Estrazione di informazione da una base documentale dell'AGCM con il software TaltaC2 : un esempio di integrazione fra strumenti di text mining e tecniche di data mining

Alessio Canzonetti[1], Federico Maria Capo[1], Valerio Ruocco[2]

[1] Università degli Studi di Roma "La Sapienza" – Via del Castro Laurenziano, 9, 00161, Rome - Italy
[2] Autorità Garante della Concorrenza e del Mercato - Piazza Giuseppe Verdi, 6/a, 00198, Rome - Italy

Abstract

This work presents a procedure for information extraction from the documents of the Italian Antitrust Authority (AGCM) making use of the TaltaC2 software.
The procedure involves the step of creating complex textual queries in addition to usual steps such as corpus building, cleaning and linking to *a priori* categorical variables. Once submitted the queries enable the finding of concepts and named entities within the documents, leading to document categorisation in answer to each query. The queries are built directly from the analysis of the corpus. The entire procedure aims at the creation of a matrix to associate each document (identified by a primary key and by any other *a priori* categorical variable) to a set of *a posteriori* textual variables. Those variables point out the presence/absence or a certain characteristic of a concept under scrutiny. Finally, once integrated with diverse macroeconomic quantitative indexes, the matrix has been used as a database to set up a regression model for Authority intervention estimation.

Riassunto

In questo lavoro si illustra una procedura di estrazione di informazione mediante il software TaltaC2, a partire da una base documentale dell'Autorità Garante della Concorrenza e del Mercato. Oltre ai normali passi riguardanti la costruzione, la pulizia del corpus e l'associazione di informazioni categoriali ai documenti costituenti il corpus, la procedura si concretizza nell'individuazione di query complesse costruite direttamente dall'analisi del testo. Le query sono finalizzate all'individuazione di concetti ed entità di interesse all'interno dei documenti ed alla loro conseguente categorizzazione. Il risultato della procedura è una matrice che associa a ciascun documento, identificato da una chiave univoca e dalle variabili categoriali *ex ante*, un insieme di variabili *ex post* indicanti la presenza/assenza o una particolare qualifica del concetto oggetto di analisi. La matrice, integrata con indici quantitativi macroeconomici, è stata adoperata come base di dati per la costruzione di un modello di regressione per la stima della probabilità di intervento dell'Autorità.

Keywords : information extraction, information retrieval, text mining, data mining, base documentale.

1. Introduzione

L'Autorità Garante della **Concorrenza** e del Mercato (AGCM)[1] è l'organismo pubblico indipendente che vigila sulle operazioni di fusione ed acquisizione tra aziende, a protezione

[1] http://www.agcm.it

della concorrenza sui mercati italiani[2]. Le decisioni finali dell'Autorità in materia di concentrazione vengono progressivamente accumulate in una base documentale.

Tra il 1990 e il 2005 l'attività dell'Autorità ha generato un corpus di oltre 5.500 provvedimenti. Le informazioni contenute in tali provvedimenti vengono estratte e classificate manualmente all'interno di una banca dati relazionale. La banca dati è incentrata sul concetto di mercato rilevante, sintesi delle tre componenti mercato del prodotto, dimensione territoriale (sovranazionale, nazionale, locale) e area specifica di riferimento, ma registra anche informazioni sulle aziende e sui dettagli delle operazioni di fusione ed acquisizione.

Nell'intenzione dell'Autorità l'analisi retrospettiva delle informazioni aggregate sulle imprese coinvolte in tali operazioni e sulle operazioni stesse potrebbe consentire di delineare diversi sottoinsiemi di mercati rilevanti, all'interno dei quali l'intervento di prevenzione dell'Autorità possa divenire anticipabile in termine di esito, sulla base dei valori (soglie di attenzione) assunti da determinati indicatori macroeconomici di sintesi.

Il buon esito dell'analisi richiedeva l'integrazione dei dati sistematicamente raccolti nella propria banca dati con informazioni di tipo qualitativo - estratte con procedure di *information retrieval* e *information extraction*[3] dal corpus dei provvedimenti - attualmente non presenti nella stessa banca dati, ma fortemente caratterizzanti il sistema di valutazione degli effetti anticoncorrenziali delle operazioni di concentrazione.

