Laurea magistrale in Ingegneria informatica (D.M. 270)

Università degli studi di Roma Tre

Categorizzazione degli open-data italiani sul profilo europeo dei metadati DCAT-AP

Simone Gasperoni

LaTeX
Roma, Italia

Contents

1 Introduzione

2 Contesto

I dati sono aperti se chiunque è libero di accederli, usarli, modificarli, e condividerli. [1]

2.1 La filosofia open-data

Gli aspetti più importanti della filosofia open-data sono:

- Disponibilità e accesso.
- Riutilizzo e ridistribuzione: i dati devono essere forniti a condizioni tali da permetterne il riutilizzo e la ridistribuzione. Ciò comprende la possibilità di combinarli con altre basi di dati.
- Partecipazione universale: tutti devono essere in grado di usare, riutilizzare e ridistribuire i dati.

Il movimento open-data si diffuse per la prima volta a seguito della promulgazione della Direttiva sull'Open government del 2009 negli Stati Uniti d'America: *Fin dove possibile e sottostando alle sole restrizioni valide, le agenzie devono pubblicare le informazioni on line utilizzando un formato aperto (open) che possa cioè essere recuperato, soggetto ad azioni di download, indicizzato e ricercato attraverso le applicazioni di ricerca web più comunemente utilizzate. Per formato open si intende un formato indipendente rispetto alla piattaforma, leggibile dall'elaboratore e reso disponibile al pubblico senza che sia impedito il riuso dell'informazione veicolata.*
In questo contesto è nato il portale Data.gov, creato con l'obiettivo di raccogliere in un unico portale tutte le informazioni rese disponibili dagli enti statunitensi in formato aperto. In Italia il portale che raccoglie i dati aperti della pubblica amministrazione è Dati.gov.it

2.2 Interoperabilità dei dati

Tra le ragioni più importanti della filosofia open data abbiamo l'interoperabilità, vale a dire, la capacità di diversi sistemi e organizzazioni di lavorare insieme. Nel caso specifico degli open data - ovviamente - ci riferiamo alla capacità di combinare una grande quantità di dati da diverse fonti. L'interoperabilità è importante perché permette a componenti diverse di lavorare insieme. L'abilità di rendere ciascun dato un componente e di combinare insieme vari componenti è essenziale per la costruzione di sistemi sofisticati. In assenza di interoperabilità ciò diventa quasi impossibile - come nel mito della Torre di Babele, in cui l'impossibilità di comunicare (e quindi di Inter-operare) dà luogo a un fallimento sistemico della costruzione della torre. Una chiara definizione di "apertura" assicura che sia possibile combinare dataset aperti provenienti da fonti diverse in modo adeguato, evitando così la "Torre di Babele".

[1] http://opendefinition.org/od/2.1/en

Il punto cruciale di un bacino di dati (o linee di codice) accessibili e utilizzabili in modo condiviso è il fatto che potenzialmente possono essere liberamente "mescolati" con dati provenienti da fonti anch'esse aperte. L'interoperabilità è la chiave per realizzare il principale vantaggio pratico dell'apertura: aumenta in modo esponenziale la possibilità di combinare diverse basi di dati, e quindi sviluppare nuovi e migliori prodotti e servizi. [2]
Le amministrazioni che aprono i propri dati utilizzano un vocabolario RDF (DCAT) progettato per garantire l'interoperabilità tra cataloghi di dati pubblicati sul web. DCAT incentiva la pubblicazione decentrata di cataloghi e facilita la ricerca federata di dati tra i siti. Usando lo standard DCAT chi pubblica insiemi di dati incrementa la propria visibilità e favorisce le applicazioni che lavorano su dati e metadati in cataloghi multipli (applicazioni cross-portali). Ad esempio il Ministero delle infrastrutture pubblica il proprio catalogo dati (in DCAT) all'indirizzo `http://dati.mit.gov.it/dcat/catalog.rdf`, i cataloghi descritti in conformità a questo vocabolario hanno la seguente formattazione:

```
<rdf:RDF>
   <dcat:Catalog rdf:about="http://dati.mit.gov.it/dcat/catalog">
   <dcat:dataset>
      ...
   </dcat:dataset>
   <dcat:dataset>
      ...
   </dcat:dataset>
   ...
</rdf:RDF>
```

Nei campi dataset ci sono i metadati per la descrizione delle risorse puntate. Infatti ogni dataset comprende i link ai dati (risorse), e dei campi standard per la descrizione di questi dati (questi campi standard sono definiti proprio dal vocabolario DCAT). Ad esempio nell'ambito dei trasporti e delle infrastrutture ci potrebbero essere delle risorse riguardanti la gestione dei porti in Italia (file csv), queste risorse saranno arricchite con dei metadati come ad esempio: license per uso dei dati, frequenza di aggiornamento dei dati, descrizione dei dati, lingua di riferimento, informazioni spaziali, e così via. La lingua e la categoria di appartenenza del dataset sono descritti nel seguente modo:

```
<dcat:theme
   rdf:resource="http://publications.europa.eu/resource/
   authority/data-theme/TRAN"/>
<dct:language
   rdf:resource="http://publications.europa.eu/resource/
   authority/language/ITA"/>
```

Pubblicare i dati sotto forma di cataloghi RDF DCAT garantisce l'interoperabilità, perché i dati sono descritti e pubblicati allo stesso modo indipendentemente dal portale di appartenenza.

[2] http://opendatahandbook.org/guide/it/what-is-open-data

2.3 DCAT-AP e DCAT-AP-it

Il DCAT Application Profile (DCAT-AP) fornisce una specifica comune per la descrizione di dati della pubblica amministrazione in Europa, questa specifica è basata sullo standard DCAT: Data Catalog Vocabulary. [3]
DCAT-AP-it è la versione italiana di DCAT-AP, il profilo nazionale dei metadati utili per descrivere i dati delle pubbliche amministrazioni, conforme alla specifica di DCAT-AP serve a favorire l'interoperabilità semantica di dati e servizi. DCAT-AP-it è un modello dati pensato per rendere omogenei in tutta la pubblica amministrazione italiana i processi di accesso e scambio delle informazioni tra le istituzioni stesse e tra le istituzioni e i cittadini e le imprese, in coerenza con il relativo framework europeo. L'utilizzo di un modello dati valorizza il patrimonio informativo pubblico nazionale in linea con la Direttiva relativa al riutilizzo dell'informazione del settore pubblico. Le categorie definite nel profilo europeo dei metadati DCAT-AP (categorizzazione adottata dal portale europeo European Data Portal) - sono le seguenti:

1. Agriculture, Fisheries, Forestry & Foods
2. Energy
3. Regions & Cities
4. Transport
5. Economy & Finance
6. International Issues
7. Government & Public Sector
8. Justice, Legal System & Public Safety
9. Education, Culture & Sport
10. Environment
11. Health
12. Population & Society
13. Science & Technology

