    - estrazione di informazione da dati strutturati
  - **TM** (Text Mining) o **KDT** (Knowledge Discovery in Texts)
    - estrazione di informazione da databases testuali non strutturati

Data Mining

Data Mining I

- Processo di estrazione di conoscenza da banche dati di grandi dimensioni tramite l'applicazione di algoritmi che individuano le associazioni "nascoste" tra le informazioni e le rendono visibili.

Data Mining II

- Col nome *data mining* si intende l'applicazione di una o più tecniche che consentono l'esplorazione di grandi quantità di dati, con l'obiettivo di individuare le informazioni più significative e di renderle disponibili e direttamente utilizzabili nell'ambito del decision making.
- L'estrazione di conoscenza (informazioni significative) avviene tramite individuazione delle associazioni, o "patterns", o sequenze ripetute, o regolarità, nascoste nei dati.
  - pattern: una struttura, un modello, o, in generale, una rappresentazione sintetica dei dati.

Origini del Data Mining

- **Gli strumenti di data mining nascono dall'integrazione di vari campi di ricerca:**
  - statistica, "pattern recognition", machine learning
- Sono stati sviluppati indipendentemente dai database, per operare su dati "grezzi"
- Recenti sviluppi vedono una sempre maggiore integrazione tra strumenti di *data mining* (visto come una *query* avanzata) e databases.
- **Implicazioni**
  - apprendimento artificiale, statistica, tecnologia dei databases, sistemi esperti, sistemi di visualizzazione di dati, etc.

Le fonti

- **Testo: trascritto di materiale verbale che occorre naturalmente**
  - es. conversazioni, documenti scritti (diari o *report* di organizzazioni), libri, enciclopedie, risposte scritte a questionari aperti, registrazioni e descrizioni verbali di osservazioni.
  - **Solitamente**: database computerizzato di parole e frasi.

Problemi di partenza

- **Mancanza di visione a lungo-termine; "cosa ci aspettiamo dai nostri file in futuro"?**
- **Integrità dei dati:**
  - alcuni dati possono essere incorretti, non aggiornati o addirittura mancanti
- **Lotta o poca collaborazione tra dipartimenti e società (pubblicità dati)**
- **Restrizioni legali e/o di privacy:**
  - alcuni dati non possono essere usati per ragioni di privatezza
- **Alcuni file possono essere difficili o impossibili da connettere:**
  - discrepanza, ad esempio, tra databases gerarchici e relazionali
- **Problemi di interpretazione:**
  - connessioni tra file senza significato e/o erronee
  - relazioni inaspettate ma esistenti (casi di frode)

Tecniche di Data Mining I

- Il data mining è stato definito come un processo, all'interno del quale si utilizzano una o più tecniche per estrarre, da grandi quantità di dati, conoscenza in termini di associazioni, "pattern", regole, o sequenze ripetute.
- Le tecniche utilizzabili sono varie e, di conseguenza, anche gli algoritmi che le implementano. La scelta dipende principalmente dall'obiettivo che si vuole raggiungere e dal tipo di dati da analizzare.

Tecniche di Data Mining II

- La **regressione** (lineare, multipla e logistica), le **reti neurali supervisionate** e gli **alberi di decisione** consentono di effettuare operazioni di classificazione utilizzando la conoscenza acquisita in fase di addestramento per classificare nuovi oggetti o prevedere nuovi eventi.
  - Nelle applicazioni di Database Marketing lo scopo della classificazione predittiva è distinguere, ad esempio, i clienti in base alla probabilità di assumere un determinato stato (acquista/non acquista).

- Gli algoritmi di **clustering** l'uso delle reti neurali non supervisionate consentono di effettuare operazioni di segmentazione sui dati, cioè di individuare gruppi omogenei, o tipologie, che presentano delle regolarità al loro interno in grado di caratterizzarli e differenziarli dagli altri gruppi.
  - Ad esempio, segmentare i clienti esistenti in gruppi ed associare un profilo diverso per ciascuno al fine di ottimizzare l'attività di vendita.