Il recupero e l'estrazione dell'informazione testuale dal corpus è avvenuta tramite il software TaltaC2 (*Trattamento Automatico Lessicale & Testuale per l'Analisi del Contenuto di un Corpus*)[4], integrato da alcune funzionalità appositamente realizzate per l'Autorità, tali da garantire un utilizzo ancora più flessibile dello strumento.

Attraverso la sperimentazione di procedure automatiche di classificazione della propria documentazione e l'impiego di query complesse, generate a partire dall'analisi del testo, l'Autorità ha potuto costruire una matrice del tipo unità/variabili su cui applicare modelli di regressione *logit* binomiale per la stima degli indicatori macroeconomici di sintesi.

2. La base documentale dell'AGCM

La maggior parte dei provvedimenti emessi dall'Autorità è caratterizzata nella forma, nella terminologia e nei concetti da un certa regolarità, dovuta ad esigenze di legge. Le regolarità che caratterizzano questi testi li rendono particolarmente adatti ad un'analisi del contenuto.

Ogni provvedimento è costituito da un singolo file di testo (in formato < .txt > o < .doc >) ed è suddiviso in 6 sezioni principali (tra cui *Le parti*, *Descrizione dell'operazione*, *Qualificazione dell'operazione* e *Valutazione della concentrazione*), ognuna corrispondente ad un preciso aspetto del dispositivo della decisione dell'Autorità. A ciascun provvedimento sono associati dei metadati (variabili categoriali *ex ante*) quali il numero del Bollettino

[2] Secondo quanto dettato dalla legge "Norme per la tutela della concorrenza e del mercato – Legge 10 ottobre 1990 n. 287, Gazzetta Ufficiale 13 ottobre 1990, n. 240".
[3] Poibeau, 2003.
[4] Il software è sviluppato a partire da ricerche condotte presso la Facoltà di Economia dell'Università "La Sapienza" di Roma (Bolasco et al., 1999 ; Bolasco, 2000). Per una breve presentazione del software si veda la pagina internet http://www.taltac.it/it/anteprima_TALTAC_2005.pdf .
Per una descrizione del programma si veda la guida del software (Bolasco et al., 2005).

mensile all'interno del quale il provvedimento è stato pubblicato, l'anno di pubblicazione, il titolo (parti acquirenti/parti acquisite).

La funzione di *creazione multifile del corpus* ha permesso di costruire automaticamente il corpus in TaltaC2 (una selezione di 3549 documenti relativi al solo periodo 1995-2003, per un'ampiezza complessiva di 28MB di file di testo), riproducendolo all'interno di una (o più) *sessioni di lavoro*. Con la stessa funzione è stato possibile associare ai rispettivi documenti le variabili categoriali *ex ante* prima descritte, acquisendole da un foglio elettronico.

Attraverso il *parsing* del corpus, ovvero la sua lettura ed acquisizione, TaltaC2 crea una tabella che mette in relazione ciascun documento con i rispettivi valori di queste variabili. La tabella continua a popolarsi con la creazione di nuovi campi (variabili *ex-post*) ad ogni operazione di *information extraction* (si veda il paragrafo 4).

3. Estrazione automatica di informazione dalla base documentale AGCM

Per l'estrazione d'informazione TaltaC2 doveva garantire il riconoscimento automatico di una serie di informazioni quali :

- le 6 sezioni principali in cui è suddiviso ciascun provvedimento. La corretta identificazione di queste sezioni è fondamentale per poter orientare gli algoritmi di ricerca e le query solo verso quelle parti nelle quali ci si aspetta che l'informazione sia presente, ottimizzando così i tempi di risposta e aumentando il tasso di *precision* ;
- il numero identificativo del caso descritto dal provvedimento (una sigla alfanumerica) ;
- il nome e il ruolo delle aziende acquirenti/acquisite, tenendo conto del fatto che una stessa azienda può essere citata in maniera diversa all'interno del documento (tramite acronimo, dizione estesa con forma giuridica o senza forma giuridica) e associando univocamente tutte queste diverse citazioni ;
- i gruppi controllanti le aziende acquirenti/acquisite ;
- ogni altra azienda citata ;
- il mercato del prodotto ;
- il mercato geografico ;
- le soglie di fatturato delle aziende, ricondotte a tre categorie ;
- la presenza di quote di mercato.

