Information retrieval e analisi delle cooccorrenze per l'estrazione di informazione specifica da documentazione giuridica*

Alessio Canzonetti[1]

[1]Sapienza Università di Roma

Abstract

This paper presents a method of Information Extraction from judicial acts. More specifically, the criminal offence under verdict is the information that needs to be extracted. The wide range of different types of crimes does not allow the prototypical textual entities to be easily detected through the use of dictionaries. For this reason, it is necessary to analyze local contexts in order to identify the main structures of desired information. Due to this large variety, the detection of local contexts cannot be performed through concordance analysis of the crimes, which are often unknown. Hence, a preliminary localization of these contexts is needed through information retrieval or automatic classification techniques. There are two stages to the process. First, the sentences which refer to the criminal offence are retrieved through a TFIDF index. To do this, it is necessary to provide a sample of acts with the sentences categorized according to the crime mention or the lack thereof. A specificities analysis on this training corpus provides the specific typeset which constitutes the retrieving query. In this way, through an analysis of co-occurrences and collocations, it is then possible to identify the structures of desired information in the retrieved sub-corpus.

Keywords: information extraction from judicial documentation, text categorization, TF IDF, co-occurrences, information retrieval.

Riassunto

Questo lavoro presenta un metodo per l'estrazione di informazione a partire da atti processuali di una giurisdizione contabile. In particolare, l'informazione desiderata è costituita dal reato che costituisce l'oggetto in giudizio. L'ampia gamma delle singole fattispecie di reato non consente di rintracciare questo tipo di entità testuali per mezzo di dizionari. Occorre pertanto passare per l'analisi dei contesti locali al fine di identificare le principali strutture portatrici dell'informazione desiderata. Sempre a causa della elevata variabilità grafica delle fattispecie di reato, l'individuazione dei contesti locali non può passare per un'analisi delle concordanze delle forme grafiche riconducibili ai reati, peraltro non note. Si rende quindi necessaria una localizzazione preventiva dei contesti operabile con tecniche di recupero di informazione o di classificazione automatica. Il lavoro si articola in due fasi. Dapprima si recuperano quei periodi che contengono la citazione del reato, mediante utilizzo dell'indice TFIDF. Occorre perciò disporre di un campione di sentenze in cui i periodi siano già categorizzati con riferimento al fatto che contengano o meno la citazione del reato. Eseguendo un'analisi delle specificità sul campione di apprendimento così partizionato, si ottiene l'insieme di forme specifiche che va a costituire la query in ordine alla quale vengono recuperati detti periodi. Sul sub-corpus così estratto, si individuano le strutture portatrici di informazione desiderata attraverso l'analisi delle cooccorrenze e delle collocazioni.

1. Introduzione

L'estrazione di informazione è l'ambito nel quale si inquadra il problema di individuare e recuperare l'informazione desiderata, intendendo con questo termine una qualsiasi entità di

* Il presente lavoro è finanziato su fondi MIUR Facoltà 2005 - C26F059955.

interesse presente all'interno di un testo o, meglio, di una collezione di testi. Uno degli scopi principali di questo procedimento è rappresentato dal popolamento di database in cui, di fatto, si ha intenzione di strutturare informazione non strutturata, qual è quella tipicamente contenuta nei dati di natura testuale. Una tale esigenza può avere più motivazioni, che vanno dalla semplice archiviazione di dati, all'attribuzione di metadati da associare ai singoli documenti, oppure per costruire matrici di dati strutturati da utilizzare in successive analisi di tipo statistico.