Tecniche di Data Mining III

- Le tecniche di **associazione e sequenze** sono utilizzate per risolvere problemi di analisi delle affinità.
  - Lo scopo è di scoprire prodotti o servizi che sono frequentemente acquistati insieme (associazioni), o per analizzare i dati degli ordini per determinare cosa i clienti sono propensi a ordinare successivamente (sequenze). Questo può portare a studiare particolari combinazioni di prodotto o strategie di promozioni.
- Le tecniche di **analisi delle associazioni** consentono di individuare delle regole nelle occorrenze concomitanti di due o più eventi.
- A queste si aggiungono "sequential patterns" (tecniche di individuazione di sequenze temporali), "naive Bayes", algoritmi genetici, …

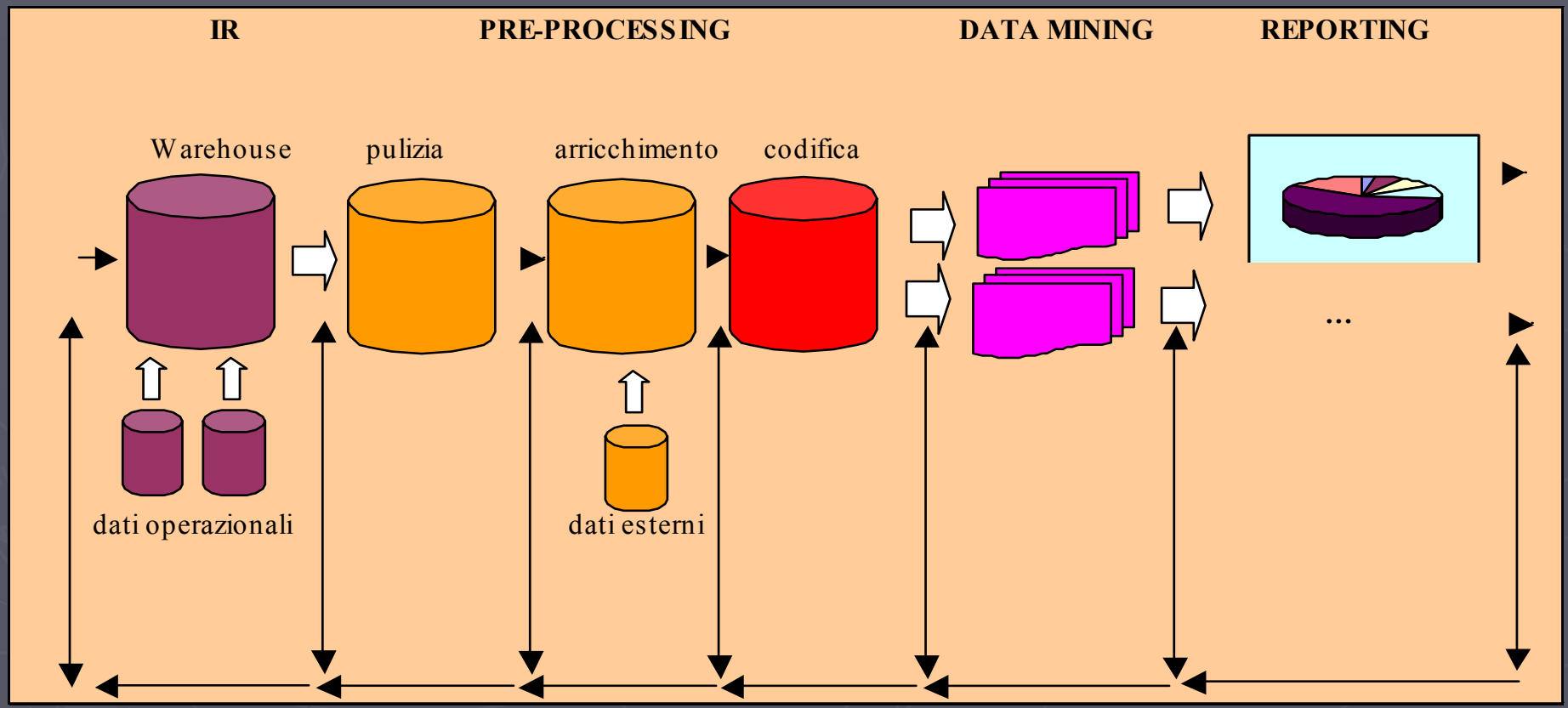
Obiettivi del Data Mining

- Sviluppare linguaggi specifici per *pattern-queries* e tecniche di ottimizzazione;
- Proporre una rappresentazione condensata per varie classi di pattern;
- Trovare strategie per lavorare con query fortemente relazionate;
- Combinare tecniche di Data Mining e statistiche;
- Utilizzare la conoscenza di fondo nel processo di KDD;
- Costruire attrezzi per selezionare, raggruppare e visualizzare la conoscenza scoperta.