Le informazioni così individuate sono state registrate in un apposito database che, in precedenza, era alimentato manualmente. Tuttavia, accanto a questa esigenza ben precisa, esisteva per l'Autorità il bisogno di poter effettuare ricerche ed indagini in maniera flessibile sui testi in suo possesso, svincolata dalle logiche del puro e semplice *database feeding*. Per soddisfare tale necessità è stata sviluppata - con TaltaC2 - una metodologia incentrata, in particolar modo, sull'utilizzo congiunto dell'analisi delle concordanze e della funzione di *Ricerca Entità* (il modulo di *information retrieval* di TaltaC2), per l'individuazione di concetti di interesse all'interno del testo e per la categorizzazione di documenti a partire dal loro contenuto testuale. Di seguito il dettaglio delle fasi del progetto.

3.1. Pre-trattamento

Una volta riprodotto nella sessione di lavoro, al corpus possono applicarsi procedure di *pre-trattamento*. In particolare i documenti AGCM, oltre ai comuni algoritmi di *normalizzazione* per la pulizia del testo, sono stati sottoposti a procedure semi-automatiche per la *standardizzazione delle grafie* e per la correzione degli errori ortografici che hanno consentito di risolvere il problema della diversa grafia delle forme giuridiche delle aziende (nel testo si trovavano scritte in diversi modi, per esempio 'Inc.', 'Incorporated', 'INC' oppure 'S.p.A'., 's.p.a' o 'SpA'), riconducendole ad un'unica grafia comune ('INC' e 'SPA').

Completate le procedure di normalizzazione, il testo pulito può essere letto in modalità *full text* direttamente in TaltaC2, mediante la funzione *Esplora il Corpus*. Attraverso una visualizzazione ad albero viene mostrato, per ogni documento, il testo completo, il testo di ciascuna sezione e le variabili categoriali ad esso associate, sia quelle derivanti dal caricamento iniziale del corpus, sia quelle ricavate *ex post* dall'analisi del contenuto (da piani di lavoro personalizzabili o da categorizzazioni dei documenti derivanti da query complesse).

Possibilità di filtro sulle sezioni e in funzione delle variabili categoriali aumentano le flessibilità di ricerca.

3.2. Estrazione automatica dell'informazione per la classificazione dei provvedimenti

Per il riconoscimento automatico delle informazioni sui provvedimenti AGCM, TaltaC2 si serve di liste esterne di lessicalizzazione (messe a disposizione dalla stessa Autorità) e di una serie di algoritmi basati sul riconoscimento di sequenze di testo e query composte da espressioni regolari che impiegano caratteri jolly ed operatori booleani. Le sequenze e le query, scritte con una semplice sintassi in un file di testo, sono modificabili per venire incontro a mutate esigenze dell'analisi.

Le diverse fasi di estrazione automatica dell'informazione sono state integrate in una procedura sequenziale, a cui si ha accesso da un menù di TaltaC2 personalizzato per l'Autorità, così articolata :

- identificazione dell'inizio (e della fine) delle 6 sezioni di cui si compongono i provvedimenti, a seconda del riconoscimento di determinate sequenze di testo che possono impiegare caratteri jolly ;
- *parsing* delle prime righe di ogni provvedimento. Queste infatti presentavano, secondo una struttura regolare, il numero del caso, il nome e il ruolo delle aziende (acquirenti/acquisite). Un apposito algoritmo di lettura automatica ha permesso di identificare e registrare agevolmente questi elementi ;
- identificazione e lessicalizzazione dei nomi delle imprese a partire da una lista esterna ;
- identificazione e lessicalizzazione di altre imprese non presenti nella lista. In questo secondo caso il riconoscimento avviene tramite un insieme di algoritmi (automa) che operano secondo la regola <incipit + nome azienda + forma giuridica> esemplificata nel grafo di *Figura 1* ;
- assegnazione del ruolo acquirente/acquisita alle imprese tramite confronto tra quanto letto dalla prima riga del provvedimento e gli oggetti <impresa> trovati nel testo e lessicalizzati tramite i due passi precedenti. Il confronto comprende i rispettivi acronimi, laddove esistano, solamente per le imprese alle quali sia stato assegnato un ruolo. In

questo modo, se si accede ai documenti dal modulo *Esplora il Corpus*, cliccando su una qualsiasi istanza di una delle aziende a cui TaltaC2 ha assegnato un ruolo, sono contestualmente visualizzate, tramite sottolineatura, tutte le citazioni della stessa azienda, indipendentemente dalla loro forma grafica ;