La costruzione di un modello di estrazione di informazione è di per sé una operazione onerosa, che generalmente non può prescindere da una preventiva osservazione diretta dei testi da cui ricavare i dati per poter approntare le regole di individuazione dell'informazione desiderata. Infatti, una volta che si sono decise quali informazioni catturare, cioè a dire la struttura dati che si deve alimentare, occorre operare un'indagine preliminare riguardante le forme e i contesti in cui tali informazioni tendono a presentarsi nei testi. È appena il caso di precisare che forme e contesti in cui l'informazione desiderata si presenta devono intendersi come strutture, grammatiche locali, piuttosto che attualizzazioni fisse. Comunque sia, il principale problema da affrontare rimane il fatto che non solo tali strutture, ma anche le semplici forme grafiche con cui l'informazione può essere espressa, non sono note a priori, se non nel caso in cui si abbia interesse ad individuare entità che per loro natura vengono espresse con una loro particolare struttura, si pensi alle date o agli indirizzi di posta elettronica.

Questo lavoro affronta, pertanto, il problema della costruzione di un modello di estrazione di informazione espressa con forme e strutture non note, proponendo una metodologia che può essere suddivisa in due fasi. La prima consiste nell'individuazione di quei periodi, sintatticamente intesi, che, con ragionevole grado di approssimazione, possono contenere l'informazione desiderata. A questo scopo possono essere utilizzate tecniche di *Information Retrieval* o di *Text Categorization*. In un secondo momento si analizza il sub-corpus recuperato nella fase precedente per individuare le principali strutture di informazione desiderata, in modo da poter formalizzare le regole generali di estrazione.

2. Il corpus in analisi

Il corpus oggetto della sperimentazione è costituito da 300 sentenze della Corte dei Conti[1] relative a giudizi in materia di responsabilità amministrativa. In particolare sono state prese in considerazione solo sentenze riguardanti il secondo grado di giudizio (appello).

In virtù di queste scelte, il corpus è costituito da 1.005.330 occorrenze (*token*) relative a 30.679 forme diverse (*type*) 42,4% e con una percentuale di hapax pari al 42,4%.

Questo genere di testi è caratterizzato da uno stile estremamente formale e sintatticamente molto elaborato, con un lessico fortemente settoriale. Nonostante la caratteristica di estrema formalità, che generalmente porta all'utilizzo di formule espressive ricorrenti, possa giocare un ruolo favorevole per i compiti di individuazione dell'informazione, occorre tenere presente che i diversi provvedimenti vengono stesi da personale diverso, non solo perché diversi sono i giudici che operano nella Corte, ma anche perché in certi casi le sentenze vengono verbalizzate da personale di segreteria, e in genere non esiste un modulo di stesura predefinito che deve essere necessariamente rispettato. Inoltre, vale la pena notare che in Italia esistono

[1] La Corte dei Conti è l'organo di giurisdizione contabile della Repubblica Italiana.

21 sezioni regionali, che si occupano dei giudizi di primo grado, e 4 sezioni centrali che si occupano dei giudizi di appello. Una tale suddivisione porta anche all'instaurazione di differenti prassi di stesura delle sentenze, con la conseguenza che non è possibile identificare una forma condivisa di presentazione dei dati contenuti nelle sentenze. Tutto ciò, ovviamente, nei limiti di contenuto effettivo che una sentenza deve comunque rispettare.

In conclusione, per quanto riguarda le caratteristiche dei testi in analisi, è possibile affermare che esiste una certa uniformità di registro unita ad una differenza di stile, sia in senso stilometrico che di dettaglio delle informazioni.

3. Recupero dei contesti

Senza perdere in generalità, si è deciso di concentrare l'attenzione su un solo tipo di informazione da individuare ed estrarre. La metodologia qui presentata è infatti senza dubbio applicabile a qualsiasi altro genere di informazione che presenti caratteristiche analoghe. Scendendo più in dettaglio, il dato che costituisce l'informazione desiderata è rappresentato dal reato che costituisce l'oggetto del giudizio di volta in volta preso in esame.

Pur esistendo una classificazione/tassonomia dei reati perseguibili dalla Corte, nei provvedimenti della stessa difficilmente, o quantomeno non sempre, la menzione del reato è espressa utilizzando la rispettiva voce tassonomica, ovvero con la forma canonica con cui il reato è giuridicamente codificato. Infatti, le fattispecie ascrivibili ad ogni singolo tipo di reato possono essere molteplici e, come tali, possono essere espresse con diverse forme, quasi sempre composte da più parole.