Il processo I

- Il processo di Data Mining non è universale e molto spesso è costruito *ad hoc*
- È possibile proporre un *framework* (o struttura) generale delle fasi principali:
  1. Definizione degli obiettivi;
  2. IR (Information Retrieval);
  3. Pre-processing (preparazione dei dati):
     - Pulizia;
     - Arricchimento;
     - Codifica;
  4. Data Mining vero e proprio;
  5. Post-processing e reporting;

Il processo II



- Processo iterativo
- La fase più impegnativa è, generalmente, quella del pre-processing.

Definizioni degli obiettivi

- Definizione degli obiettivi a cui l'attività di analisi è preposta.
  - In campo aziendale, un tipico esempio è la selezione di un target per la promozione di un prodotto avente determinate caratteristiche

IR (Information Retrieval)

- Reperire i dati necessari per il raggiungimento degli obiettivi sopra definiti.
  - Le fonti dei dati possono essere interne, esterne, oppure una combinazione dettata dalla necessità di arricchire i dati con nuove dimensioni descrittive non presenti nel sistema informativo in esame.
  - Questa fase di ricerca è facilitata dalla presenza di un Data Warehouse organizzato per soggetti e contenente dati certificati.

Pre-processing

- La fase di pre-processing è fondamentale per la struttura di Data Mining
- I dati, in qualunque forma siano, vengono preparati per l'utilizzo successivo a seconda del tipo di trattamento a cui sono rivolti, del modello scelto e del software a disposizione.
- In generale, distinguiamo tre fasi principali di pre-processing: la **pulizia**, **l'arricchimento** e la **codifica**.

Pre-processing - pulizia

- Ci sono diversi tipi di processi di **pulizia** (*cleaning*), alcuni dei quali possono essere eseguiti in principio mentre altri sono utilizzati solo dopo che si è rilevato un disturbo nelle altre fasi del processo di Data Mining.
- Presenza di dati doppi
  - ad esempio un utente di un sito internet può essere registrato in due record a causa di una doppia registrazione o di un errore nel database clienti (nel database compaiono due Sig. Rossi con due numeri clienti diversi e uno stesso indirizzo: sorge il forte dubbio che i due clienti siano in realtà la stessa persona e che ci sia un errore nel numero del cliente. Non possiamo averne la certezza, ma un algoritmo di de-duplicazione che utilizza tecniche di riconoscimento di pattern potrebbe automaticamente identificare la situazione e presentarla all'utente).
- Mancanza di consistenza del dominio
  - Alcuni dati possono non essere veritieri ad es. una data di nascita improbabile o non corretta. Un buon programma dovrebbe essere in grado di catturare questi errori

Pre-processing - Arricchimento

- L'**arricchimento** è una fase a cui si dovrebbe poter sempre tornare in qualsiasi momento del processo di Data Mining, poiché in molti casi più informazioni si hanno più è possibile migliorare l'analisi.
- Le informazioni possono riguardare i clienti di un'organizzazione
  - ad esempio potrebbe essere utile, in un secondo tempo, sapere se il cliente possiede una carta di credito per valutare la possibilità di "vendita in rete", oppure possono essere informazioni aggiuntive che provengono dall'esterno.

Pre-processing - Codifica

- A seconda del tipo di dati possiamo proporre trasformazioni, o **codifiche**, differenti, ad esempio:
  - Selezione *record* con informazione di valore (cancellazione di quelli con dati mancanti)
  - Selezione record con dati mancanti (specialmente nelle analisi di scoperta di frode, infatti, dove ci può essere una connessione tra la mancanza di dati ed un certo comportamento del soggetto in questione)
  - Campionamento dei dati
- La codifica è un processo creativo: c'è, infatti, un numero infinito di codici differenti in relazione al numero di pattern che vogliamo trovare.