- identificazione dei gruppi industriali delle aziende con ruolo assegnato, verificando la presenza di un set di espressioni "sensibili" dopo i nomi di queste società (per esempio *"è LAG7 controllat? LAG7 da*"* o *"partecipat? LAG7 da*"*[5]). Eventuali nomi di società che seguono tali espressioni identificano una controllante. Si reitera il processo per la società identificata come controllante al fine di ottenere o una controllante della controllante o l'effettiva capogruppo finale. Le query per l'identificazione dei gruppi sono applicate alla sola sezione *Le parti* ;
- identificazione della categoria di soglie di fatturato superate dalle parti per l'obbligo di comunicazione dell'operazione. Le query per l'identificazione della categoria sono applicate alla sola sezione *Qualificazione dell'operazione* ;
- identificazione e lessicalizzazione dei mercati del prodotto da lista esterna ;
- identificazione e lessicalizzazione dei mercati del prodotto non presenti nella lista. L'individuazione avviene, anche in questo caso, attraverso una o più espressioni "sensibili" che precedono il nome del mercato, per esempio (*"part? LAG8 attiv? n* settor? d*" OR "part? LAG8 attiv? n* mercat? d*"*). Anche queste espressioni sono (come le precedenti) editabili ;
- identificazione del mercato geografico tramite query di ricerca, per esempio *"dimension? geografic* LAG15 local*"* per un mercato locale. Le query per l'identificazione dei mercati (del prodotto e geografico) sono applicate alla sola sezione *Valutazione della concentrazione* ;
- individuazione della presenza della quota di mercato.

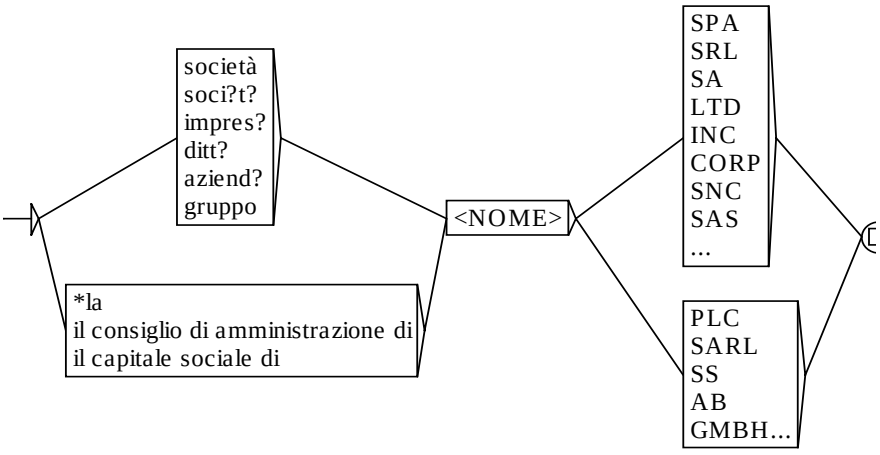

Figura 1 – Grafo di una regola di riconoscimento del nome di impresa utilizzato dall'automa

[5] Nella sintassi di Taltac2, l'operatore LAG sta ad indicare la presenza eventuale di un numero di parole che va da zero al numero che segue l'operatore. L'espressione "partecipat? LAG7 da*" sta quindi ad indicare "partecipata dalla" oppure "partecipata con quota di maggioranza dalle".

4. Integrazione fra strumenti e tecniche di Text Mining e Data Mining

Una volta identificate dall'Autorità le tematiche d'interesse per la formulazione del modello, sono state messe a punto query complesse per catturare e categorizzare quei documenti in cui si riscontravano o meno le sequenze di testo identificate dalle query.