Per quanto appena detto, non è possibile conoscere a priori tutte le forme, semplici e composte, in cui un reato può essere espresso, e ciò rende impossibile utilizzare l'analisi delle concordanze per recuperarle. Inoltre, anche se si conoscessero le principali forme semplici che entrano nella costruzione delle espressioni di reato, l'analisi delle concordanze di queste forme sarebbe comunque lunga e laboriosa, in quanto è plausibile ipotizzare che i contesti di utilizzo di dette forme semplici non si limitino alle espressioni di reato, ma si estendano ad altri ambiti.

Da ciò, è sembrato naturale per prima cosa individuare i contesti in cui compaiono le espressioni relative ai reati[2]. A questo scopo si rende necessaria la messa a punto di un campione di apprendimento. In particolare, ogni periodo delle sentenze di questo campione di apprendimento viene marcato, per mezzo di una apposita variabile dicotomica, relativamente al fatto di contenere o meno la menzione dell'informazione desiderata, nel nostro caso del reato oggetto di giudizio. Pertanto, la base documentale viene frammentata non a livello di sentenza, bensì a livello di singoli periodi[3].

Sul corpus di apprendimento così predisposto si effettua un'analisi delle specificità (Lafon, 1984) considerando la partizione definita da questa variabile, ottenendo così il linguaggio specifico dei periodi contenenti la menzione di reato, ovvero di tutti i contesti di interesse. A partire da questa base è possibile applicare tecniche di *Information Retrieval* per recuperare lo stesso genere di contesti anche su altri insiemi di sentenze, anch'esse ovviamente frammentate a livello di singoli periodi.

[2] In (Balbi e Di Meglio, 2004), pur con finalità diverse, viene seguito un medesimo approccio.

[3] Ogni periodo possiederà, oltre alla variabile appena menzionata, anche un identificativo univoco e una variabile relativa alla sentenza di appartenenza del periodo stesso.

Lo strumento utilizzato a questo scopo è stato il calcolo del TFIDF (Salton et Buckley, 1988). Sulla matrice *Frammenti x Forme*[4] del corpus del campione di validazione si calcola la matrice dei pesi espressi dai TFIDF[5] delle forme per ogni frammento. Successivamente si ordinano i frammenti in senso decrescente rispetto alla somma dei TFIDF delle forme appartenenti al linguaggio specifico individuato in precedenza. Questo linguaggio, di fatto, opera come interrogazione di recupero. La graduatoria che si ottiene mostra ai ranghi più bassi i frammenti che risultano essere maggiormente pertinenti, quanto a contenuto, alla query di recupero, ossia sono i frammenti del campione di validazione che, con buona probabilità, contengono la menzione del reato.

Per l'individuazione del sub-corpus su cui effettuare la successiva analisi di identificazione delle strutture portatrici di informazione, sorge un problema riguardo alla quantità di frammenti, a rango più basso ovviamente, da selezionare a tale scopo. Non è infatti possibile formulare ipotesi circa valori del TFIDF che possano avere il ruolo di soglie di discriminazione tra frammenti "pertinenti" e frammenti "non pertinenti" rispetto all'interrogazione effettuata. In una situazione del genere, una soluzione ammissibile può essere quella di considerare un numero di frammenti pari al numero di sentenze del corpus in esame. Questa quantità può essere anche leggermente aumentata in virtù del fatto che, in alcuni casi, nel campione di apprendimento si era osservato che in alcune sentenze la menzione del reato si estendeva su due periodi/frammenti.

In alternativa, si possono utilizzare per il medesimo scopo tecniche di *Text Categorization* (Basili et Moschitti, 2005), che consentirebbero di individuare i frammenti pertinenti in maniera anche più raffinata e forniscono maggiore ausilio anche per l'individuazione di eventuali falsi positivi, al costo di una maggiore complessità computazionale. Tuttavia, la validità generale del metodo non viene inficiata dal tipo di tecnica adottata per il recupero dei contesti di interesse.