Data Mining

- Si parte dall'assunto che c'è più conoscenza nascosta nei dati di quella che si mostra in superficie.
- Ogni tecnica che ci aiuta a estrarre informazione dai dati è utile, ecco perché le tecniche proposte formano un gruppo abbastanza eterogeneo.
- Utilizzo tecniche statistiche, simboliche, sub-simboliche e di visualizzazione

Post-processing e reporting

- Il post-processing della conoscenza scoperta consiste in vari passi: dalla selezione ulteriore all'ordinamento, dalla visualizzazione all'estrazione di meta-informazione.
- Il processo di Data Mining non si ferma quando, ad esempio, sono stati scoperti dei pattern in un database. L'utente deve essere in grado di capire cosa è stato scoperto, vedere i dati e i pattern simultaneamente, confrontare i pattern scoperti con la conoscenza di fondo, etc.

Text Mining

Text Mining I

- Il Text Mining o Text Data Mining (TM o TDM) è l'estensione del Data Mining tradizionale su dati testuali non strutturati

- Obiettivo principale: estrazione di informazione implicitamente contenuta in un insieme di documenti e la visualizzazione di grossi set di testi.

Text Mining II

- Il TM è un campo più complicato del DM, perché lavora con i testi che non sono strutturati
- È un campo multidisciplinare, che impiega:
  - **l'Information Retrieval** (la raccolta di informazioni),
  - l'analisi testuale,
  - l'**Information Extraction** (l'estrazione di informazioni),
  - il clustering,
  - le tecniche di visualizzazione,
  - le tecniche di trattamento dei database,
  - l'apprendimento artificiale,
  - il Data Mining (l'accoppiamento della tecnologia della lingua con gli algoritmi del data mining)

Perché ha successo

- Le ragioni dell'attuale successo del text mining sono da ricercarsi:
  - nei recenti progressi delle tecniche di NLP (Natural Language Processing) e nella loro formalizzazione matematica,
  - nella disponibilità di applicazioni complesse e di potenza elaborativa attraverso gli ASPs (Application Services Providers),
  - nell'attenzione corrente di accademici, multinazionali del software, produttori di motori di ricerca verso tecniche di gestione della lingua, che ci fanno prevedere un forte sviluppo di questa tecnologia

Applicazioni I

- Le tecniche di text mining sono applicabili a qualsiasi ambito di indagine
- In generale trovano applicazione tutte le volte che siamo di fronte a grandi quantità di dati e abbiamo l'esigenza di conoscerne il contenuto.

Applicazioni II

Alcune delle applicazioni più comuni sono:

- Segmentazione della clientela (Database Marketing)
  - applicazione di tecniche di clustering al fine di individuare gruppi omogenei in termini di comportamento d'acquisto e di caratteristiche socio-demografiche; l'individuazione delle diverse tipologie di clienti permette di effettuare campagne di marketing diretto e di valutarne gli effetti, nonché di ottenere indicazioni su come modificare la propria offerta, e rende possibile monitorare nel tempo l'evoluzione della propria clientela e l'emergere di nuove tipologie
- Analisi delle associazioni (Basket Analysis)
  - applicazione di tecniche di individuazione di associazioni a dati di vendita al fine di conoscere quali prodotti sono acquistati congiuntamente; questo tipo d'informazione consente di migliorare l'offerta dei prodotti (disposizione sugli scaffali) e di incrementare le vendite di alcuni prodotti tramite offerte sui prodotti ad essi associati

Applicazioni III

- Analisi testuale (Text Mining)
  - applicazione di tecniche di clustering al fine di individuare gruppi omogenei di documenti in termini di argomento trattato; consente di accedere più velocemente all'argomento di interesse e di individuarne i legami con altri argomenti
- Technology Watch (Competitive Intelligence)
  - applicazione di tecniche di clustering a banche dati di tipo tecnico-scientifico al fine di individuare i gruppi tematici principali (nel caso di banche dati di brevetti, un gruppo tematico indica una particolare tecnologia), le loro relazioni, l'evoluzione temporale, le persone o le aziende coinvolte
- Applicazioni in rete
  - applicazione nei motori di ricerca o di tecniche di filtraggio di informazioni indesiderate (es. POESIA Project)