[3] https://www.w3.org/TR/vocab-dcat

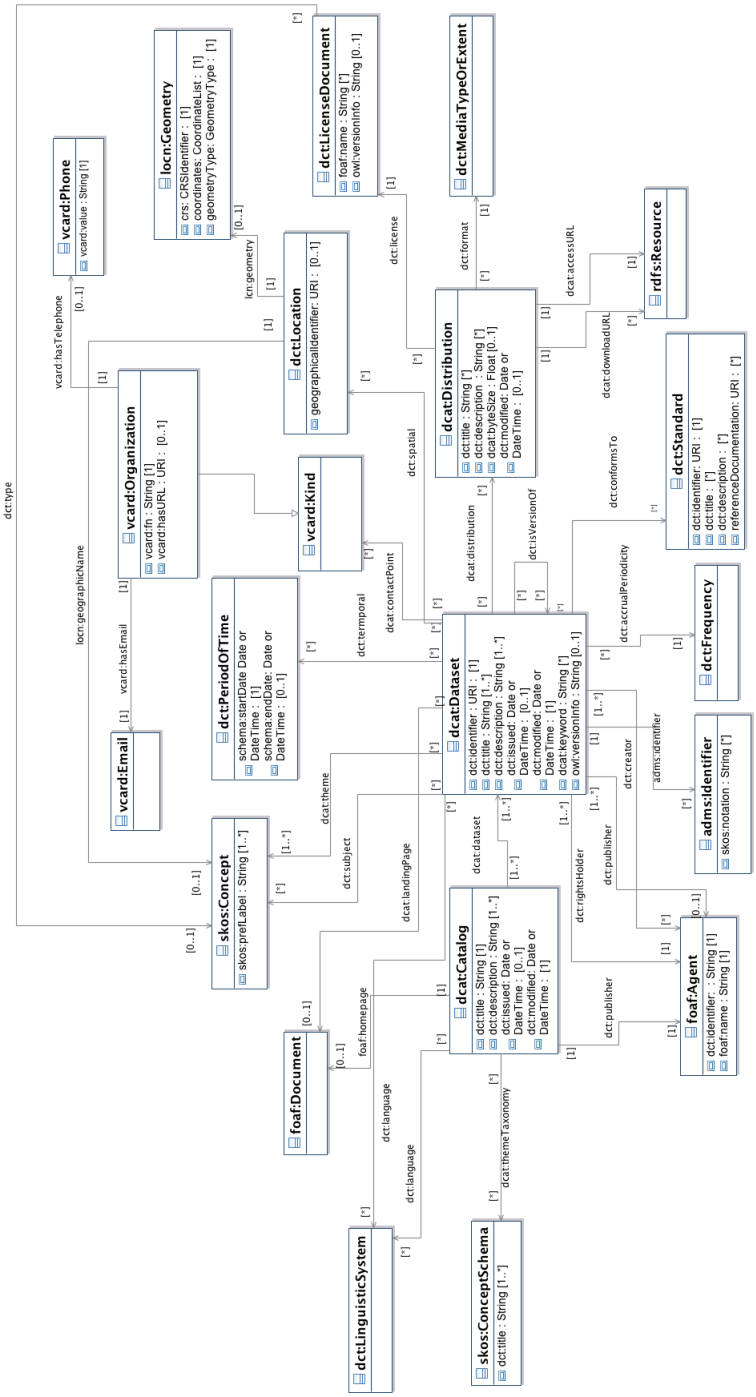

Figure 1: Modello dati DCAT-AP-it

2.4 CKAN

CKAN (Comprehensive Knowledge Archive Network) è una sistema open source (`https://github.com/ckan/ckan`) pensato per l'immagazzinamento, la catalogazione e la distribuzione dei dati in diversi formati; quali ad esempio fogli di calcolo, csv, pdf, xml, rfd, ... CKAN è scritto in Python ed è ispirato dal sistema di gestione dei pacchetti comune a sistemi operativi open source come quelli della famiglia Linux. CKAN è un sistema performante e manutenuto da una comunità di sviluppatori online (Open Knowledge Foundation).
Il sistema è usato sia come piattaforma pubblica su Datahub (`https://datahub.io`), sia da varie pubbliche amministrazioni nell'ambito della loro strategia di pubblicazione di dati aperti, ogni portale web basato su CKAN espone delle Application programming interface API mediante le quali è possibile interrogare i dataset e fruire dei dati di ciascun portale.
Molte pubbliche amministrazioni europee utilizzano CKAN (standard de facto) per la pubblicazione, gestione, e presentazione dei propri dati aperti. Riporto i link ad alcune pubbliche amministrazioni che utilizzano CKAN:

- `https://data.gov.uk`
- `http://dati.emilia-romagna.it`
- `http://dati.trentino.it`
- `http://dati.comune.lecce.it`
- ...

Oltre alle amministrazioni esiste un portale europeo (anch'esso basato su CKAN):

`https://europeandataportal.eu`

Il portale europeo raccoglie i metadati delle informazioni del settore pubblico disponibili sui portali di dati pubblici dei vari paesi europei.
Come detto in precedenza il profilo DCAT-AP garantisce l'interoperabilità, CKAN salvaguarda questa importante caratteristica dei dati aperti mettendo a disposizione un'estensione (plugin):

`https://github.com/ckan/ckanext-dcat`

Il plugin è pensato per abilitare l'esportazione dei metadati in formato DCAT. La corrispondenza tra i campi che descrivono i dati in CKAN ed i campi in DCAT è realizzata mediante un mapping:

`https://github.com/ckan/ckanext-dcat#rdf-dcat-to-ckan-dataset-mapping`

il quale prende i valori dai campi CKAN e li mappa in quelli DCAT.

2.4.1 European Data Portal

European Data Portal (`https://europeandataportal.eu`) è il portale europeo degli open-data dataset di tutta europa. Il portale europeo oltre a raccogliere i metadati europei produce delle analisi riguardanti la fornitura di dati e i vantaggi del loro riutilizzo. In tutta Europa, i dati sono sempre più aperti e disponibile per il riutilizzo. Un certo numero di studi ha incrementato le aspettative per quanto riguarda i benefici finanziari dei dati aperti, la strada da percorrere per sfruttare pienamente i dati aperti è ancora lunga. Il portale europeo esplora approfonditamente il tema del riutilizzo dei dati aperti, al fine di sostenere la trasformazione dei dati in valore (economico, sociale e politico).

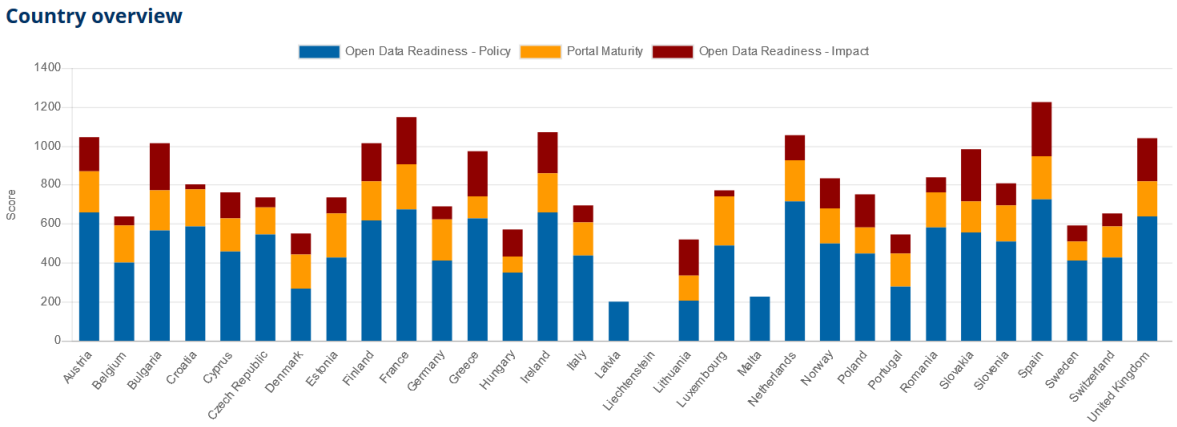

Figure 2: Alcune analisi prodotte dal portale europeo

2.4.2 Open Data Hub Italy

Open Data Hub Italy (`http://sciamlab.com/opendatahub/it`) è una piattaforma per l'indicizzazione e la ricerca di dati aperti (open-data), è il più grande portale italiano di raccordo dei dati aperti delle pubbliche amministrazioni italiane.
Questo portale - anch'esso basato su CKAN - raccoglie più di 34'000 dataset, l'ammontare dei dati e delle organizzazioni tracciate cresce continuamente in quanto il ciclo di aggiornamento del catalogo è contenuto grazie all'utilizzo di algoritmi di processamento parallelo e distribuito.
La piattaforma utilizza algoritmi basati su esecuzione parallela per l'analisi dei testi, l'arricchimento dei metadati di categoria - è realizzato automaticamente mediante il classificatore JRC-JEX (categorizzazione con le voci del tesauro EuroVoc).
Nella fase di harvesting, quando i dati non sono disponibili in maniera esplicita (mediante API), si utilizzano tecniche di "scraping" per reperire le informazioni direttamente dal codice HTML.

2.4.3 Dati.gov.it

Il portale Dati.gov.it (`http://dati.gov.it`) è il portale ufficiale dei dati aperti italiani, raccoglie circa 10'000 dataset open-data, è il portale italiano che fornisce i dati italiani al portale europeo. Il portale italiano è gestito dall'Agenzia per l'Italia digitale, la quale fornisce anche le linee guida per la modellazione del profilo italiano dei metadati DCAT-AP-it (il quale è un'estensione del profilo europeo DCAT-AP).

2.5 EuroVoc

EuroVoc è un tesauro multilingue e pluridisciplinare che comprende la terminologia dei settori d'attività dell'Unione europea. Contiene termini in 23 lingue dell'UE. EuroVoc è in linea con le raccomandazioni del W3C e con gli ultimi sviluppi negli standard di classificazione. Il thesaurus EuroVoc viene utilizzato dalle istituzioni dell'Unione europea, dall'Ufficio delle pubblicazioni dell'UE, da parlamenti nazionali e regionali in Europa, come pure da amministrazioni nazionali e utenti privati di tutto il mondo. [4] I descrittori del thesaurus multilingua EuroVoc sono usati da molti parlamentari europei e centri di documentazione per indicizzare manualmente le loro collezioni di documenti, i descrittori assegnati ai documenti sono utilizzati per cercare ed individuare documenti in una gerarchia semantica divisa in 21 domini, 127 sottodomini (microtesauri) e circa 7000 descrittori (concetti).