4. Le strutture portatrici di informazione

Una volta che è stato recuperato il sub-corpus costituito dai frammenti che, con buona approssimazione, contengono l'oggetto dell'analisi, si passa all'individuazione delle strutture portatrici di informazione desiderata. Si tratta cioè di costruire dei *templates* che sintetizzino i modelli espressivi utilizzati nelle menzioni dei reati.

Si è detto che tali menzioni sono quasi sempre espresse per mezzo di forme composte, segmenti di testo. Nell'ambito di ogni medesima tipologia di reato, le singole fattispecie si

[4] È una matrice in cui gli elementi di riga sono costituiti dai frammenti in cui il corpus è suddiviso, mentre in colonna figurano le forme dell'intero vocabolario. Il contenuto della matrice rappresenta il numero di occorrenze con cui ogni forma compare in ogni singolo frammento.

[5] Nello specifico, è stata utilizzata la seguente formulazione del TFIDF:

$$TFIDF(i,j) = \frac{tf_i * \log\frac{N}{n}}{\sqrt{\sum_i \left(tf_i * \log\frac{N}{n}\right)^2}}$$

dove: tf_i è la frequenza della forma i-esima nel frammento j-esimo, N è il numero totale dei frammenti, n è il numero dei frammenti in cui compare la forma i-esima.

trovano ad essere graficamente simili ma non identiche. È plausibile pensare, quindi, che le principali associazioni tra forme semplici siano dovute proprio a queste entità.

L'individuazione delle strutture tipiche dell'informazione da estrarre è stata condotta attraverso lo studio delle cooccorrenze e delle collocazioni esistenti tra le forme semplici.

Si ha una cooccorrenza quando due parole compaiono nel testo entro un medesimo contesto. La definizione di questo contesto può essere diversa a seconda dei casi. Nel nostro caso, un contesto è definito come un ambito che al massimo può estendersi fino a *n* parole, e comunque non può estendersi oltre i limiti definiti dalla punteggiatura presente all'interno del testo. Per le nostre analisi è stato utilizzato un $n = 14$.

Il risultato principale del calcolo delle cooccorrenze è una matrice quadrata del tipo *Forme x Forme* di cui uno stralcio è presentato in *Tabella 1*.

	impugnata	***erariale***	***euro***	***giurisdizionale***	***rivalutazione***	***citazione***	***sentenza***
danno	**0**	**20**	**2**	**0**	**0**	**2**	**0**
lire	**0**	**3**	**0**	**0**	**1**	**0**	**2**
veniva	**0**	**3**	**0**	**0**	**0**	**0**	**0**
interessi	**0**	**2**	**1**	**0**	**12**	**0**	**4**
Comune	**0**	**2**	**1**	**0**	**0**	**1**	**1**
Regione	**1**	**1**	**0**	**4**	**1**	**1**	**3**
euro	**0**	**1**	**4**	**0**	**3**	**0**	**0**

Tabella 1 – Matrice delle cooccorrenze (stralcio)

Concentrando l'attenzione sulle coppie di forme che mostrano i più alti valori di cooccorrenze, in termini assoluti o relativi (Canzonetti, 2007), si possono individuare le relazioni più forti esistenti in questo sub-corpus. Tuttavia, la matrice delle cooccorrenze non presenta alcuna informazione circa la posizione relativa delle forme, nel testo, nel momento in cui si verifica una cooccorrenza. Pertanto, dopo aver valutato quali siano le associazioni più interessanti, si passa all'analisi delle collocazioni esistenti tra una particolare forma e tutte le altre.

Le collocazioni permettono di ottenere la distribuzione posizionale delle cooccorrenze di una data forma, ovvero il numero di volte che una forma compare *n* posizioni precedenti o successive rispetto ad una forma perno.

