Arricchimento semantico e classificazione di informazioni testuali

Giuliano Armano

DIEE – Univ. di Cagliari

Macomer, 25 Ottobre 2016

Sommario

- Contesto operativo
- Qualche nota sulle tecnologie
- Focus sull'arricchimento semantico
- Focus sulla classificazione automatica
- Conclusioni

- Contesto operativo
- Qualche nota sulle tecnologie
- Focus sull'arricchimento semantico
- Focus sulla classificazione automatica
- Conclusioni

Contesto Operativo

- Le biblioteche e i bibliotecari dovranno essere aperti sempre di più verso il WEB per ...
  - Fornire supporto ad attività di ricerca di testi
  - Fornire suggerimenti di lettura
  - Gestire e orientare di gruppi di interesse
  - ...

Contesto Operativo

- Fornire supporto ad attività di ricerca di testi
  - Gli utenti sono ormai abituati a usare motori di ricerca e si aspettano che anche per la ricerca di testi si utilizzi lo stesso approccio
- Tra le tecnologie utilizzate ricordiamo …
  - Classificazione automatica di testi (ed eventualmente di materiale multimediale)
  - Clustering

Contesto Operativo

- Fornire suggerimenti di lettura
  - La modalità “push” è ormai uno standard nella pubblicità per la fornitura di servizi
  - In altre parole non si aspetta che sia l'utente a chiedere, ma si inviano informazioni per invogliarlo/a a comprare merci o fruire di servizi
- Tra le tecnologie utilizzate ricordiamo …
  - Profilazione utente

Contesto Operativo

- Gestire e orientare gruppi di interesse
  - Siamo nell'era dei social media (facebook, twitter, instagram, ecc.), quindi occorre adeguare le modalità di interazione con l'utenza in accordo con questa prospettiva
- Tra le tecnologie utilizzate ricordiamo ...
  - Clustering
  - Reti complesse
  - Sentiment analysis

- Contesto operativo
- **Qualche nota sulle tecnologie**
- Focus sull'arricchimento semantico
- Focus sulla classificazione
- Conclusioni

Qualche Nota sulle Tecnologie

- In precedenza, a seconda del problema, abbiamo citato varie metodologie / tecniche …
  - Classificazione automatica
  - Clustering
  - Profilazione utente
  - Reti complesse
  - Sentiment analysis

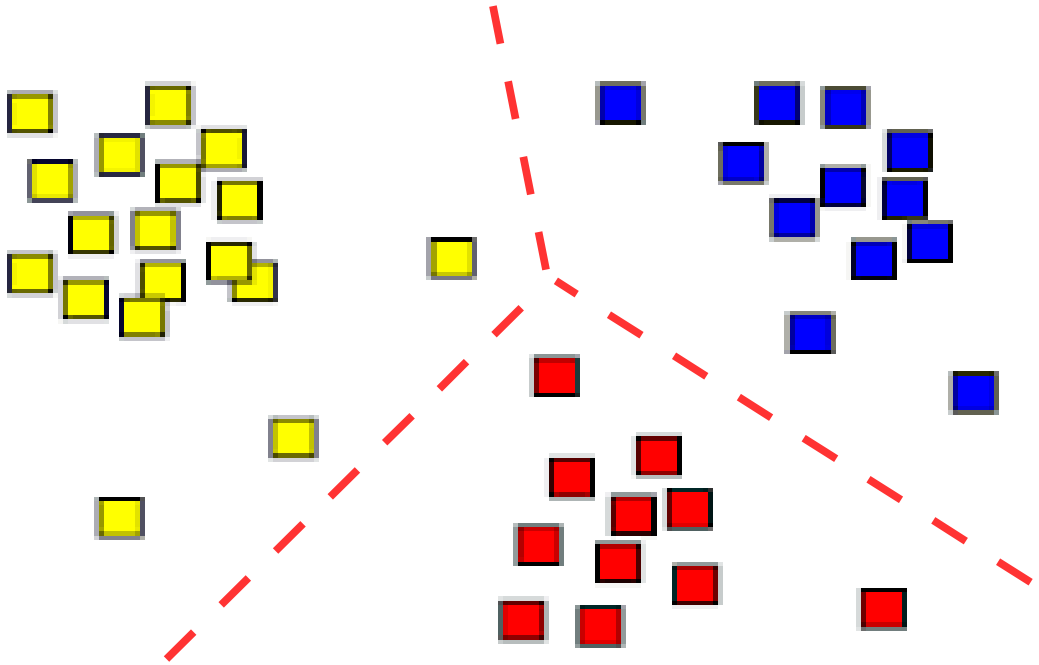
Vediamo sommariamente di che si tratta …

Qualche Nota sulle Tecnologie

- **Classificazione automatica (un esempio)**
  - Supponiamo di avere la descrizione di un oggetto e di volerne trovare la categoria di appartenenza tra quelle date
  - In questo caso addestrerò un sistema automatico insegnandogli a riconoscere l'appartenenza di un libro alle varie categorie date
  - Si tratta di un problema di apprendimento (automatico) supervisionato