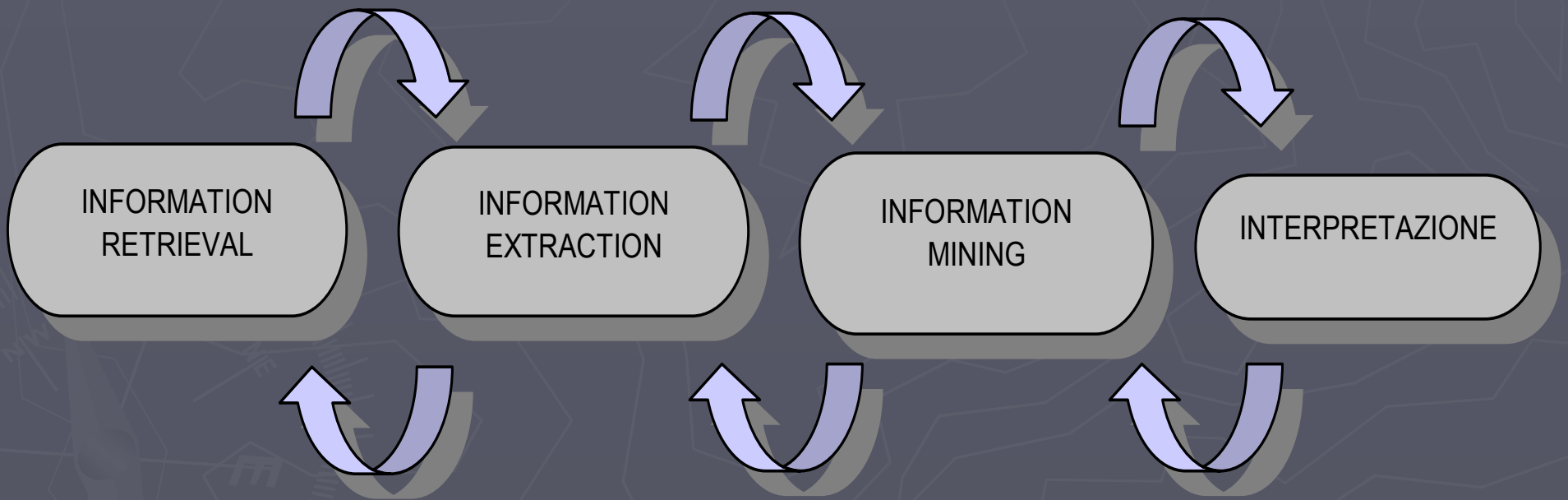
Fonti Text Mining I

- Web Data (siti web)
  - Internet sta diventando il principale "media" attraverso cui è possibile ottenere documenti, dati ed informazioni. I siti web liberamente raggiungibili via Internet sono una delle fonti principali della documentazione da analizzare (filtraggio informazioni)
- Banche dati online
  - Le banche dati online costituiscono collezioni di informazioni specializzate, generalmente accessibili via Internet tramite abbonamento. Esempi tipici di queste banche dati sono quelle dedicate alle pubblicazioni, ai brevetti o agli articoli scientifici (di chimica, fisica o matematica) rese disponibili in modo diretto o attraverso information broker.
- Sorgenti informative private
  - Una banca dati privata di documenti elettronici (costruita negli anni) può essere resa disponibile ed essere opportunamente usata insieme alle altre sorgenti informative. Il formato ed i contenuti dei documenti di una banca dati privata sono generalmente completamente differenti da quelli dei documenti ottenuti attraverso le banche dati online.