2.6 JRC-Acquis

JRC-Acquis è un corpus parallelo multilingua che contiene documenti di natura governativa (riguardanti varie tematiche) utilizzati/prodotti dalle istituzioni europee. La versione corrente di JRC-Acquis (3.0) contiene più di 23'000 documenti annotati manualmente con i descrittori dei concetti del tesauro EuroVoc.
`https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis`
JRC-Acquis può essere usato per addestrare e/o testare sistemi di classificazione basati su algoritmi di text mining o software per la keyword-assignment basati su varie tecnologie semantiche. Il corpus è codificato in XML, in accordo con lo standard TEI per la descrizione dei campi. [5]
I vantaggi che possiamo trarre utilizzando JRC-Acquis come corpus di addestramento per la catalogazione di open data sono molteplici, tra i vantaggi principali abbiamo:

1. Standard TEI

[4] http://eurovoc.europa.eu

[5] Steinberger Ralf, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski & Signe Gilbro. An overview of the European Union's highly multilingual parallel corpora. Language Resources and Evaluation Journal, 2014

2. Corpus aggiornato e manutenuto
3. Corpus multilingua
4. Bridge dinamico tra i concetti e le categorie DCAT-AP

2.6.1 Standard TEI

JRC-Acquis ha un formato che rispetta lo standard TEI. Text Encoding Initiative (TEI) è un consorzio che sviluppa e mantiene uno standard per la rappresentazione di testi in formato digitale. La codifica TEI garantisce che il corpus sia machine-readable, le modalità di interazione dei sistemi che utilizzano JRC-Acquis (ad esempio un modulo XQuery) non variano al variare delle versioni di JRC-Acquis.

Il titolo del documento è contenuto nel tag head:

```
<head n="1">
   Decisione n. 2046/2002/CE del Parlamento europeo e del Consiglio,
   del 21 ottobre 2002, che modifica la decisione n. 1719/1999/CE
   relativa ad una serie di orientamenti, compresa l'individuazione
   di progetti di interesse comune, per reti transeuropee per lo
   scambio elettronico di dati fra amministrazioni (IDA)
</head>
```

Il corpo testuale del documento è contenuto nel tag div di tipo body

```
<div type="body">
   ...
</div>
```

Segue la parte relativa alla classificazione EUROVOC:

```
<profileDesc>
   <textClass>
      <classCode scheme="eurovoc">206</classCode>
      <classCode scheme="eurovoc">3010</classCode>
      <classCode scheme="eurovoc">453</classCode>
      <classCode scheme="eurovoc">616</classCode>
      <classCode scheme="eurovoc">4424</classCode>
      <classCode scheme="eurovoc">5864</classCode>
   </textClass>
</profileDesc>
```

Ogni documento contiene anche metadati che potrebbero essere utili come la data di emissione dei documenti:

```
<date>
   2007-03-29
</date>
```

e la sorgente:

```
<div type="signature">
   <p n="77">Fatto a Lussemburgo, addì 21 ottobre 2002.</p>
   <p n="78">Per il Parlamento europeo</p>
   <p n="79">Il Presidente</p>
   <p n="80">P. Cox</p>
   <p n="81">Per il Consiglio</p>
   <p n="82">Il Presidente</p>
   <p n="83">P. S. Møller</p>
   <p n="84">(1) GU C 332 E del 27.11.2001, pag. 287.</p>
   <p n="85">(2) GU C 80 del 3.4.2002, pag. 21.</p>
   ...
</div>
```

2.6.2 Corpus aggiornato e manutenuto

JRC-Acquis è un corpus manutenuto e aggiornato manualmente, decine di migliaia di documenti riguardanti politiche europee su tematiche trasversali sono arricchiti da descrittori EuroVoc da specialisti del dominio. JRC-Acquis viene manutenuto perché utilizzato da diversi sistemi basati su tecnologie semantiche, nell'ambito della classificazione è utilizzato come corpo di addestramento del classificatore JRC-JEX, sistema ufficiale di classificazione dei documenti utilizzato dal parlamento europeo.

2.6.3 Corpus parallelo multilingua

JRC-Acquis è un corpus multilingua, questo vuol dire che è disponibile una versione per ciascuna lingua dell'Unione Europea. Il parallelismo è dato dal fatto che molti documenti sono disponibili in più lingue. Utilizzare corpus multilingua facilita il compito di sviluppare sistemi basati su tecnologie semantiche multilingua.

2.6.4 Bridge dinamico tra EuroVoc e le categorie DCAT-AP

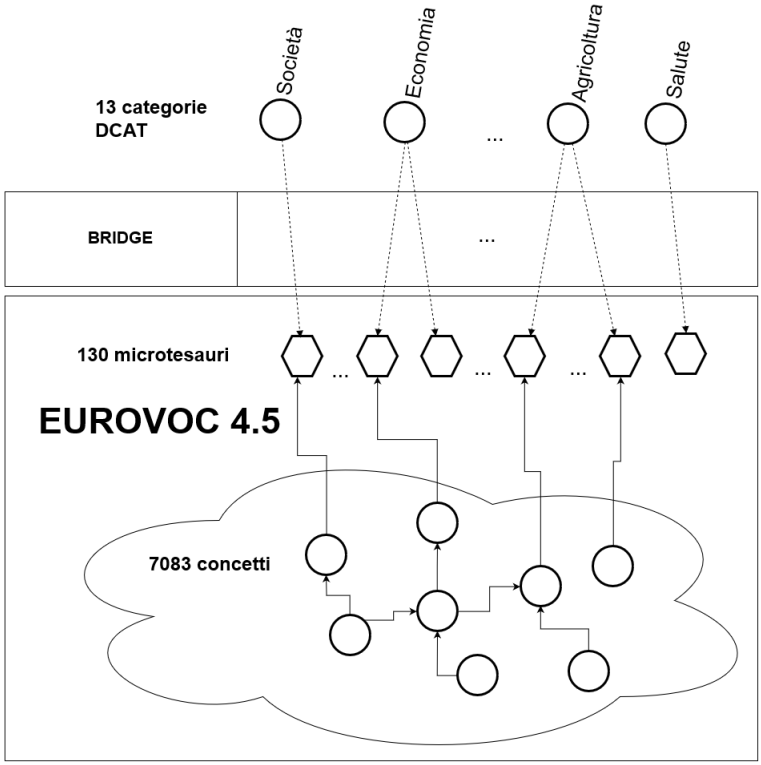

Figure 3: Bridge tra le categorie DCAT-AP ed EUROVOC

In appendice ("Associazione dei Temi DCAT-Microtesauri EuroVoc") è possibile prendere nota delle corrispondenze tra i 13 temi definiti nell'ambito di DCAT-AP e i concetti (microtesauri) disponibili nel vocabolario EuroVoc (come indicato dal portale europeo https://www.europeandataportal.eu).

JRC-Acquis è indicizzato con i concetti del vocabolario EuroVoc, pertanto è stato necessario sviluppare un modulo specializzato per la navigazione del tesauro al fine di ottenere i descrittori dei microtsauri associati ai descrittori dei concetti, per poi permutarli a loro volta nei rispettivi temi DCAT-AP.
Il seguente modulo java: `https://github.com/simonegasperoni/cata/blob/master/code/src/main/java/com/sciamlab/it/eurovoc/EVoc.java`
è dedicato alla gestione di EuroVoc, fornisce funzionalità di base per interrogare

[5]Profilo italiano di DCAT-AP (DCAT Application Profile for data portals in Europe), Versione 1.0, Agenzia per l'Italia digitale, 2006

il tesauro:

1. Dato un concetto si ottengono i microtesauri associati
2. Dato un microtesauro si ottiene il relativo tema DCAT-AP (EU Data Theme)

Un corpus etichettato con le voci di un tesauro come EuroVoc permette la realizzazione di associazioni tra concetti a vario livello di astrazione in modo dinamico, ad esempio qualora il microtesauro EuroVoc "2831 culture and religion" originariamente associato al tema Population & Society (Popolazione e Società) dovesse passare in Education, culture and sport (Educazione, cultura e sport), basterebbe modificare il mapping tra i microtesauri e i temi all'interno del modulo specializzato.

```
Map<String, EUNamedAuthorityDataTheme.Theme> EUROVOC_TO_DCAT_CATEGORIES
    = new HashMap<String, EUNamedAuthorityDataTheme.Theme>(){
     {
        put("0406",EUNamedAuthorityDataTheme.Theme.GOVE);
        ...
        ...
        ...
        put("7626",EUNamedAuthorityDataTheme.Theme.INTR);
     }
};
```

Il tesauro (EuroVoc) e il corpus (JRC-Acquis) sono i due componenti fondamentali per la costruzione dei dati di addestramento del classificatore, essendo questi due componenti disaccoppiati è possibile pensare di aggiornare uno e mantenere immutato l'altro qualora siano disponibili versioni più aggiornate del corpus o del tesauro. Anche questa è una forma di flessibilità, anche se in realtà dall'aggiornamento del tesauro non si traggono benefici se i descrittori dei concetti aggiunti non compaiono anche nel corpus.

3 Elaborazione del corpus

Una fase preliminare in qualsiasi ambito di classificazione di documenti è l'elaborazione dei corpi testuali, si tratta di scegliere le modalità con le quali estrarre le "feature" dai testi - in base alla loro formattazione, al fine di rendere le "feature" fruibili al sistema di classificazione in modo opportuno.