Qualche Nota sulle Tecnologie

- **Classificazione automatica (un esempio)**



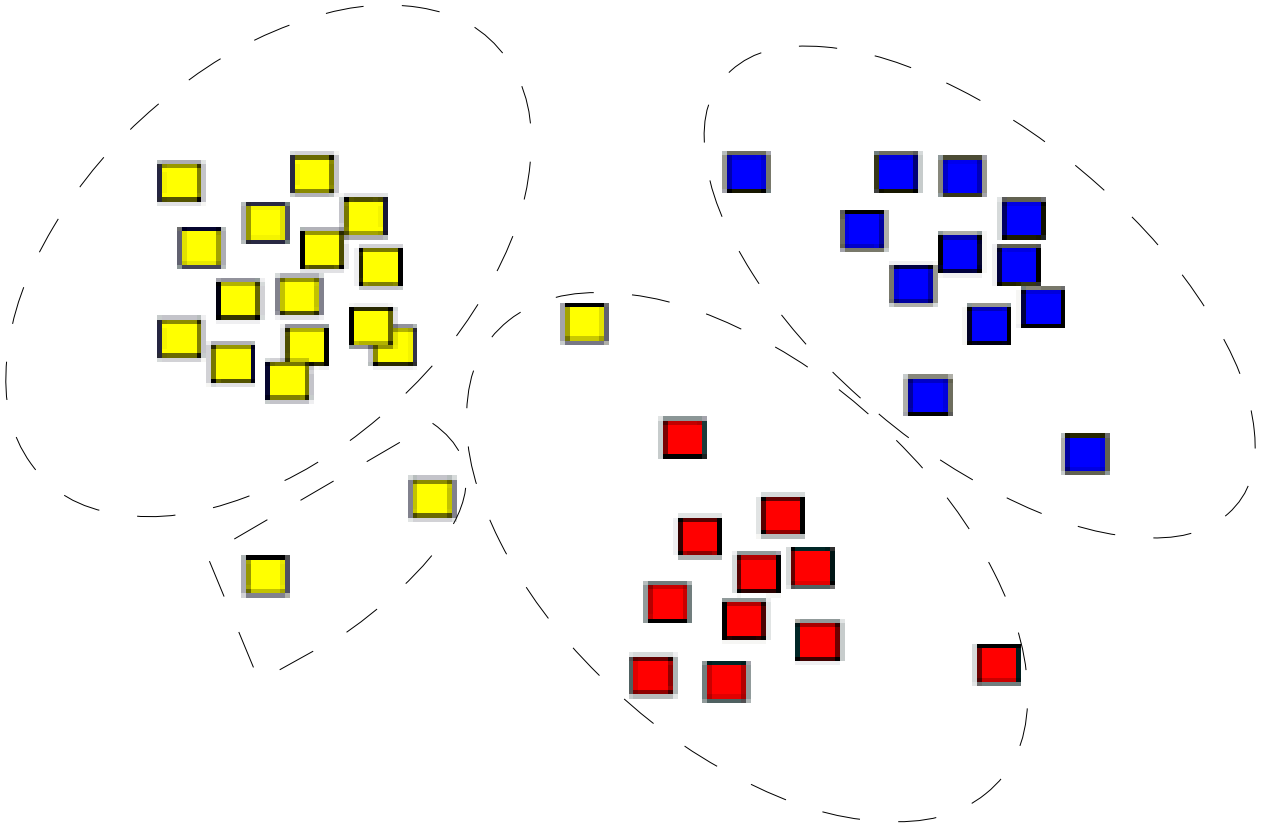
Il sistema automatico dovrà apprendere la linea di separazione tra gli oggetti appartenenti alle tre classi (marcate in giallo, blu e rosso)

Qualche Nota sulle Tecnologie

- **Clustering (un esempio)**
  - Supponiamo di avere la descrizione di tanti oggetti e di volerli raggruppare, senza però avere alcuna idea a priori di come farlo
  - In questo caso chiederò a un sistema automatico di individuare i raggruppamenti (al più indicandogli quanti ne voglio ottenere)
  - Si tratta di un problema di apprendimento (automatico) non supervisionato

Qualche Nota sulle Tecnologie

- **Clustering (un esempio)**



Il sistema automatico non sa che gli oggetti in realtà appartengono a classi diverse, e quindi tenta di raggruppare gli elementi in base a caratteristiche comuni

Qualche Nota sulle Tecnologie

- **Profilazione utente**
  - Quando un utente accede a un sito WEB, tipicamente gli viene chiesto di confermare l'uso dei cookies
  - Quando un utente si iscrive a facebook (o ad altro social network), lo si avvisa che i suoi dati personali saranno utilizzati per varie attività informative
  - Bene, questi sono soltanto due tra i tanti esempi in cui siamo avvertiti del fatto che qualcuno laggiù nel profondo WEB ci “profilerà” per cercare di capire quali acquisti di beni o servizi raccomandarci …