4.1 Integrazione fra strumenti di Text Mining e tecniche di Data Mining

L'*analisi delle concordanze semplici e complesse* è stata di ausilio alla compilazioni delle query suddette (e di quelle per la procedura automatica di cui al paragrafo precedente, in particolare per l'identificazione del gruppo di appartenenza delle imprese coinvolte nel provvedimento, per l'individuazione del mercato del prodotto e del mercato geografico di riferimento). Tali concordanze sono effettuabili in TaltaC2 rispettivamente su singole parole o sequenze, su insiemi di unità lessicali (mediante l'uso di caratteri *jolly*) o in base a tag grammaticali (ottenuti dalla procedura di *tagging*) o semantici (ottenuti sottoponendo, tramite un'apposita funzione di TaltaC2, dizionari tematici personalizzabili che consentono di associare un dato concetto a tutte le forme del corpus in analisi presenti nel dizionario).

L'analisi delle concordanze ha permesso di identificare ed isolare tutti i contesti (le frasi) riferiti ad una serie di termini *pivot* legati ad un aspetto di studio. Ad esempio, uno degli aspetti che l'Autorità aveva interesse ad indagare era la presenza/assenza di barriere all'entrata in un mercato. Pertanto, le concordanze della parola "barriere" hanno consentito di individuare, per ogni provvedimento, i riferimenti a questo specifico concetto. In tal modo i contesti più frequenti ed effettivamente associabili al concetto di interesse sono stati utilizzati per costruire le query.

Con riferimento al concetto <assenza di barriere all'entrata> la query ha assunto la seguente forma : *"non LAG4 present* LAG3 barriere" OR "inesisten* LAG2 barriere"*, permettendo così di individuare alcuni quasi-segmenti altamente selettivi quali *"non risultano essere presenti particolari barriere"*, *"non presenta barriere"* o *"inesistenza di rilevanti barriere"*. Le query sono state applicate nel modulo *Ricerca Entità* di TaltaC2 che, oltre ad effettuare il recupero dei documenti, permette di creare e valorizzare nuove variabili mediante le quali operare categorizzazioni dei provvedimenti sulla base del loro contenuto testuale.Come detto, infatti, nel caso AGCM il *retrieving* è finalizzato proprio alla creazione e all'alimentazione di una nuova variabile categoriale *ex post* in grado di identificare o qualificare con un qualsiasi valore i documenti recuperati per mezzo della query. Poiché le informazioni qualitative d'interesse per l'Autorità si concentrano all'interno di una specifica sezione dei documenti, TaltaC2 ha permesso di limitare la ricerca solo in questa parte dei documenti.

Eseguendo le medesime operazioni su tutti i concetti di interesse si è ottenuta una matrice che associa a ciascun documento, identificato da una chiave univoca e dalle variabili categoriali *ex ante* (metadati), un insieme di variabili *ex post* indicanti la presenza/assenza o una particolare qualifica del concetto oggetto di analisi.

Questa matrice risultante, integrata con indici quantitativi macroeconomici, è stata adoperata come base di dati per la costruzione dei modelli *logit*.

4.2 I risultati del modello logit

A seguito dell'istituzione dell'Unione Europea, la valutazione dei casi di concentrazione tra imprese da parte delle Autorità antitrust europee segue linee guida che riflettono i principi

comunitari e, più in generale, quelli consolidatisi nella disciplina antitrust e nella prassi applicativa delle leggi in materia di controllo delle concentrazioni

Obiettivo del lavoro era quello di verificare la coerenza delle decisioni adottate dall'Autorità, esaminando i provvedimenti conclusivi dei casi di concentrazione trattati nel periodo 1995-2003, al fine di rendere più trasparenti i criteri decisionali seguiti.

La specificazione del modello ha condotto all'individuazione dei fattori che meglio spiegano l'approccio decisionale dell'Autorità, evidenziando il ruolo di indicatori quali *quote di mercato detenute dalle parti*, *grado di concentrazione dell'offerta* e suo incremento, a seguito dell'operazione di concentrazione. Accanto a questi fattori di natura quantitativa, è emerso il ruolo dei fattori di natura qualitativa come le *barriere all'entrata di nuovi operatori*, l'integrazione dei processi produttivi (*effetti verticali*), l'ampliamento della gamma produttiva (*effetti orizzontali*) e il *potere negoziale dal lato della domanda*.