Posizioni →	**-7**	**-6**	**-5**	**-4**	**-3**	**-2**	**-1**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
erariale	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**20**	**0**	**0**	**0**	**0**	**0**	**0**
veniva	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**1**	**1**	**1**	**0**	**0**	**0**	**0**
condannare	**0**	**0**	**1**	**2**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
ente	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**1**	**2**	**0**	**0**	**0**
giudizio	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Tabella 2 – Matrice delle collocazioni della forma <danno> (stralcio)

Attraverso lo studio dei profili delle collocazioni, in particolare tra quelle forme che presentano maggiori valori di cooccorrenza, si è in gradi di giungere alla individuazione delle strutture che identificano gli schemi più ricorrenti con cui l'informazione desiderata si presenta nel testo (Canzonetti, 2007).

Questi schemi possono essere tradotti in interrogazioni di ricerca sul testo in modo da giungere all'estrazione di tutte le attualizzazioni testuali corrispondenti. Ad esempio, intense

associazioni tra le parole <maggiori> ed <erogati>, in questo ordine e a bassa distanza, e tra <erogati> e <Consiglio>, con distanza più elevata, può portare alla costruzione della seguente regola di estrazione[6]:

```
"maggiori LAG2 erogati LAG10 Consiglio"
```

che consente di individuare fattispecie di reato quali <maggiori compensi erogati a danno del Consiglio comunale> ma anche <maggiori emolumenti erogati per i lavori di pulizia degli Uffici del Consiglio Regionale>, entrambe riconducibili alla tipologia di reato *somme non dovute*.

Queste strutture si possono ottenere anche con forme maggiormente generalizzabili utilizzando le categorie grammaticali e/o categorizzazioni semantiche. La seguente regola:

```
"maggiori CATGR(N) corrisposti" OR "maggiori CATGR(N) erogati"
```

porta all'individuazione di <maggiori compensi corrisposti> e di <maggiori interessi erogati>. Se si fossero poste in relazione di sinonimia, attraverso una categorizzazione semantica, le parole <corrisposti> ed <erogati>, la seguente regola sarebbe ancora più generale:

```
"maggiori CATGR(N) CATSEM(corrispostierogati)"
```

L'esistenza di strutture maggiormente complesse può essere meglio apprezzata ricorrendo alle poli-coooccorrenze (Martinez, 2003), ovvero alle cooccorrenze calcolate tra tre o più forme grafiche, e non solo su semplici coppie. Il risultato del calcolo delle poli-coooccorrenze è costituito da un inventario, attestato in frequenza, di tutte le sequenze osservate nel contesto considerato.

Sequenza	*Poli-cooccorrenze*
danno, conseguente, omessa, esclusione, assistiti, deceduti, elenchi, periodicamente, determinate, retribuzioni, quota, capitaria, erogate, medici, di base, pediatri	2
illecito, produttivo, danno, contestato, convenuti, scaturito, violazione, articolo, comma, settimo, legge, numero	2
illegittimo, inquadramento	5
illegittimo, inquadramento, dipendente	2
spesa, illegittima	2
maggiori, somme	3
maggiori, emolumenti	2
maggiori, importi	2
maggiori, oneri	2
lievitare, spesa, iniziale, maggiori, importi	2

Tabella 3 – Inventario delle poli-cooccorrenze (stralcio)

In Tabella 3 sono mostrate alcuni casi paradigmatici. I primi due casi rappresentano regole molto complesse e dettagliate, che consentono di individuare casi molto particolari di illecito.

A seguire, <illegittimo,inquadramento,dipendente> rappresenta un tipico caso di estensione della sequenza <illegittimo,inquadramento> con maggior livello di dettaglio[7].

[6] La regola è scritta secondo la sintassi per le interrogazioni testuali adottata dal software utilizzato per l'analisi TaLTaC[2] (www.taltac.it). In particolare l'operatore LAG indica che la parola di destra può comparire, nel testo, con un ritardo massimo, rispetto alla parola di sinistra, indicato dal numero che segue l'operatore stesso.