Qualche Nota sulle Tecnologie

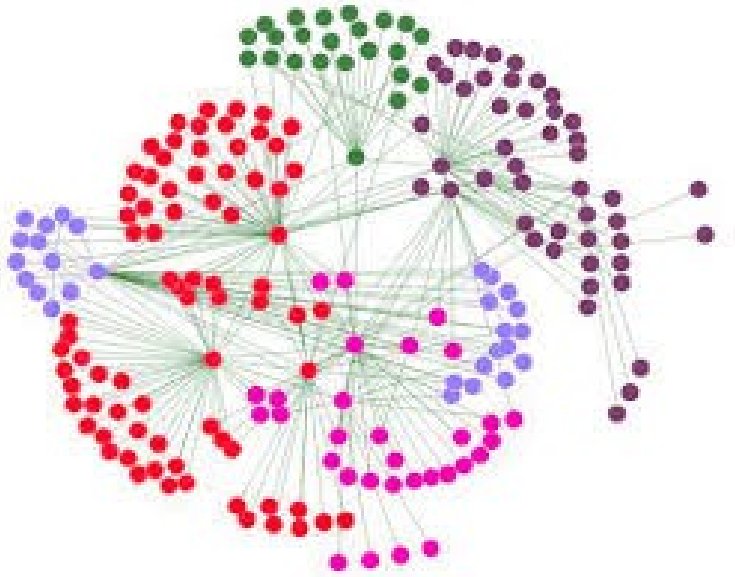
- **Profilazione utente**
  - **Ci sono due principali tecniche di profilazione:**
    - **Basata sul contenuto**
      - **Per esempio quando usiamo un motore di ricerca tutte le ns "query" possono essere analizzate per capire le ns preferenze e/o interessi …**
    - **Collaborativa**
      - **Si usa tipicamente una matrice utenti-prodotti (o utenti-servizi), in cui una riga rappresenta anche le ns scelte passate …**
      - **Al momento di capire se un prodotto o servizio può essere di ns interesse si guarda in che misura utenti simili a noi hanno già effettuato quella scelta …**

Qualche Nota sulle Tecnologie

- **Reti complesse**
  - **Sono uno strumento essenziale per lo studio dei comportamenti emergenti generati da numerosi elementi (cose, concetti o persone) tra i cui membri esistono una o più relazioni**

Qualche Nota sulle Tecnologie

Un esempio di rete complessa



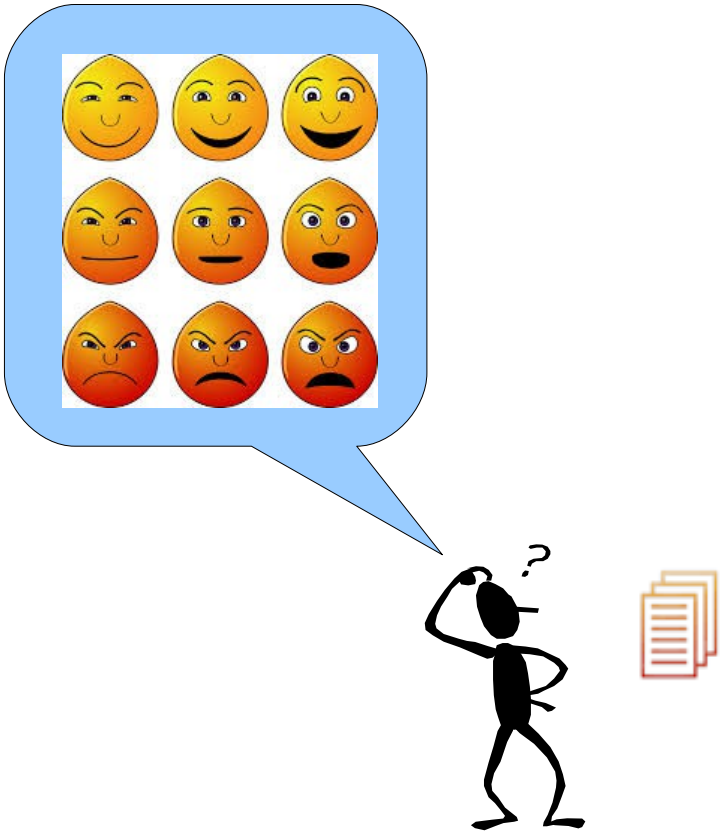
- ✓ La rete potrebbe rappresentare interessi culturali di persone appartenenti a un determinato social networking
- ✓ Immaginiamo che i vari interessi culturali siano rappresentati dai colori diversi
- ✓ Si vede chiaramente che esistono degli “hub”, ovvero persone che costituiscono un riferimento per ogni interesse culturale
- ✓ Poi si possono calcolare anche vari parametri della rete, che ci danno maggiori informazioni sul fenomeno (per esempio: topologia, robustezza, resilienza)

Qualche Nota sulle Tecnologie

- **Sentiment Analysis**
  - **È finalizzata a identificare l'umore di un utente ("mood"), tipicamente analizzando i contenuti che l'utente stesso lascia sul WEB (per esempio annotazioni sulla sua pagina facebook)**
  - **La sentiment analysis è ampiamente applicata per analizzare social media per una varietà di applicazioni, dal marketing al servizio clienti**

Qualche Nota sulle Tecnologie

Sentiment Analysis



- ✓ Supponiamo di voler identificare il contenuto “emozionale” di messaggi scambiati tra componenti di un social networking
- ✓ In particolare vogliamo classificare ogni messaggio in base a tre distinte emozioni: a) apprezzamento, b) neutralità e c) contrarietà rispetto all'argomento discusso nel messaggio
- ✓ A sua volta ogni emozione potrebbe avere (per esempio) tre gradi diversi ...