Nel contesto di un sistema di categorizzazione dei testi vogliamo ottenere delle "feature" ben rappresentative del contenuto semantico dei documenti dai quali sono estratte. Le tappe che dobbiamo percorrere sono essenzialmente due: innanzitutto è necessario utilizzare degli algoritmi per l'estrazione di "feature", dopodiché queste "feature" potranno essere stemmerizzate o rielaborate in vario modo, ad esempio utilizzando n-grammi di parole.

Il processo di elaborazione viene suddiviso in fasi diverse, tuttavia simili a quelle che si possono incontrare nel processo di elaborazione di un linguaggio di programmazione, i seguenti passi rappresentano un possibile canovaccio da seguire per l'elaborazione dei testi:

- **Analisi lessicale:** scomposizione di un'espressione linguistica in token (in questo caso le parole).
- **Analisi grammaticale:** associazione delle parti del discorso a ciascuna parola nel testo.
- **Analisi sintattica:** arrangiamento dei token in una struttura sintattica (ad albero: parse tree).
- **Analisi semantica:** assegnazione di un significato (semantica) alla struttura sintattica e, di conseguenza, all'espressione linguistica.

3.1 Lemmatizzazione

Una tecnica di elaborazione delle "feature" è la lemmatizzazione, vale a dire, "quel complesso di operazioni che conducono a riunire tutte le forme sotto il rispettivo lemma", intendendo per lemma "ciascuna parola-titolo o parola-chiave di un dizionario" e per forma ogni possibile diversa realizzazione grafica di un lemma.[6]

Esistono delle convenzioni di lemmatizzazione proprie di ciascuna lingua: ad esempio in italiano è uso convenzionale che il lemma verbale sia la forma coniugata all'infinito presente attivo.
La lemmatizzazione è, dunque, una pratica apparentemente facile, se non, addirittura, ovvia ed intuitiva: più, o meno tutti, infatti, esercitiamo quotidianamente la conoscenza della differenza tra lemma e forma. Tuttavia, alla prova dei fatti, la lemmatizzazione rivela una serie di problemi, il più delle volte non

[6] R. Busa, Fondamenti di informatica linguistica. Istituto Geografico De Agostini, 1987

immaginabili prima di averne fatto esperienza diretta: queste difficoltà rendono il lemmatizzare un esercizio interessante, in quanto costringe chi lo esercita a riflettere su:

- Quanti e quali automatismi l'uomo metta inconsciamente in atto ogni volta che parla, o scrive.
- Quanto sia difficile formalizzare questi automatismi.

Esistono due tipi di lemmatizzazione: la lemmatizzazione morfologica analizza le forme di parole in isolamento, ovvero fuori dal contesto sintattico, fornendone tutti i valori che sono possibili in un dato sistema linguistico.

La lemmatizzazione morfo-sintattica analizza le forme di parole entro il contesto sintattico. Non è mai ambigua, ma sempre univoca, in quanto l'immersione della forma nella frase ne precisa il valore. Quindi, mentre la lemmatizzazione morfologica è indipendente dal testo, la lemmatizzazione sintattica è, invece, legata al contesto.

Questo processo è reso particolarmente difficile e complesso a causa delle caratteristiche intrinseche di ambiguità del linguaggio umano. Il problema della lemmatizzazione richiama il problema della disambiguazione, infatti talvolta nell'assegnazione dei lemmi ai termini è necessario contestualizzare il termine.

3.2 Disambiguazione

Nell'analisi semantica la procedura automatica che attribuisce all'espressione linguistica un significato tra i diversi possibili è detta disambiguazione. La comprensione del linguaggio naturale è spesso considerata un problema IA-completo, poiché si pensa che il riconoscimento del linguaggio richieda una conoscenza estesa del mondo e una grande capacità di manipolarlo. Per questa ragione, la definizione di "comprensione" è uno dei maggiori problemi dell'elaborazione del linguaggio naturale.[7]

Molte parole nel linguaggio naturale sono delle polisemie, vale a dire, possono avere più significati. Le tecniche finalizzate alla disambiguazione sono conosciute in letteratura come *Word Sense Disambiguation* (WSD).

La WSD può essere affrontata come un problema di classificazione, il senso corretto è la classe da predire la parola è rappresentata con un insieme (vettore) di feature in ingresso al classificatore, l'ingresso è di solito costituito da una rappresentazione della parola da disambiguare (target) e del contesto in cui si trova (un certo numero di parole a sinistra e destra della parola target). Il classificatore può essere stimato con tecniche di apprendimento automatico a partire da un dataset etichettato. Si possono usare diversi modelli per costruire

[7] I. Chiari, Introduzione alla linguistica computazionale. Laterza, 2007

il classificatore (Naïve Bayes, reti neurali, alberi di decisione...).

Il problema dell'approccio appena descritto è che il modello di apprendimento supervisionato che scegliamo di adottare potrebbe richiedere un training set troppo grande per essere sufficientemente preciso, per questo motivo esistono tecniche alternative come i metodi Dictionary-based.

Tra i metodi Dictionary-based abbiamo l'algoritmo di Lesk (1986) che è un metodo molto semplice, basato sull'intuizione secondo cui un dizionario può fornire informazioni sul contesto legato ai sensi delle parole (le glosse).

3.3 Snowballstem

Lo stemming è il processo usato per ridurre parole flesse al loro tema, il tema non deve necessariamente coincidere con la radice morfologica della parola: l'importante è che parole con una semantica strettamente correlata vengano mappate sullo stesso tema. Le tecniche di stemming sono studiate in informatica da quarant'anni, sono uno dei metodi di base per ridurre la dimensionalità dei documenti di testo.

In questo campo è stato fondamentale il contributo di Martin Porter, inventore dello "stemmer di Porter", uno dei più comuni algoritmi per la stemmerizzazione in lingua inglese, e inventore del framework Snowball:
`http://snowballstem.org`

Porter definì il suo algoritmo di stemming in un articolo del 1980 "An algorithm for suffix stripping", il quale è stato citato più di 8'000 volte secondo Google Scholar.

3.4 Snowball

Since it effectively provides a "suffix STRIPPER GRAMmar", I had toyed with the idea of calling it "strippergram", but good sense has prevailed, and so it is "Snowball" named as a tribute to SNOBOL, the excellent string handling language of Messrs Farber, Griswold, Poage and Polonsky from the 1960s.
Martin Porter

Il progetto Snowball è un progetto open source:
`https://github.com/snowballstem/snowball`
Il progetto Snowball raccoglie le implementazioni di vari algoritmi di stemming (per più lingue e in più linguaggi di programmazione). Snowball è utilizzato in diversi progetti Apache Software Foundation, tra i quali abbiamo Apache Solr e Apache Lucene.
Le stopword della lingua italiana sono riportate in formato txt al seguente indirizzo:
`http://snowballstem.org/algorithms/italian/stop.txt`

Tartarus Snowball contiene anche un adattamento dell'algoritmo di Porter per la lingua italiana, riporto di seguito l'algoritmo di stemming per la lingua italiana:

La lingua italiana prevede le seguenti forme accentate:

á, é, í, ó, ú, à, è, ì, ò, ù

Gli accenti acuti vanno rimpiazzati con gli accenti gravi. Le vocali sono:

a, e, i, o, u, à, è, ì, ò, ù

Nell'algoritmo di stemming della lingua italiana - ma non solo - si utilizzano le due seguenti definizioni di "regione":

- **R1**: è la regione che segue la prima non vocale seguita da una vocale, potrebbe essere nulla qualora non ci fosse una non vocale.
- **R2**: è la regione che segue la prima non vocale seguita da una vocale nella regione R1, anche questa potrebbe essere una regione nulla.
- **RV**: Se la seconda lettera è una consonante, RV è la regione che segue la prossima vocale, oppure se le prime due lettere sono vocali, RV è la regione che segue la prossima consonante; altrimenti (nel caso consonante-vocale) RV è la regione dopo la terza lettera.

Riporto di seguito i passi dell'algoritmo:

1. Ricerca del più lungo suffisso tra i seguenti:
   ci, gli, la, le, li, lo, mi, ne, si, ti, vi, sene, gliela, gliele, glieli, glielo, gliene, mela, mele, meli, melo, mene, tela, tele, teli, telo, tene, cela, cele, celi, celo, cene, vela, vele, veli, velo, vene

   a seguito di questi altri suffissi:

   (a) *andо, endo*
   (b) *ar, er, ir*

   In RV, nel caso di (a) il suffisso è cancellato, nel caso di (b) è rimpiazzato dalla "e":

   Ad esempio abbiamo:
   guardandogli → guardando
   accomodarci → accomodare

2. Ricerca del più lungo suffisso tra i seguenti suffissi:
   Cancellare se in R2:
   anza, anze, ico, ici, ica, ice, iche, ichi, ismo, ismi, abile, abili, ibile, ibili, ista, iste, isti, istà, istè, istì, oso, osi, osa, ose, mente, atrice, atrici, ante, anti

   Cancellare se in R2:
   azione, azioni, atore, atori

   Sostituire con *log*, se in R2:
   logia, logie

   Sostituire con *u*, se in R2:
   uzione, uzioni, usione, usioni

   Sostituire con *ente*, se in R2:
   enza, enze

   Cancellare se in RV:
   amento, amenti, imento, imenti

   Cancellare *amente* se in R1
   se preceduto da *iv*, cancellare se in R2 (anche se preceduto da *at*, cancellato se in R2), altrimenti, se preceduto da *os, ic, abil* cancellare se in R2

   Cancellare *ità* se preceduto da *abil, ic, iv* in R2

   Cancellare *ivo, ivi, iva, ive* in R2

   Si passa al passo tre se non abbiamo più nulla da rimuovere.