Fonti Text Mining II

- e-mail
  - Le e-mail sono la forma più ricca dal punto di vista informativo e più semplice da analizzare. E' il mezzo attraverso cui le persone comunicano all'interno ed all'esterno di aziende ed organizzazioni. Possono essere analizzate sia le e-mail interne ad una organizzazione sia quelle ricevute dall'esterno od inviate all'esterno dell'organizzazione.
- Opinion surveys
  - Spesso le opinion surveys sono analizzate con cura nella parte codificata, dove è prevista la risposta: SI, NO, o numerica. Sono invece analizzate in maniera superficiale nella parte testuale, ove si raccolgono le risposte in testo libero alle domande aperte.
- Newsgroups, Chatlines, Mailing Lists
  - Importanti e ricche fonti di informazione dato che riguardano i temi più disparati, dai consumi alla politica. Il problema con questo tipo di informazione è che l'informazione pertinente è all'interno di frasi e/o affermazioni di scarsa importanza, espresse con linguaggio spesso gergale. Grazie al text mining queste affermazioni/opinioni possono essere analizzate e filtrate al fine di conoscere quali sono le opinioni di chi scrive.

Il processo


INFORMATION RETRIEVAL
INFORMATION EXTRACTION
INFORMATION MINING
INTERPRETAZIONE

Information retrieval I

- Localizzare e recuperare documenti che possono essere considerati rilevanti alla luce degli obiettivi prefissati.
- L'utente del sistema può specificare il set di documenti, ma l'operazione necessita comunque di un sistema che filtri i testi irrilevanti.
- Solitamente col termine "Information Retrieval" si identifica la raccolta di testi tra quelli che ipotizziamo trattare lo stesso argomento, ma più genericamente possiamo intendere anche la semplice raccolta di informazioni testuali per una successiva analisi.

Information retrieval II

- Ha come obiettivo la selezione di un sottoinsieme rilevante di documenti da un insieme più grande e tenta di rappresentare tutto il contenuto informativo di una forte parte delle informazioni contenute nel testo.
- Il termine IR fa riferimento all'attività di ricerca di documenti attraverso delle parole chiavi o composizioni logiche delle stesse (query), le quali a loro volta sono utilizzate per indicizzare i documenti.

Information extraction I

- Estrazione di informazioni dai documenti selezionati.
- Di solito si tratta di riempire specifici template di informazioni, ma in questa fase stanno anche tutte le tecniche di pruning e di estrazione di conoscenza generica.
  - *Template*: tabelle che contengono dei dati semi-strutturati. Possono esserci informazioni quantitative e qualitative. Solitamente si utilizza un template di base che verrà compilato per ogni documento testuale che viene analizzato.
  - *Pruning*: letteralmente significa "potatura". Si tratta di un'applicazione di una serie di tecniche atte a pulire i dati da elementi non interessanti per alleggerirne il trattamento.

Information extraction II

- L'IE può, quindi, essere considerata come un'attività di supporto all'IR.
- L'IR fa riferimento all'attività di ricerca di documenti attraverso delle parole chiavi, ma spesso questo metodo non porta al recupero di documenti realmente interessanti per il nostro scopo perché le chiavi sono scelte da terzi (nella maggior parte dei casi dall'autore del testo).
- L'IE cerca di superare questa differenza tra le due logiche in modo da avere un meccanismo di ricerca che sia basato su una rappresentazione oggettiva della conoscenza.

IE in Internet

- Un sistema di IE risulta utile come passo successivo per i motori di ricerca per il Web nell'adempiere alle necessità del ritrovamento di informazione.
- L'IE mira a sviluppare delle metodologie capaci di elaborare il testo dei vari documenti e di estrarre, come risultato di questa elaborazione, dei concetti che permettono di descrivere il contenuto del documento stesso.

Il processo di IE I

- Il processo relativo ad un sistema di Information Extraction si suddivide in due parti principali:
  - prima il sistema estrae fatti individuali dal documento attraverso un'analisi locale del testo;
  - poi i fatti estratti vengono integrati con l'analisi di coreferenza e di inferenza.
- Infine, dopo tale fase di integrazione, i fatti pertinenti vengono tradotti nel formato di output richiesto