Qualche Nota sulle Tecnologie

- Tipicamente le tecniche citate richiedono una fase preliminare di codifica dei dati, che si ottiene tramite:
  - Analisi del testo
  - Analisi di metadati
  - Arricchimento semantico

Qualche Nota sulle Tecnologie

- **Analisi di testo**
  - **Può essere effettuata …**
    - **su base statistica, ovvero a partire dalla frequenza con cui le varie parole occorrono all'interno dei documenti in fase di studio**
    - **su base semantica, ovvero analizzando i documenti in fase di studio tramite tecniche e algoritmi di analisi del linguaggio naturale**

Qualche Nota sulle Tecnologie

Analisi di metadati

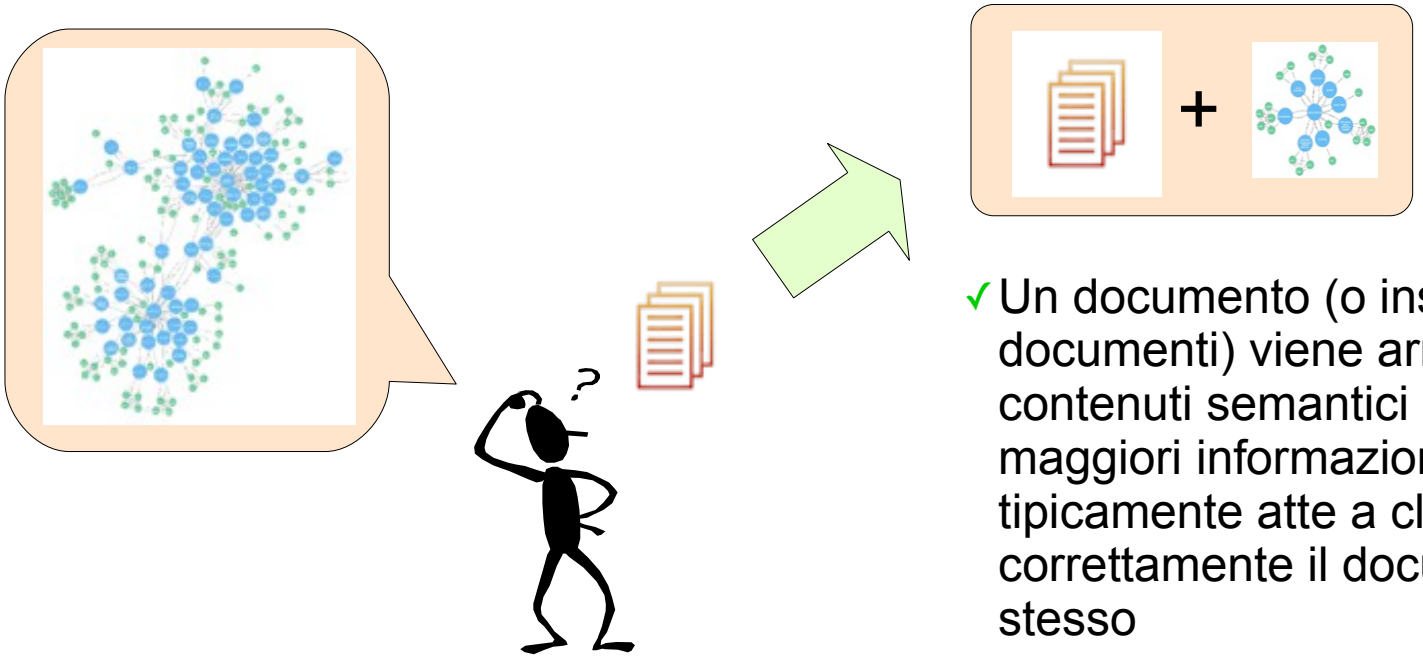
- **Tipicamente i metadati sono già presenti in forma standard (per esempio tramite codifiche XML), e quindi è relativamente facile utilizzarli per gli scopi che ci si prefigge**

Vediamo un semplice esempio di codifica XML di metadati ...

```
<?xml version="1.0" encoding="ISO-8859-1"?>
  <libri>
    <libro titolo="I tre moschettieri" autore="A. Dumas"
           edizione="Donzelli, Roma" anno="2014" />
    … // qui metto altri libri
  </libri>
</xml>
```

Qualche Nota sulle Tecnologie

Arricchimento semantico



- ✓ Un documento (o insieme di documenti) viene arricchito con contenuti semantici per fornire maggiori informazioni, tipicamente atte a classificare correttamente il documento stesso
- ✓ Esempio di arricchimento semantico: si usa **WordNet** per accedere ai sinonimi, agli iperonimi, agli iponimi ecc.