3. Per quanto riguarda i suffissi dei verbi si va alla ricerca del più lungo suffisso in RV, e, se trovato, si cancella:

   ammo, ando, ano, are, arono, asse, assero, assi, assimo, ata, ate, ati, ato, ava, avamo, avano, avate, avi, avo, emmo, enda, ende, endi, endo, erà, erai, eranno, ere, erebbe, erebbero, erei, eremmo, eremo, ereste, eresti, erete, erò, erono, essero, ete, eva, evamo, evano, evate, evi, evo, Yamo, iamo, immo, irà, irai, iranno, ire, irebbe, irebbero, irei, iremmo, iremo, ireste, iresti, irete, irò, irono, isca, iscano, isce, isci, isco, iscono, issero, ita, ite, iti, ito, iva, ivamo, ivano, ivate, ivi, ivo, ono, uta, ute, uti, uto, ar, ir

4. Cancellare *a, e, i, o, à, è, ì* oppure *ò* se in RV

crocchi → crocch
crocchio → crocch

Sostituire infine *ch, gh* con *c, g* se in RV

crocch → crocc

```
public List<String> execute(String text){
    String resultString = text.replaceAll("\\P{L}+", " ").toLowerCase();
    List<String> result=new ArrayList<String>();
    StringTokenizer tk=new StringTokenizer(resultString);
    while(tk.hasMoreTokens()){
        String s=tk.nextToken();
        if(!stopwords.contains(s)){
            result.add(this.stems(s));
        }
    }
    return result;
}
```

4 Classificazione bayesiana

I classificatori bayesani sono metodi statistici di classificazione, predicono la probabilità che una data istanza appartenga ad una certa classe. Nel gergo della classificazione di testi o Text Categorization, con il termine classificatore bayesiano ci si riferisce convenzionalmente al classificatore bayesiano naïve (Naïve Bayes Classifier), ossia un classificatore bayesiano semplificato con un modello di probabilità sottostante che fa l'ipotesi di indipendenza delle feature, ovvero assume che la presenza o l'assenza di una particolare feature in un documento testuale non è correlata alla presenza o assenza di altre feature.
I classificatori bayesiani sono metodi incrementali: ogni istanza dell'insieme di addestramento modifica in maniera incrementale la probabilità che una ipotesi sia corretta. La conoscenza già acquisita può essere combinata facilmente con le nuove osservazioni basta aggiornare i conteggi. Questi metodi sono utilizzati ad esmpio in Mozilla o SpamAssassin per riconoscere le mail spam dalle mail ham.
Una probabilità è una misura su un insieme di eventi che soddisfa tre assiomi:

$$0 \leq P(E = e_i) \leq 1$$

$$\sum_{n=1}^{n} P(E = e_i) = 1$$

$$P(E = e_1 \cup E = e_2) = P(E = e_1) + P(E = e_2)^8$$

Dove gli eventi e1 ed e2 sono disgiunti. Un modello probabilistico consiste in uno spazio di possibili esiti mutualmente esclusivi insieme alla misura di probabilità associata ad ogni esito. Che tempo fa domani? esiti: {SOLE, NUVOLE, PIOGGIA, NEVE}, l'evento corrispondente ad una precipitazione è il sottoinsieme {PIOGGIA, NEVE}.

Definizione 1. La probabilità condizionale è definita come:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Definizione 2. A e B sono condizionalmente indipendenti se

$$P(B|A) = P(B)$$

o il suo equivalente

$$P(A|B) = P(A)$$

4.1 Formula di Bayes

Il teorema di Bayes (formula di Bayes o teorema della probabilità delle cause), proposto da Thomas Bayes, deriva da due teoremi fondamentali delle probabilità: il teorema della probabilità composta e il teorema della probabilità assoluta. Viene impiegato per calcolare la probabilità di una causa che ha scatenato l'evento verificato.

Teorema 1. *gli eventi A (per ogni i) sono stocasticamente indipendenti e sono spesso chiamati cause di E, vale a dire:*

$$E \subset \bigcup_{j=1}^{n} A_j$$

abbiamo

$$P(A_i|E) = \frac{P(E|A_i)P(A_i)}{P(E)} = \frac{P(E|A_i)P(A_i)}{\sum_{j=1}^{n} P(E|A_j)P(A_j)}$$

Segue la dimostrazione.[9] *Dalla teoria delle probabilità condizionali abbiamo:*

$$(1)\ \ P(A_i|E) = \frac{P(A_i \cap E)}{P(E)}$$

$$(2)\ \ P(E|A_i) = \frac{P(A_i \cap E)}{P(A_i)} \Rightarrow P(A_i \cap E) = P(E|A_i)P(A_i)$$

andiamo a sostituire la (2) a numeratore della (1) ottenendo:

$$P(A_i|E) = \frac{P(E|A_i)P(A_i)}{P(E)}$$

dato che

$$E \subset \bigcup_{j=1}^{n} A_j$$

abbiamo che

$$P(E) = \sum_{j=1}^{n} P(E|A_j)P(A_j)$$

perché

$$E = E \cap (\bigcup_{j=1}^{n} A_j) \Rightarrow E = \bigcup_{j=1}^{n} E \cap A_j$$

infine abbiamo che

$$P(E) = P(\bigcup_{j=1}^{n} E \cap A_j)$$

essendo gli eventi A incompatibili abbiamo

$$P(E) = \sum_{j=1}^{n} P(E \cap A_j) \Rightarrow \sum_{j=1}^{n} P(E|A_j)P(A_j)$$

[9] Introduzione alla probabilità, Enzo Orsingher, Luisa Beghin, Carocci editore

4.2 Classificatori bayesiani naïve

L'assunzione di indipendenza rende i calcoli possibili consente di ottenere classificatori ottimali quando è soddisfatta ma è raramente soddisfatta in pratica. Questa assunzione consentono di considerare le relazioni causali tra gli attributi, in realtà, si è visto che anche quando l'ipotesi di indipendenza non è soddisfatta, il classificatore naïve Bayes spesso fornisce ottimi risultati.

Sia X una istanza da classificare, e C1,...,Cn le possibili classi. I classificatori Bayesiani calcolano

$$P(C_i|X)$$

come

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Andiamo a cercare l'indice i per massimizzare la probabilità condizionata, ottenendo

$$Max_i[P(C_i|X)]$$

P(X) è uguale per tutte le classi per cui non occorre calcolarla, P(Ci) si può calcolare facilmente sull'insieme dei dati di addestramento, si conta la percentuale di istanze di classe Ci sul totale. Assunzione dei classificatori naïve: indipendenza degli attributi. Se X è composta dagli attributi 'a' con indice da 1 a m, otteniamo

$$P(X|C_i) = \prod_{j=1}^{m} P(A_j = a_j|C_i)$$

per il calcolo di

$$P(A_j = a_j|C_i)$$

La formula che verrà utilizzata nella fase di predizione dal classificatore sarà pertanto:

$$v = Max_i[P(C_i) \prod_{j=1}^{m} P(A_j = a_j|C_i)]$$

dove v è la categoria predetta.

Se gli A sono categorici, viene stimato come la frequenza relativa delle istanze che hanno quel determinato valore a con indice j tra tutte le istanze di C. Se A è continuo, si assume che la probabilità segue una distribuzione Gaussiana, con media e varianza stimata a partire dalle istanze di classe C.

```
#pseudo codice classificatore naif
#vettore dei target v[j]
def v[]
#vettore degli attributi a[i]
def a[]
#p(a[i]|v[j])
#probabilita' che occorra a[i] quando il documento etichettato v[j]
#p(v[j]) probab. che occorra v[j]
```

```
def double p(e){
   "ritorna la stima probabilistica per l'evento e"
}

def void naif_bayes_learner(){
   for each j do {
      stima p(v[j])
      for each i do {
         stima p(a[i]|v[j])
      }
   }
}

#entry da classificare: x
def v nuova_classificazione(x){
      "ritorna v[j] tale che
      p(v[j])*p(a[1]|v[j])*..*p(a[i]|v[j])*..*p(a[n]|v[j])
      sia il massimo possibile"
}
```

4.3 Stima delle probabilità (m-estimate)

Quando calcoliamo le probabilità dobbiamo avere alcune accortezze, infatti, se un certo valore di un attributo non si verifica mai per una data classe quando arriva una nuova istanza X la probabilità sarà sempre nulla indipendentemente da quanto siano probabili i valori per gli altri attributi. Il problema delle frequenze nulle non è il solo nella stima delle probabilità in un classificatore bayesiano. Le probabilità tendono ad essere sottostimate in alcune circostanze, ad esempio:

$$P(wind = strong|playTennis = no)$$

stimato come

$$\frac{n_C}{n}$$

Questa stima è buona in molti casi ma se abbiamo pochi esempi con playTennis=no la stima tenderà a zero. Un modo per far fronte a tutte queste problematiche è mediante l'uso di una stima chiamata "m-estimate":

$$\frac{n_C + mp}{n + m}$$

p è la probabilità a priori, solitamente si assume p come il reciproco di k dove k è il numero di valori diversi per attributo

$$\frac{n_C + m\frac{1}{k}}{n + m}$$

m è la *equivalent sample size*, come si può vedere dalla formula serve a determinare il peso di incidenza di p sui dati osservati.