[7] Si fa notare che le quantità riportate nella colonna Poli-cooccorrenze della Tabella 3 non sono affette da ridondanza.

Generalmente, le sequenze più lunghe in termini di forme componenti sono quelle che meglio riescono ad individuare una struttura di interesse. Tuttavia, le sequenze più corte tendono ad avere una frequenza più elevata, e ciò le rende maggiormente individuabili. Questo fatto permette di sfruttare le sequenze più corte per "tracciare" l'esistenza di sequenze più lunghe, più idonee allo scopo di estrarre informazione maggiormente dettagliata.

Infine, tra le sequenze più corte, troviamo alcuni esempi di ciò che si associa con *maggiori* (*somme*, *emolumenti*, *importi*, *oneri*). In questo caso è evidente come una categorizzazione semantica, ovvero il porre in relazione di sinonimia i termini di destra delle relazioni in esame, avrebbe senza dubbio migliorato la rilevazione del concetto relativo al <maggior esborso> causato ai danni della Pubblica Amministrazione.

5. Conclusioni e futuri sviluppi

Il metodo presentato offre la possibilità di estrarre un tipo di informazione caratterizzato da una variabilità, a livello grafico, relativamente ampia. Per arrivare a questa estrazione è necessario prima localizzarla, ovvero recuperare i contesti entro i quali può esistere. A questo scopo si ricorre a tecniche di *Information Retrieval*. Vale la pena sottolineare come questo *task* relativo alla localizzazione sia perseguibile anche attraverso altre tecniche, quali, ad esempio, quelle offerte dalla categorizzazione automatica dei testi (Basili et Moschitti, 2005). Il ricorso all'analisi delle specificità congiuntamente a quello del TFIDF presentato in questo lavoro è stato dettato principalmente dalla semplicità di utilizzo e dalla scarsa complessità computazionale, senza comunque che ciò vada a togliere validità generale al metodo.

Successivamente, sul sub-corpus individuato, si effettuano analisi dei contesti locali, per mezzo di cooccorrenze e collocazioni, al fine di giungere a regole di estrazione sufficientemente sintetiche ed esustive.

Uno scenario di sviluppo possibile è costituito dalla categorizzazione dell'informazione estratta in categorie predefinite. Nel caso in esame, infatti, l'informazione estratta rappresenta tutta una serie di fattispecie di reato riferibili a diverse tipologie. Pertanto, con l'obiettivo di giungere alla definitiva strutturazione dell'informazione estratta dal testo, tecniche di categorizzazione automatica potrebbero consentire di attribuire ogni fattispecie alla rispettiva tipologia di reato, e quindi classificare con modalità standard ogni sentenza della Corte dei Conti in base alla stessa.

Bibliografia

Balbi S., Di Meglio E. (2004). A Text Mining Strategybased on Local Contexts of Words. In Purnelle G., Fairon C., Dister A. (eds.), *Les poids des mots*, Actes JADT'2004, Louvain, Presses universitaires de Louvain, 79-87

Basili R. & Moschitti A. (2005). *Automatic Text Categorization – From Information Retrieval to Support Vector Learning*. Aracne.

Canzonetti A. (2007). Semantic classification and cooccurrences: a method for the rules production for the information extraction from textual data. In *Classification and Data Analysis 2007 – Book of Short Paper (CLADAG 2007)*, Macerata, pages 259-262

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Slatkine-Champion.

Martinez W. (2003). Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels. Thèse de Doctorat en Sciences du Langage, Université de la Sorbonne nouvelle - Paris 3, Paris.

Rajman M., Lebart L. (1998). Similarités pour données textuelles. In Mellet S., editor, *Actes des 4es Journées internationales d'Analyse statistique des Données Textuelles (JADT98)*, Nice, pages 545-555.

Salton G. & Buckley C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5): 513-523.