- Contesto operativo
- Qualche nota sulle tecnologie
- **Focus sull'arricchimento semantico**
- Focus sulla classificazione
- Conclusioni

Arricchimento Semantico

- **Un modo tipico per realizzare l'arricchimento semantico è quello di usare un dizionario semantico-lessicale**
  - **WordNet (Fellbaum, 1998) è uno dei più diffusi dizionari semantico-lessicali**
  - **Ora disponibile anche in formato multilingua (^), si avvale di raggruppamenti di termini con significato affine, chiamati synset (= synonym set)**

(^) Disponibile anche per l'italiano su → http://multiwordnet.fbk.eu

Arricchimento Semantico

- I vari synset sono collegati tra loro attraverso diversi tipi di relazioni definite in particolare per …
  - Sostantivi
  - Verbi

Arricchimento Semantico

- L'arricchimento semantico e l'uso di WordNet
  - Principali tipi di relazioni per i sostantivi
    - Iperonimia – Y iperonimo di X se ogni X è anche Y
    - Iponimia – Y iponimo di X se ogni Y è anche X
    - Coordinazione – Y termine coordinato di X se entrambi hanno un iperonimo in comune;
    - Olonimia – Y olonimo di X se X è parte Y
    - Meronimia – Y meronimo di X se Y è parte X

Arricchimento Semantico

- L'arricchimento semantico e l'uso di WordNet
  - Principali tipi di relazioni per i verbi
    - Iperonimia – Y è iperonimo di X se l'attività X è "una specie di" Y
    - Troponimia – Y è troponimo di X se nel fare Y si fa anche X
    - Implicazione – Y implica il verbo X se nel fare X si deve fare anche Y
    - Coordinazione – Y termine coordinato di X se X e Y hanno un iperonimo in comune

Arricchimento Semantico

- Come si utilizzano i synset ?
  - Scenario tipico:
    - Da un documento di testo posso calcolare facilmente la cosiddetta “bag-of-words”, cioé l'insieme delle parole contenute nel testo (con eventualmente anche la loro frequenza interna)
    - Dalla bag-of-words posso poi passare alla “bag-of-synsets”, estraendo per ogni parola i synset a cui partecipa
    - ATTENZIONE però! In realtà avrebbe senso inserire nella bag-of-synsets soltanto synset pertinenti, ma identificare (per ogni parola) il synset semanticamente pertinente non è mai un'operazione agevole …

Arricchimento Semantico

- **Come si utilizzano i synset?** (seguito)
  - **Per esempio, se nel ns documento compare la parola "aquila", su MultiWordNet troviamo che appartiene a 2 synset …**
    - **1. Pos: n 1. aquila, cervello, cima, genio, ingegno, mente, mente_superiore (Factotum) [someone who has exceptional intellectual ability and originality]**
    - **2. Pos: n 2. aquila (Animals, Biology) [any of various large keen-sighted diurnal birds of prey noted for their broad wings and strong soaring flight]**
  - **Quale sarà allora il synset da selezionare ?**

NB In questo caso la parola scelta ha soltanto due synset, ma tipicamente il numero di synset a cui una parola può partecipare è più elevato ...

- Contesto operativo
- Qualche nota sulle tecnologie
- Focus sull'arricchimento semantico
- **Focus sulla classificazione automatica**
- Conclusioni

Classificazione Automatica

- Come già accennato, …
  - Dato un insieme di categorie e un oggetto da classificare, la classificazione automatica consiste nell'identificare la categoria (o le categorie) a cui l'oggetto appartiene
  - Quindi, con N numero di categorie, abbiamo …
    - N=2 :: Classificazione binaria (o dicotomica)
    - N>2 :: Classificazione multiclasse, quando a ogni oggetto può essere associata una sola categoria
    - N>2 :: Classificazione multilabel, quando a ogni oggetto possono essere associate più categorie

Classificazione Automatica

- Di cosa abbiamo bisogno per generare un classificatore automatico? Di un algoritmo che lo generi!
  - Ci sono quindi (tipicamente) due fasi distinte:
    - Apprendimento

      in cui a partire dai dati a disposizione (^) viene generato il classificatore
    - Uso

      in cui il classificatore viene utilizzato per identificare la categoria (o le categorie) da associare a ogni nuovo oggetto da classificare