Fasi del processo

- Analisi lessicale
  - consente di assegnare alle singole parole part-of-speech ed altre caratteristiche attraverso l'analisi morfologica
- Riconoscimento di nomi
  - ha lo scopo di identificare i nomi ed altre speciali strutture lessicali (ad esempio date, locuzioni, ecc.)
- Analisi sintattica (completa o parziale)
  - consente di individuare i gruppi nominali, i gruppi verbali, altre strutture sintattiche di interesse, e le teste di tali gruppi
- Individuazione dei fatti di interesse
  - i fatti vengono integrati e combinati con altri fatti presenti nel documento, attraverso l'analisi del discorso. Tale analisi risolve le relazioni di coreferenza che vi sono, ad esempio fra i pronomi o fra descrizioni multiple di uno stesso evento. Vengono anche "inferiti" nuovi fatti a partire da quelli già esplicitamente asseriti nel testo
- Generazione dei template
  - i dati vengono infine ordinati e rappresentati sotto forma di una tabella di output

IE: Analisi lessicale

- Il testo viene prima diviso in frasi e token. Ciascun token viene ricercato all'interno di un dizionario per determinarne i possibili part-of-speech ed altre caratteristiche.
- Generalmente tali dizionari includono una raccolta di nomi di società, abbreviazioni, suffissi di compagnie ed altro.
- Questa fase è composta dall'identificazione del linguaggio, dalla tokenizzazione, dall'analisi morfologica e dal part-of-speech tagging.

IE: Riconoscimento di nomi

- La fase successiva del processo identifica i vari tipi di nomi propri ed altre forme speciali, come dati e cifre.
- I nomi propri appaiono frequentemente in molti tipi di testi e la loro identificazione e classificazione semplifica le successive fasi di elaborazione.
- I nomi vengono identificati tramite un set di pattern (espressioni regolari) espresse nei termini del part-of-speech, delle caratteristiche sintattiche e delle caratteristiche ortografiche (ad es. l'iniziale maiuscola).

IE: Analisi sintattica I

- Consiste nell'identificazione di legami sintattici elementari fra i diversi elementi della frase.
- Un'analisi sintattica profonda di una frase ha generalmente come risultato una foresta di alberi di derivazione sintattica, ciascuno dei quali fornisce una possibile interpretazione sintatticamente corretta della frase stessa.
- Gli argomenti da estrarre spesso corrispondono a frasi di nomi nel testo, mentre le relazioni di solito corrispondono a relazioni grammaticali.

IE: Analisi sintattica II

- Alcuni sistemi di IE tentano di costruire un parsing completo della frase. La maggior parte di questi ultimi falliscono in ciò, e costruiscono, allora, diversi strati di parsing.
- Una delle più importanti strutture, formate da più parole, che si possono facilmente riconoscere dopo la fase di "part-of-speech tagging" è la semplice frase nominale (cioè, una porzione di frase in cui compaiono nomi, ma non verbi) in quanto l'individuazione di strutture sintattiche complete si rivela piuttosto difficile.

IE: Pattern maching

- Il pattern matching consiste nell'estrazione di eventi o relazioni rilevanti per lo scenario di interesse

IE: analisi di coreferenza

- L'analisi di coreferenza si pone come obiettivo la risoluzione dei riferimenti dei pronomi ed anche di frasi di nomi che esprimono cose già dette nel testo.

IE: inferenze

- Può accadere che informazioni relative ad uno stesso evento siano sparse in diverse frasi.
- È necessario, allora, riunire tali informazioni prima della generazione dei template o degli output.
- Quando invece sono presenti delle informazioni non esplicitamente indicate nel testo si fa uso del meccanismo dell'inferenza per renderle esplicite.

IE: generazione dei template

- Tutte le informazioni finora ricavate dal testo sono sufficienti per l'estrazione dei template, un particolare tipo di output.
- Questi sono frame (tabelle) con slot da riempire con le informazioni richieste.
- Da una stessa porzione di testo possono essere estratti più template in base al numero di eventi di interesse citati nello stesso.

Information Mining

- Una volta compilato un template per ogni documento analizzato, abbiamo, di fatto, un database che è compatibile con le tecniche usuali di Data Mining.
- In questo passo cerchiamo se esistono dei patterns o delle relazioni fra i dati. Nel caso di analisi di un testo unico, questa fase corrisponde alle tecniche di analisi della conoscenza estratta, comprendenti metodi statistici e metodi simbolici.