4.4 Classificazione bayesiana di testi

Il classificatore byaesiano sopradescritto trova applicazione nel campo della categorizzazione dei testi essendo - ad oggi - uno dei metodi più efficaci conosciuti. Segue una trattazione sull'algoritmo che sfrutta le intuizioni probabilistiche bayesiane, altri esempi sono descritti da Lewis (1991), Lang (1995), Joachims (1996)[10]

Le istanze X che abbiamo considerato fin'ora possono ora essere considerati documenti testuali. Il training set da considerare è una collezione di documenti etichettati (classificati), su questa base di conoscenza si dovrà costruire un sistema di predizione per le entry dello spazio X.

Prendiamo in considerazione una collezione di testi, ad esempio 1000, dei quali 300 interessano ad una certa persone, mentre invece, gli altri 700 sono considerati non interessanti, questo può essere considerato un dataset per addestrare il nostro classificatore dove le categorie sono "like" e "dislike".

Due problemi fondamentali nella progettazione del classificatore bayesiano sono: la rappresentazione dei documenti, e la modalità di stima delle probabilità.

4.5 Rappresentazione testi e stima delle probabilità

Il modo più semplice di rappresentare i testi è mediante una raccolta - senza considerare l'ordine - di parole. Testi lunghi daranno luogo ad insiemi di attributi molto grandi, come vedremo questo non è un problema. Questo tipo di approccio è personalizzabile introducendo n-grammi di parole o la lemmatizzazione. Ricollegandoci all'esempio di prima abbiamo:

$$v = Max_i[P(C_i) \prod_{j=1}^{m} P(A_j = a_j|C_i)]$$

dove le categorie sono due:

$$C_1 : like, C_2 : dislike$$

avremo dunque

$$P(C_1) = \frac{300}{1000}, \;\; P(C_2) = \frac{700}{1000}$$

Le probabilità condizionate sono semplicemente proporzionali alle frequenze con cui occorre un parola dentro tutti i documenti di una categoria. La stima delle probabilità è una m-estimate nella quale consideriamo

$$m = p = |Vocabolario|$$

$$\frac{w + 1}{n + |Vocabolario|}$$

[10]Machine learning, Mc Graw Hill, 1997, Tom M. Mitchell

con n numero di parole di tutti i documenti di una determinata categoria, w numero di occorrenze di una data parola nell'insieme di parole di una data categoria.

In sintesi l'algoritmo di classificazione dei testi usa un classificatore bayesiano naïve con l'assunzione che la probabilità di occorrenza della parola è indipendente dalla posizione dentro i documenti.

Segue lo psudocodice di un approccio minimale:

```
#pseudo codice classificatore naif per la categorizzazione testi
#vettore dei target v[j]
def v[]

#inizializzo vocabolario con tutti i vocaboli
def vocabolario=init_vocabolario()

def void naif_bayes_TEXT_learner(){
   #per ogni target v[j]
   for each j do {
      docs[j]="insieme dei documenti etichettati con v[j]"
      p(v[j])=|docs[j]|/|Esempi|
      text[j]="concateno tutti i docs[j]"
      n="numero di parole distinte dentro text[j]"

      #qui si calcolano i pesi per le parole
      for each parola in Vocabolario do{
         w="numero di volte che la parola occorre in text[j]"
         p(parola|v[j])=(w+1)/(n+|Vocabolario|)
      }
   }
}

#entry da classificare: x
nuova_classificazione(x)
```

5 Feature selection

La selezione delle feature è il processo che ci porta a selezionare un sottoinsieme di feature (termini nella text classification), solo questo sottoinsieme sarà utilizzato come training set per i nostri classificatori. I motivi principali per cui si procede ad una selezione delle feature sono due: innanzi tutto alcuni modelli possono essere addestrati solo con un insieme di feature contenuto, data la loro complessità computazionale in fase di addestramento o predizione. In secondo luogo dobbiamo considerare che i classificatori tendono ad essere più precisi quando il numero delle feature è ridotto, molte feature, non solo non sono determinanti nell'individuazione della classe di appartenenza di un documento, ma possono addirittura introdurre un rumore (noise feature). Le noise feature sono quelle feature che occorrendo accidentalmente in una sola classe si rendono responsabili di errate generalizzazioni che colpiscono l'accuratezza della classificazione (overfitting). Introduction to Information Retrieval[11]descrive tre procedure per la selezione di feature:

- Mutual information
- Chi square
- Frequency based

Queste tre procedure ci permettono di ottenere una misura di utilità di ciascuna feature.

5.1 Mutual information feature selection

Mutual information di un termine t e una classe c è la misura di quanto la feature contribuisce a determinare la corretta classificazione

5.2 χ^2 feature selection

Nella selezione delle feature χ^2 si sfrutta l'intuizione statistica del test χ^2 che serve a determinare il grado di indipendenza tra eventi, nel nostro caso tra feature e classe di appartenenza.

$$\chi^2(t,c) = \sum_{t\in[0,1]} \sum_{c\in[0,1]} \frac{(N_{c,t} - E_{c,t})^2}{E_{c,t}}$$

Il calcolo delle N riguarda le frequenze osservate: $N_{1,1}$ è il numero di documenti nei quali occorre il termine t nella classe c, $N_{1,0}$ è il numero di documenti nei quali non occorre il termine t nella classe c, $N_{0,1}$ è il numero di documenti nei quali occorre il termine t in tutte le classi eccetto la classe c, $N_{0,0}$ è il numero di documenti nei quali non occorre il termine t in tutte le classi eccetto c.

[11]C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008

$E_{c,t}$ sono le frequenze attese che il termine t e la classe c occorrano nello stesso documento essendo indipendenti:

$$E_{c,t} = N \cdot P(t) \cdot P(c)$$

ad esempio

$$E_{1,1} = N \cdot \frac{N_{1,1} + N_{1,0}}{N} \cdot \frac{N_{1,1} + N_{0,1}}{N}$$

5.3 Frequency-based feature selection

Tra le tre procedure è la più semplice, consiste nell'andare a selezionare solo le feature che occorrono di più nelle varie classi. La frequenza può essere considerata come "Document frequency" (numero di documenti della classe che contengono una determinata feature), o come "Collection frequency" (numero di occorrenze della feature in una classe, senza considerare i documenti). "Document frequency" è più appropriata per il modello di Bernoulli, mentre invece la "Collection frequency" è più appropriata per il modello multinomiale. Sebbene questo metodo sia molto meno complesso rispetto agli altri due, questo approccio introduce una limitazione, nel selezionare le feature potrebbe includere feature trasversali alle classi, queste feaure non danno un contributo informativo circa il legame tra la feature e la classe.

5.4 Benefici della feature selection

In "Introduction to Information Retrieval (Manning, Raghavan, Schütze)" è presente uno studio sui benefici delle tecniche di feature selection. Il grafico che segue evidenzia i benefici prodotti su alcuni corpora REUTERS, in particolare gli algoritmi di feature selection sono applicati in modo combinato con diversi algoritmi (Naïve Bayes) di classificazione:

- Mutual information feature selection - Bayes multinomiale
- Chi square feature selection - Bayes multinomiale
- Frequency based feature selection - Bayes multinomiale
- Mutual information feature selection - Bayes binomiale

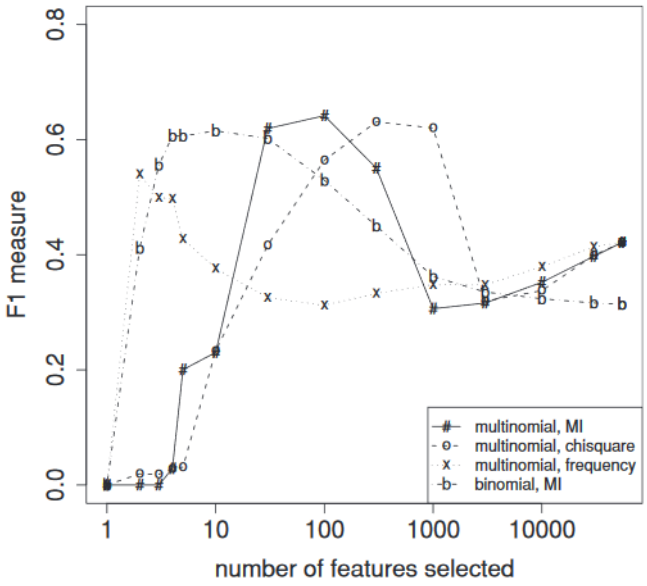

Figure 4: Benefici della feature selection su alcuni corpora REUTERS

Come si può notare, tramite la limitazione del numero di feature si raggiunge una F-measure più alta, in questo caso sia il classificatore multinomiale che quello binomiale è più accurato con circa 100 feature.