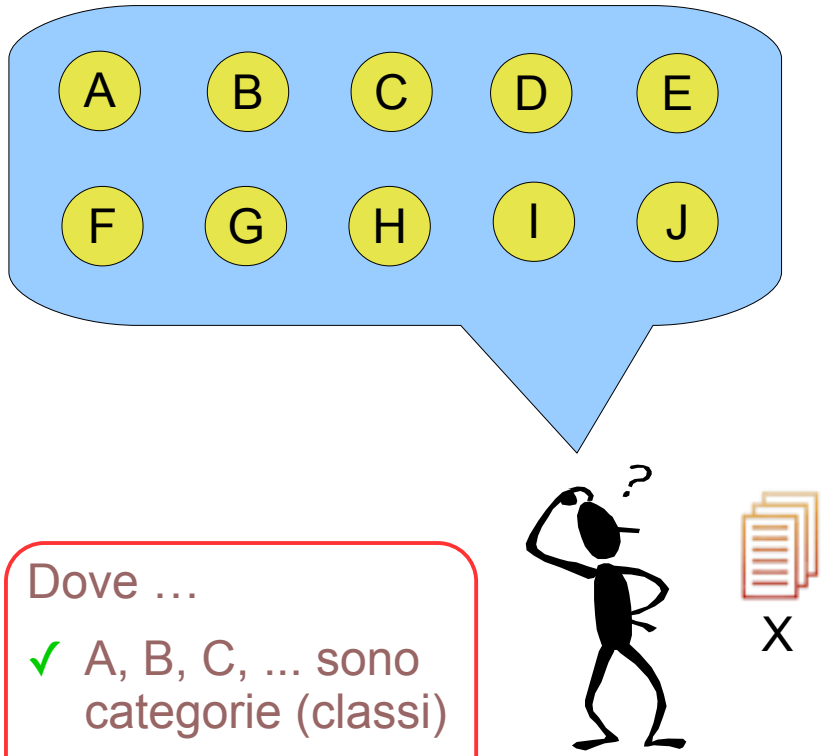
(^) … di cui dobbiamo conoscere l'associazione con le categorie di interesse

Classificazione Automatica

- **Prima ...**
  - Vediamo il problema della classificazione per il caso specifico in cui gli oggetti da classificare siano documenti testuali (per esempio la descrizione di un libro)
- **Poi ...**
  - faremo un breve cenno alle tecnologie utilizzabili per generare un classificatore capace di fornire risultati accettabili ...

Classificazione Automatica

- **Un caso classico: la classificazione di Dewey**



Dove …

- ✓ A, B, C, ... sono categorie (classi)
- ✓ X è un oggetto da classificare

Per esempio…

- ✓ A = Informatica, ecc.
- ✓ B = Filosofia e discipline connesse
- ✓ C = Religione
- ✓ D = Scienze sociali
- ✓ E = Linguistica
- ✓ F = Scienze pure
- ✓ G = Tecnologia (Scienze applicate)
- ✓ H = Arti, belle arti e arti decorative
- ✓ I = Letteratura
- ✓ J = Geografia, storia ecc.

Mentre ...