La simulazione di diversi scenari condotte con il modello evidenzia che (Tavola 1) :

- in assenza di barriere all'entrata è altamente improbabile un intervento dell'Autorità ;
- in presenza di barriere all'entrata, la soglia di quota di mercato detenuta dalle parti oltre la quale comincia a diventare probabile un intervento correttivo è pari a circa il 40%. La probabilità di intervento aumenta con l'aumentare di altri fattori (sovrapposizione dei mercati, misurato dalla variazione dell'indice di Herfindahl-Hirschman (HHI)[6], e in presenza di effetti verticali o orizzontali) ;
- pur in presenza di barriere all'entrata, l'effetto restrittivo della concentrazione dal lato dell'offerta si bilancia con la presenza di un potere negoziale dal lato della domanda, ad eccezione dei casi in cui la presenza di barriere è rafforzata dall'esistenza di effetti verticali o orizzontali.

		Variazione di HHI				
		0-99	**100-199**	**200-499**	**500-999**	**≥ 1000**
Quota di mercato	**0-19%**	3.63				
	20-39%	19.49	25.55	29.95		
	40-59%	38.09	46.92	52.48	57.59	59.58
	60-79%	53.25	62.34	67.43	71.83	73.48
	80-100%	64.56	72.76	76.96	80.46	81.74

Tabella 1 – Probabilità di intervento dell'Autorità con presenza di barriere all'entrata, potere negoziale della domanda ed effetti verticali/orizzontali (Le aree ombreggiate evidenziano gli intervalli di quote e di variazione di HHI a cui è associata una probabilità di intervento dell'Autorità superiore al 50%. Le aree nere si riferiscono ad eventi non osservati nella realtà esaminata e, come tali, non rilevanti per i risultati delle simulazioni).

[6] L'indice di Herfindahl-Hirschman è un indice di concentrazione comunemente accettato risultato della somma dei quadrati delle quote di mercato individuali di tutti gli operatori presenti in un dato mercato.

Références

Allison P.D. (1999). *Logistic Regression Using the SAS System. Theory and Application.* SAS Institute Inc.

Baiocchi F., Bolasco S., Canzonetti A., Capo F. M., della Ratta-Rinaldi F., Morrone A. (2005). *Guida di TaltaC 2.0.* Roma

Bolasco S. (2002). Integrazione statistico-linguistica nell'analisi del contenuto. In Mazzara, B. (editor), *Metodi qualitativi in Psicologia Sociale. Prospettive teoriche e strumenti operativi.* Carocci Ed., Roma : 329-342.

Bolasco S. (2000). TALTAC : un environnement pour l'exploitation de ressources statistiques et linguistiques dans l'analyse textuelle. Un exemple d'application au discours politique. *JADT2000, EPFL*, Lausanne 9-11 marzo, tome 2 : 342-353.

Bolasco S., Canzonetti A., Capo F. M. (2005). *Text Mining : uno strumento strategico per imprese ed istituzioni.* CISU, Roma.

Bolasco S., Morrone A., Baiocchi F. (1999). A Paradigmatic Path for Statistical Content Analysis Using an Integrated Package of Textual Data Treatment. In Vichi, M., Opitz, O. (eds.), *Classification and Data Analysis. Theory and Application.* Springer-Verlag, Heidelberg : 237-246.

La Noce M., Allegra E., Ruocco V., Capo F. M., (2005). Merger Control in Italy 1995-2003 : A Statistical Study of the Enforcement Practice by Mining the Text of Authority Resolutions. Social Science Electronic Publishing.

http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID831664_code26713.pdf?abstractid=771986&mirid=1

Maddala G.S. (2001). *Introductions to Econometrics*. John Wiley & Sons Ltd.

Poibeau T. (2003). *Extraction Automatique d'Information : du texte brut au web semantique*. Hermes-Lavoisier, Paris.