[11] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008

6 Sperimentazioni e validazione

La validazione di un sistema di text categorization è basata sulla classificazione di un campione di esempi che sono stati categorizzati manualmente da esperti del dominio. Quando si vuole valutare un sistema di text categorization si devono confrontare le categorie assegnate manualmente con le categorie assegnate dal sistema di text categorization.

Consideriamo la seguente casistica, nella quale diamo una rappresentazione binaria alle varie possibilità di assegnazione di una determinata categoria ad un determinato documento, distinguendo l'assegnazione del sistema da quella dell'esperto del dominio.

classe assegnata dall'esperto del dominio:	SI	NO
classe assegnata dal classificatore: SI	TP	FP
classe assegnata dal classificatore: NO	FN	TN

Abbiamo TP: true positive, FP: false positive, FN: false negative e TN: true negative; questi valori vanno interpretati nel seguente modo:

- **true positive:** il sistema e l'esperto del dominio sono coerenti nell'assegnazione della categoria.
- **false positive:** la categoria è assegnata dal sistema ma non dall'esperto del dominio.
- **true negative:** la categoria non è assegnata né dal sistema né dall'esperto del dominio.
- **false negative:** la categoria è assegnata dall'esperto ma non dal sistema.

6.1 Metriche per l'evaluation

Piuttosto che avere tutti questi valori binari per ciascuna coppia documento-categoria siamo interessati ad avere quattro valori che rappresentino ciascuna delle quattro tipologie (true positive, false positive, true negative, false negative), pertanto è necessario sommare questi quattro valori per tutte le coppie documento-categoria. Questi quattro valori sono le grandezze che stanno alla base di varie metriche utili per l'evaluation del sistema di classificazione.

$$Precision: \quad P = \frac{TP}{TP + FP}$$

La Precision è una misura utile per quantificare la precisione del sistema sul campione di categorie selezionate a valle della classificazione, è il rapporto tra le categorie correttamente assegnate e tutte quelle assegnate.

$$Recall: \quad R = \frac{TP}{TP + FN}$$

La Recall (in italiano "recupero") è utile per misurare la frazione di categorie assegnate dall'esperto a valle della classificazione del sistema. La recall è il rapporto tra categorie assegnate e tutte le categorie di appartenenza di un documento.
Recall e Precision sono due metriche basilari nei sistemi di information retrieval, trovare un equilibrio tra queste due grandezze non è facile, è necessario un compromesso, in quanto, incrementando la precision abbiamo il declino della recall e viceversa. In una categorizzazione multi-label nella quale restituiamo poche categorie per documento è verosimile avere precision alta e recall bassa, mentre invece, se assegniamo molte categorie sarà più probabile avere recall alta a scapito di precision.

$$Fallout: \quad F = \frac{FP}{FP + TN}$$

$$Accuracy: \quad F = \frac{TP + TN}{TP + FP + FN + TN}$$

Utile per quantificare l'accuratezza generale del sistema.

$$Error: \quad F = \frac{FP + FN}{TP + FP + FN + TN}$$

F-measure e Interpolated-break-even sono due grandezze che sintetizzano in un solo valore la precision e la recall di un sistema di classificazione:
La F-measure (anche nota come F1-score) è la media armonica

$$F - measure: \quad F1 = \frac{2PR}{P + R}$$

$$Interpolated - break - even: \quad IB = \frac{P + R}{2}$$

Tra le due metriche è preferibile usare l'F1, in quanto più sensibile a valori di precision o recall molto bassi.

6.2 K-fold-cross-validation

Gli algoritmi di apprendimento supervisionato per la text categorization (Bayes-multinomiale, Bayes-Bernoulli, VSM-Rocchio, VSM-Knn, SVM, sono tutti supervisionati) hanno bisogno di un test dataset ed un training dataset per poter essere valutati. Questi due insiemi devono essere disgiunti per avere delle buone stime delle performance del sistema di classificazione.
Training set: è un insieme di esempi usato per l'addestramento (apprendimento) del classificatore.
Validation set: è un insieme di esempi usato per regolare i parametri del classificatore.
Test set: è uni insieme di esempi usato per assegnare le performance di un classificatore addestrato.

Se non separassimo gli insiemi di test e di validation la stima di errore di un modello sulla validazione sarà parziale (inferiore) in quanto il validation set è utilizzato per selezionare il modello finale, infine si può procedere alla valutazione del modello finale sul test set.
La k-fold-cross-validation (in italiano validazione incrociata) è una tecnica utilizzabile in presenza di un training set abbastanza grande. In particolare la k-fold cross-validation consiste nella suddivisione del dataset totale in k parti di uguale numerosità e, ad ogni passo, la k-esima parte del dataset viene ad essere il test dataset, mentre la restante parte costituisce il training dataset. Così, per ognuna delle k parti (di solito k=10, k=5) si allena il modello, evitando quindi problemi di overfitting, ma anche di campionamento asimmetrico (e quindi affetto da bias) del training dataset, tipico della suddivisione del dataset in due sole parti (ovvero training e test dataset). In altre parole, si suddivide il campione osservato in gruppi di egual numerosità, si esclude iterativamente un gruppo alla volta e lo si cerca di predire con i gruppi non esclusi. Ciò al fine di verificare la bontà del modello di predizione utilizzato.

[11] Ricardo Gutierrez-Osuna, Introduction to Pattern Analysis, Texas A&M University

7 Sistemi di catalogazione

I sistemi di categorizzazione dei testi coprono differenti aeree di interesse. I sistemi di categorizzazione sono stati realizzati per rispondere alla necessità di catalogare ed indicizzare i documenti. I task di categorizzazione multi label sono molto complessi ed i risultati ottenuti spesso sono molto distanti dai livelli accettabili di accuratezza. Molti approcci sono stati esplorati ed implementati, tra questi abbiamo i sistemi keyword-based, i quali sono stati pensati per la realizzazione di abstract, vale a dire l'individuazione di parole chiave che diano una descrizione semantica e sintetica del contenuto dei documenti. Un'altra classe di algoritmi utilizzati per la categorizzazione dei testi è quella basata su tecniche di machine learning come i classificatori binari, algoritmi margin-based, tecniche bayesiane, ...
Nel seguito verranno descritti i sistemi che hanno avuto maggiore successo.

7.1 MeSH

MeSh (acronimo di Medical Subject Headings), è un indicizzatore automatico utilizzato dalla U.S National Library of Medicine, il quale utilizza circa 18'000 termini del campo medico. `https://www.nlm.nih.gov/pubs/factsheets/mesh.html`

7.2 BIOSIS

Uno dei più sofisticati programmi per l'indicizzazione automatica sviluppato negli studi BIOSIS è stato discusso da Vleduts-Stokolov; le parole che occorrono nei titoli delle pubblicazioni che sono anche contenute all'interno di un vocabolario semantico. Il quale vocabolario semantico rappresenta la terminologia scientifica rappresentata da circa 15'000 termini. I termini coprono i seguenti campi di studio: agricoltura, anatomia, comportamento, biochimica, bioingegneria, biofisica, biotecnologie, botanica, biologia cellulare, biologia ambientale, medicina clinica sperimentale, genetica, immunologia, microbiologia, patologia, farmacologia, fisiologia e tossicologia.
`www.cas.org/File%20Library/Training/STN/DBSS/biosis.pdf`
Questi 15'000 termini sono mappati su un vocabolario di 600 concetti (Concept Headings), pertanto un Concept Heading può essere assegnato dal sistema sulla base della presenza o meno di alcuni termini nei titolo. Vleduts-Stokolov ha documentato l'accuratezza di BIOSIS, notando che il sistema si comporta come un umano nella assegnazione dei concetti del vocabolario nel 61% dei casi. Al fine di rendere più accurato il sistema si è pensato di utilizzare più livelli di astrazione dei concetti che stanno al di sopra dei termini (primario, secondario e terziario), portando così l'accuratezza al 75%.

[11] Natasha Vleduts-Stokolov, Concept recognition in an automatic text-processing system for the life sciences, 1987

7.3 NASA MAI System

Nasa Center for Aerospace Information (CASI) ha un database contenente tre milioni di record, almeno due milioni sono report tecnici e articoli di giornale. La CASI indicizza i documenti al fine di ottenere degli abstract. Nasa MAI System (MAI è acronimo di Machine Aided Indexing) è un sistema di sostegno utilizzato dagli utenti (specialisti del dominio) per navigare e selezionare un'insieme di possibili parole chiave da utilizzare per annotare i documenti. Questo sistema è stato pensato per produrre delle annotazioni più coerenti, in quanto l'esperto del dominio è guidato dalle connessioni tra i concetti di un vocabolario.