- ✓ X = A. Dumas, I tre moschettieri, Donzelli, Roma 2014

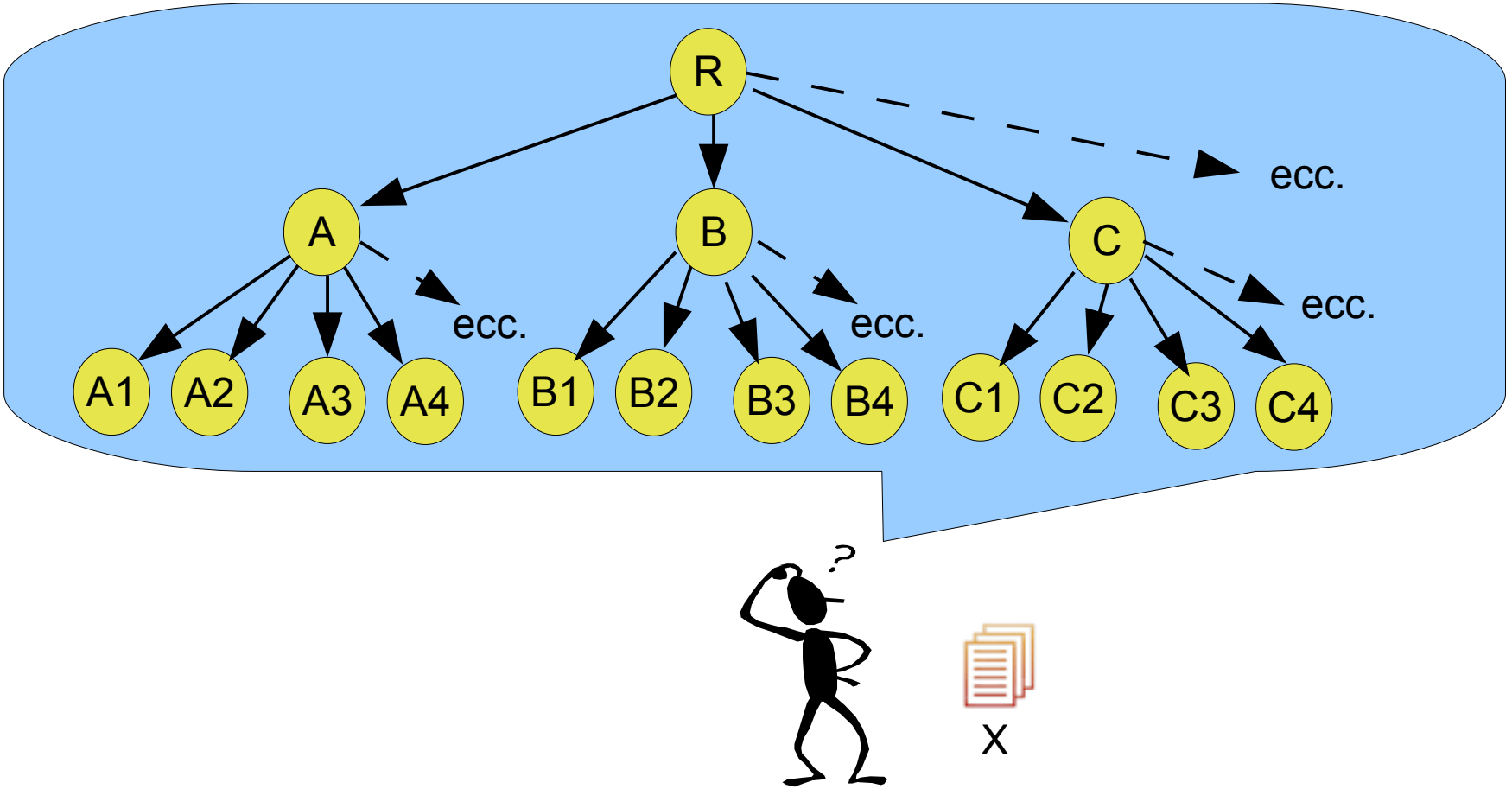
Classificazione Automatica

- **Notiamo però che …**
  - **La numerazione Dewey ha diversi livelli di profondità, quindi in linea di principio si tratta di un problema di classificazione gerarchica, ovvero di un problema in cui le categorie sono parte di una tassonomia**

Vediamo come si può riformulare il problema in questo caso …

Classificazione Automatica

- Dewey in termini gerarchici ...



Classificazione Automatica

- Purtroppo …
  - A volte è molto difficile stabilire l'appartenenza a una categoria senza avere informazioni addizionali (per es. “letteratura francese” se il libro è scritto in italiano)
- Per fortuna …
  - Esistono spesso informazioni addizionali (metadati) che semplificano di molto il problema (esempio di metadati: titolo, autore, nazionalità autore, edizione, casa editrice)
  - Anche questo tipo di dati (oltre ai dizionari semantico-lessicali stile WordNet) possono essere utilizzati per realizzare un arricchimento semantico

Classificazione Automatica

- Il principale problema è comunque dato dal fatto che …
  - … le categorie sono in numero elevato, e questo restringe molto le scelte possibili da adottare a livello tecnologico
- Tra le (poche) opzioni disponibili per risolvere il problema citiamo in particolare …
  - La classificazione gerarchica
  - Le tecniche ECOC (^)

(^) **ECOC** = Error Correcting Output Codes

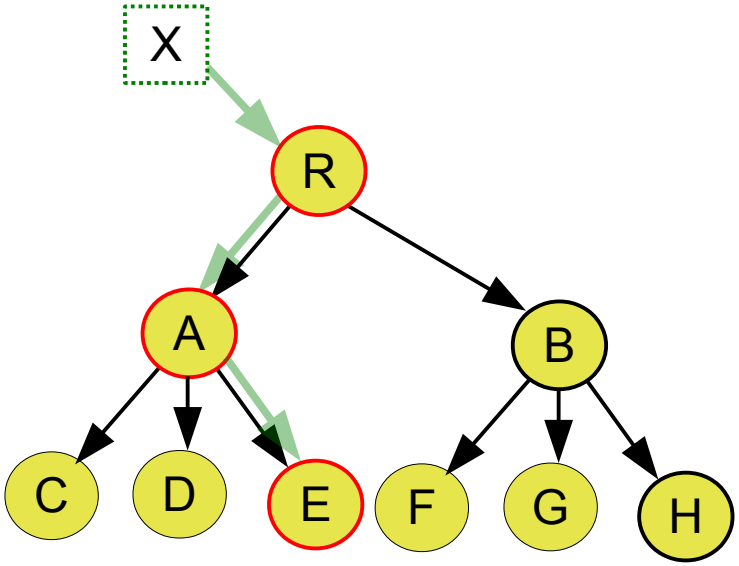
Classificazione Automatica

- **Classificazione gerarchica**
  - **Nella sua formulazione più semplice la classificazione gerarchica avviene lasciando "cadere" l'oggetto da classificare dalla sommità della gerarchia verso il basso ...**
  - **Per ogni nodo della gerarchia dovrà essere stato addestrato un corrispondente classificatore binario (dicotomico), cioè in grado di rispondere soltanto con "si" / "no"**

NB Non cambia nulla se si assume che la risposta, invece di "si" / "no" sia "vero" / "falso" oppure 1/0 ...

Classificazione Automatica

Classificazione gerarchica



Un esempio di classificatore gerarchico costruito su una tassonomia a due livelli:

- ✓ R = Radice della tassonomia (livello zero)
- ✓ A, B = Nodi di primo livello
- ✓ C, D, E, F, G, H = Nodi di secondo livello

Mentre ...