Interpretazione

- Il passo finale consiste nell'analizzare i risultati e interpretare i pattern scoperti durante la fase di mining.
- Idealmente, l'interpretazione dovrebbe essere in formato di linguaggio naturale.

The POESIA Project

http://www.poesia-filter.org

Internet e pornografia I

- L'utilizzo di Internet si è rapidamente diffuso tra i giovani
- Educatori e famiglie sono preoccupati per la crescita dei siti a carattere pornografico i quali attirerebbero l'attenzione degli adolescenti
- Crescente scetticismo nei confronti della Rete

Internet e pornografia II

- Dato che non è possibile controllare totalmente la diffusione di materiale osceno via Internet è necessario limitarne o controllarne l'accesso

POESIA Project I

- Public Open-Source Environment for a Safer Internet Access (iniziato nel Febbraio 2001)
- Fondato dalla commissione europea nell'ambito del "Information Society and Technology Safer Internet Action Plan" (con l'obiettivo di controllare il diffondersi in Rete di materiale pericoloso, illegale, osceno e con contenuti razzisti)
- Il progetto POESIA ha lo scopo di sviluppare, testare, valutare e promuovere dei metodi di filtraggio delle informazioni diffuse via Internet
- È un software completamente open-source quindi modificabile e aggiornabile.

POESIA Project II

- Partners del POESIA Project
  - Istituto di Linguistica Computazionale (Italy)
  - Commissariat à l'Energie Atomique (France)
  - Ecole Nouvelle d'Ingénieurs en Communication (France)
  - M.E.T.A. S.r.l. (Italy)
  - Universidad Europea de Madrid CEES (Spain)
  - University of Sheffield (UK)
  - Fundació Catalana per a la Recerca (Spain)
  - PIXEL Associazione (Italy)
  - Liverpool Hope University College (UK)
  - Telefónica Investigación y Desarrollo (Spain)

POESIA Project III

- I creatori di POESIA si augurano che questo possa diventare uno standard nei metodi di filtraggio dei contenuti della Rete
- POESIA è progettato per supportare le attività di scuole, librerie e uffici dove vi sono gruppi di computer collegati tra loro e con l'accesso ad Internet

POESIA Project IV

- I filtri di POESIA operano su diversi canali
  - Web
  - E-mail
  - News
- Sono utilizzati diversi e sofisticati metodi di filtraggio dati quali ad esempio:
  - Filtraggio dei testi (natural language text filtering)
  - Filtraggio delle immagini
  - Controllo indirizzi URL
  - Filtraggio JavaScript
- Sono supportati diversi linguaggi quali inglese, italiano e spagnolo

Filtraggio dei testi I

- POESIA utilizza alcuni algoritmi di Text Mining per analizzare i contenuti delle pagine web
- Ad esempio utilizza un disambiguatore linguistico
  - Alcune espressioni multiword possono essere utilizzate in contesti diversi – Esempio:
    - Siti informazione sessuale
    - Siti pornografici

Filtraggio dei testi II

- Il filtraggio dati avviene in due fasi
  - Inizialmente un semplice filtering agent che implementa funzioni di NLP analizza rapidamente i dati (anche in elevate quantità) e, dopo averli classificati, individua quelli che dovranno essere ulteriormente analizzati
  - Un altro agente più sofisticato e preciso analizza e scansiona i dati che il primo non è stato in grado di classificare con esattezza

Metodi analisi

- Metodi di NLP utilizzati dagli agenti di POESIA:
  - Estrazione automatica da un corpus di dati (singole parole, espressioni particolari, parole multiword, parole ambigue, categorie ecc.)
  - Costruzione di un dizionario semantico e lessicale di dominio
  - Algoritmi di riconoscimento di espressioni linguistiche
    - Tokenizzazione
    - Analisi morfologica e sintattica
    - Riconoscimento di entità linguistiche
    - Segmentazione di testi
    - Riconoscimento relazioni grammaticali

Adattabilità

- I filtri di POESIA sono naturalmente dinamici e si adattano alla natura del linguaggio che devono analizzare
- I contenuti del Web sono infatti in costante aggiornamento e non mancano i tentativi di aggirare i metodi di filtraggio
- È previsto un addestramento all'utilizzo di POESIA