Il progetto NASA MAI è stato pensato per minimizzare lo sforzo per la reindicizzazione di documenti già indicizzati da altre agenzie, infatti, quando il progetto è iniziato circa la metà dei report tecnici all'interno del NASA STI (database) erano stati indicizzati da agenzie diverse. Questi report, ricevuti in forma machine-readable, sono convertiti utilizzando tecniche di processamento del linguaggio naturale.

Lo scopo principale di NASA MAI System è quello di fornire un'insieme di termini "adeguati" per l'annotazione dei documenti. Molti documenti provenienti da altre agenzie utilizzavano dei termini non corrispondenti al vocabolario utilizzato dalla NASA. L'utilizzo di un unico vocabolario specializzato permette di collegare i concetti a vari livelli di astrazione, e sfruttare i sinonimi nella ricerca dei documenti. Per ogni termine, MAI, propone una serie di termini candidati corrispondenti ai "NASA terms".

Il NASA MAI System non è basato su sofisticate tecniche di processamento del linguaggio naturale e parsing grammaticale, segue un approccio rule-based che approssima i risultati dei parsing più approfonditi, e, che richiedono un notevole sforzo computazionale. Le regole sono sviluppate al fine di ottenere dei termini (feature testuali) che meglio rappresentino i concetti.

MAI usa una grande knowledge base (manutenuta), oltre 170'000 termini e frasi, oltre alle regole. I valori di precision e recall sono attorno al 50% (break-even point), sebbene possa sembrare basso questo risultato è piuttosto buono se si considera la vastità di argomenti coperti e la diversità dei documenti processati.

7.4 JRC-JEX (EuroVoc)

Il classificatore JEX, sviluppato presso i laboratori JRC (Joint Research Centre), è un classificatore multilabel basato sul tesauro EuroVoc. La scelta di basare il classificatore su EuroVoc rende il sistema multilingua e garantisce la copertura di tutte le tematiche di interesse della politica europea. Le annotazioni prodotte

da JEX sono i descrittori dei concetti di EuroVoc. Sul sito di JEX si descrivono in sintesi le funzionalità di questo sistema:

"Multilingual Eurovoc thesaurus descriptors are used by a large number of European Parliaments and Documentation Centres to manually index their large document collections. The assigned descriptors are then used to search and retrieve documents in the collection and to summarise the document contents for the users. As Eurovoc descriptors exist in one-to-one translations in almost thirty languages, they can be displayed in a language other than the text language and give users cross-lingual access to the information contained in each document. At the same time, EuroVoc is an ideal means to search in the user's language and to retrieve documents in other languages. The European Commission's (EC) Joint Research Centre (JRC) has developed - and makes available - software that automatically assigns EuroVoc descriptors to documents in currently 22 languages. The system uses statistical Machine Learning methods that learn the multi-label categorisation rules from previously manually indexed documents. The method used can be described as profile-based category ranking. This software, called JRC EuroVoc Indexer, or short JEX, has been trained for 22 languages and is available for download from this site. The software allows users to re-train the software on their own data, even using their own, alternative classification systems"

8 Aspetti software

8.1 Architettura

8.2 Deployment

9 Bibliografia e risorse

9.1 Bibliografia

-
-
-
-
-

9.2 Link utili

- https://github.com/simonegasperoni/cata
- http://opendefinition.org/od/2.1/en
- http://opendatahandbook.org/guide/it/what-is-open-data
- http://dati.mit.gov.it/dcat/catalog.rdf
- https://www.w3.org/TR/vocab-dcat
- https://ckan.org
- https://github.com/ckan/ckan
- https://datahub.io
- https://europeandataportal.eu
- http://sciamlab.com/opendatahub/it
- http://dati.gov.it
- https://github.com/ckan/ckanext-dcat
- https://github.com/ckan/ckanext-dcat#rdf-dcat-to-ckan-dataset-mapping
- http://eurovoc.europa.eu
- https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis
- https://ec.europa.eu/jrc/en/language-technologies/jrc-eurovoc-indexer
- http://tei-c.org
- http://snowballstem.org
- https://github.com/snowballstem/snowball

- https://www.nlm.nih.gov/pubs/factsheets/mesh.html
- www.cas.org/File%20Library/Training/STN/DBSS/biosis.pdf
-
-
-
-
-
-

10 Appendice

10.1 Associazione Temi DCAT-Microtesauri EuroVoc

Microtesauri EuroVoc associati al tema DCAT-AP (EU Data Themes): AGRI (Agriculture, Fisheries, Forestry & Foods)

```
5606 agricultural policy
5611 agricultural structures and production
5616 farming systems
5621 cultivation of agricultural land
5626 means of agricultural production
5631 agricultural activity
5636 forestry
5641 fisheries
6006 plant product
6011 animal product
6021 beverages and sugar beverage sugar
6026 foodstuff
6031 agri-foodstuffs
6036 food technology
```

Microtesauri EuroVoc associati al tema DCAT-AP (EU Data Themes): ENER (Energy)

```
6606 energy policy
6616 oil industry
6621 electrical and nuclear industries
6626 soft energy
6611 coal and mining industries
```

Microtesauri EuroVoc associati al tema DCAT-AP (EU Data Themes): REGI (Regions & Cities)

```
7206 Europe
7211 regions of EU Member States
7216 America
7221 Africa
7226 Asia and Oceania
7231 economic geography
7236 political geography
7241 overseas countries and territories
```

Microtesauri EuroVoc associati al tema DCAT-AP (EU Data Themes): TRAN (Transport)

```
4806 transport policy
4811 organisation of transport
4816 land transport land transport
```

4821 maritime and inland waterway transport
4826 air and space transport

Microtesauri EuroVoc associati al tema DCAT-AP (EU Data Themes): ECON (Economy & Finance)

1606 economic policy
1611 economic growth
1616 regions and regional policy
1621 economic structure
1626 national accounts
1631 economic analysis
2406 monetary relations
2411 monetary economics
2416 financial institutions and credit
2421 free movement of capital
2426 financing and investment
2431 insurance
2436 public finance and budget policy
2441 budget
2446 taxation
2451 prices
2006 trade policy
2011 tariff policy
2016 trade supply
2021 international trade
2026 consumption
2031 marketing
4006 business organisation
4011 business classification
4016 legal form of organisations
4021 management
4026 accounting
4031 competition
6806 industrial structures
6811 chemistry
6816 iron
6821 mechanical engineering
6826 electronics and electrical engineering
6831 building and public works
6841 leather and textile industries
6846 miscellaneous industries

Microtesauri EuroVoc associati al tema DCAT-AP (EU Data Themes): INTR (International Issues)

7606 United Nations

```
7611 European organisations
7616 extra-European organisations
7621 world organisations
7626 non-governmental organisations
0811 cooperation policy
0816 international balance
0821 defence
```

Microtesauri EuroVoc associati al tema DCAT-AP (EU Data Themes): GOVE (Government & Public Sector)

```
0406 political framework
0411 political party
0416 electoral procedure and voting election
0421 parliament
0426 parliamentary proceedings
0431 politics and public safety
0436 executive power and public service
1006 Community institutions and European civil service
1011 European Union law
1016 European construction
```

Microtesauri EuroVoc associati al tema DCAT-AP (EU Data Themes): JUST (Justice, Legal System & Public Safety)

```
1206 sources and branches of the law
1211 civil law
1216 criminal law
1221 justice access to the courts
1226 organisation of the legal system
1236 rights and freedoms
```

Microtesauri EuroVoc associati al tema DCAT-AP (EU Data Themes): ENVI (Environment)

```
5206 environmental policy
5211 natural environment
5216 deterioration of the environment
```

Microtesauri EuroVoc associati al tema DCAT-AP (EU Data Themes): EDUC (Education, Culture & Sport)

```
3206 education
3211 teaching
3216 organisation of teaching
3221 documentation
3226 communications
3231 information and information processing
3236 information technology and data processing
2831 culture and religion arts cultural policy culture
```

Microtesauri EuroVoc associati al tema DCAT-AP (EU Data Themes): HEAL (Health)

```
2841 health
```

Microtesauri EuroVoc associati al tema DCAT-AP (EU Data Themes): SOCI (Population & Society)

```
2806 family
2811 migration internal
2816 demography and population composition
2821 social framework
2826 social affairs
2836 social protection
2846 construction and town planning
4406 employment
4411 labour market
4416 organisation of work and working conditions
4421 personnel management and staff remuneration
4426 labour law and labour relations
```

Microtesauri EuroVoc associati al tema DCAT-AP (EU Data Themes): TECH (Science & Technology)

```
3606 natural and applied sciences
3611 humanities behavioural
6406 production
6411 technology and technical regulations
```