- ✓ X = Oggetto da classificare

NB A ogni nodo della tassonomia è associato un classificatore binario, che se consultato potrà decidere di accettare o meno X. L'oggetto da classificare viene propagato verso il basso soltanto se è accettato. Nel caso in figura X viene accettato dai nodi R, A, E mentre viene rigettato da C, D e B. Si noti che F, G e H non sono neppure coinvolti, dato che B non ha accettato X. Si noti inoltre che la radice R non ha alcuna funzione di decisione (dice sempre “si”).

Classificazione Automatica

- **Classificazione gerarchica: Criticità (^)**
  - **Non è detto che un elemento sia riconosciuto come appartenente a una categoria (mentre in ogni caso almeno una categoria ci dovrebbe essere)**
  - **Se un oggetto da classificare viene erroneamente rifiutato (ovvero è un falso negativo) non potrà più essere recuperato. Questo può costituire un problema se il rifiuto avviene tipicamente ai livelli "alti" della gerarchia (cioè vicino alla radice)**

(^) Ci sono varie criticità, la cui illustrazione necessiterebbe però di una conoscenza approfondita della materia. Pertanto ci limitiamo a segnalare le più evidenti.

Classificazione Automatica

- **Classificazione ECOC**
  - **Le tecniche ECOC sono particolarmente adatte a trattare problemi di classificazione dove le categorie sono molte o moltissime**
  - **L'idea è quella di decomporre un problema multiclasse in un insieme di problemi dicotomici**

Vediamo (molto in generale) come si fa ...

Classificazione Automatica

Matrice di codifica (coding matrix)

Classificazione ECOC

	d1	d2	d3	d4	d5
A	0	0	1	1	0
B	0	1	0	1	0
C	0	1	0	0	1
D	1	1	0	0	1
E	1	0	0	0	1
F	0	0	1	1	0
G	1	0	1	0	1
H	1	1	0	0	0

NB Le categorie sono riportate sulle righe (A, B, C, ecc.). I classificatori dicotomici sulle colonne (d1, d2, d3, ecc.). Gli zeri e uni presenti in tabella indicano se gli esempi di addestramento siano da considerare positivi (1) o negativi (0) per il classificatore dicotomico in questione.

Classificazione Automatica

Classificazione ECOC Addestramento

	d1
A	0
B	0
C	0
D	1
E	1
F	0
G	1
H	1

Prendiamo per esempio il classificatore dicotomico d1:

- ✓ I suoi esempi positivi sono quelli di addestramento appartenenti alle categorie D, E, G e H (marcati con “1”)
- ✓ I suoi esempi negativi sono quelli di addestramento appartenenti alle categorie A, B, C ed F (marcati con “0”)

NB Una volta identificati gli esempi positivi e quelli negativi il classificatore dicotomico potrà essere addestrato. Ovviamente il processo viene ripetuto per tutti i classificatori, in questo specifico caso d1, d2, d3, d4 e d5.

Classificazione Automatica

Classificazione ECOC: classificazione

classificatore	d1	d2	d3	d4	d5
A	1	0	1	1	0

oracolo	d1	d2	d3	d4	d5
A	0	0	1	1	0

- ✓ Supponiamo che il classificatore ECOC, nel predire la categoria associata a un oggetto X risponda, per la categoria A, come indicato nella prima riga
- ✓ Le risposte di classificatori che si comportano come “oracoli” associate alla categoria A sono invece riportate nella seconda riga
- ✓ In questo caso il punteggio (score) della categoria A sarà elevato, poiché soltanto d1 differisce rispetto all'oracolo
- ✓ Quindi cosa si fa per decidere la categoria?
- ✓ Semplicemente si calcola il punteggio di tutte le righe (cioé di tutte le categorie) e poi si sceglie quella con punteggio migliore ...

Classificazione Automatica

- **Classificazione ECOC: Criticità (^)**
  - **La decisione di quanti classificatori dicotomoci utilizzare (quelli sulle colonne, per intenderci) è spesso non banale**
  - **Il problema principale è comunque quello di definire la matrice di codifica, che deve soddisfare vari vincoli …**

(^) Ci sono varie criticità, la cui illustrazione necessiterebbe però di una conoscenza approfondita della materia. Pertanto ci limitiamo a segnalare le più evidenti.

- Contesto operativo
- Qualche nota sulle tecnologie
- Focus sull'arricchimento semantico
- Focus sulla classificazione automatica
- **Conclusioni**

Conclusioni

- Abbiamo visto da vicino, pur senza entrare in profondità, …
  - Alcuni aspetti importanti legati alle tecniche utilizzate per l'arricchimento semantico e la classificazione, in particolare quando applicate a testi / documenti

Conclusioni

- Ci sono però ancora molti problemi da indagare, … d'altra parte noi universitari (almeno in teoria) siamo qui per questo!
  - Oltre naturalmente a …
    - Tenere corsi universitari
    - Scrivere progetti per chiedere finanziamenti
    - Gestire gruppi di ricerca
    - Avere un qualche ruolo addizionale nella struttura
    - Farci da segretari
    - Et cetera …

… Molte Grazie …