

TECNICHE DI WEB SCRAPING E TEXT MINING PER LA DEFINIZIONE DEI SETTORI INDUSTRIALI: UN'APPLICAZIONE A PIEMONTE E VENETO

Gianluca Toschi¹, Giampaolo Vitali², Duccio Schiavon³, Diego Chinellato⁴, Nicola Ianuale⁵

SOMMARIO

La definizione di settore industriale ha una grossa rilevanza per gli economisti regionali in quanto le analisi sulle caratteristiche economiche di un territorio, e in particolare sulle condizioni di tipo strutturale, si basano spesso sulla comparazione di sistemi economici suddivisi in settori industriali. La letteratura economica identifica in maniera chiara il concetto di settore ma nel momento in cui si affronta la questione da un punto di vista operativo le scelte appaiono limitate dalla carenza di informazioni, dalla lentezza di aggiornamento delle stesse e dalla rigidità delle tassonomie impiegate. Nel presente lavoro viene proposto un metodo per arricchire le informazioni che possono essere utilizzate ai fini dell'attribuzione di un'impresa ad un settore industriale basato sull'uso di *Qiba*, un *software* che permette la raccolta e successiva analisi delle informazioni attraverso tecniche di *web scraping*, *entity extraction*, *text mining*, *text analysis* e *document clustering*. La metodologia viene applicata a titolo sperimentale all'analisi del settore della meccanica (fabbricazione di macchine utensili) di Piemonte e Veneto.

¹ Fondazione Nord Est e IRCRES-CNR, via Torino 151/C, 30172, Venezia gianluca.toschi@fnorddest.it (corresponding author).

² IRCRES-CNR, Via Real Collegio, 30, 10024 Moncalieri TO

³ Quantitas Srl, Parco Scientifico Tecnologico VegaPark, Marghera (VE)

⁴ Quantitas Srl, Parco Scientifico Tecnologico VegaPark, Marghera (VE)

⁵ Quantitas Srl, Parco Scientifico Tecnologico VegaPark, Marghera (VE).

Introduzione

In un saggio del 1998 P.A. Geroski sottolineò il difficile rapporto che esiste tra gli economisti industriali e l'attività di definizione dei confini del mercato⁶.

“L'identificazione dei confini del mercato è considerata dagli economisti industriali come un lavoro degno ma noioso [...] In effetti, per una professione che pretende di essere interessata a ciò che accade nei mercati, dedichiamo sorprendentemente poche risorse alla definizione di quella che potrebbe essere naturalmente considerata la nostra unità di analisi.” Geroski (1998)

Ciò è dovuto, sempre secondo l'economista, a due fattori: il primo è che risulta decisamente più interessante concentrarsi sulle cose che accadono in un mercato piuttosto che sul luogo in cui queste accadono (il mercato stesso), il secondo è che molti considerano l'attività di definizione del mercato come un tentativo di organizzare il modo in cui pensiamo l'attività economica artificiale e arbitrario e quindi poco interessante da analizzare. Qualunque sia la motivazione, il risultato è che gli economisti industriali hanno rivolto i propri studi in questo ambito da una parte alle proprie necessità e dall'altra ai bisogni dei *policy makers*, in particolare a quelli delle autorità *antitrust*. Le definizioni di mercato più utilizzate fanno infatti riferimento a “lo spazio nel quale vale la legge del prezzo unico” (l'esistenza e l'ampiezza di un mercato può quindi essere rilevata attraverso l'analisi dell'elasticità incrociata della domanda rispetto al prezzo) e al “mercato rilevante” (di prodotto e geografico) nelle applicazioni che riguardano la politica della concorrenza⁷. Due definizioni che al di fuori dell'ambito nelle quali sono state elaborate risultano scarsamente applicabili. A questo si deve aggiungere che i tentativi di definire, da un punto di vista operativo, i mercati e i settori industriali risultano spesso limitati dalla carenza di informazioni, dalla lentezza di aggiornamento delle stesse e dalla rigidità delle tassonomie proposte che mal si adattano sia alla velocità di cambiamento che alle diverse necessità di chi le deve utilizzare. Partendo da queste considerazioni è lo stesso Geroski a suggerire di lavorare con maggiore creatività nell'attività di definizione dei mercati, tenendo in considerazione anche gli utilizzi alternativi rispetto a quelli più vicini alla ricerca economica.

“I confini del mercato sono linee immaginarie che imponiamo alla realtà e le disegniamo per isolare alcuni tipi di attività dalle altre al fine di dare un senso e pensare in modo creativo a ciò che osserviamo. Il punto in cui tracciamo i confini dipende dal motivo per cui siamo interessati a farlo”. Geroski (1998)

Sulla scia di tali considerazioni il presente lavoro si propone di identificare una strategia di analisi che permetta di arricchire le informazioni che possono essere utilizzate per attribuire un'impresa ad un settore in modo flessibile rispetto ai diversi scopi di ricerca. La metodologia proposta utilizza la raccolta sul web di informazioni (*web scraping*) e la loro successiva analisi con tecniche di *Text Mining* e *Text Analysis*. La crescente mole di informazioni che le imprese pubblicano sui siti Web li rende una preziosa fonte di indagine per i ricercatori. Pur nella consapevolezza che i siti Web sono “auto-rapporti”, le informazioni disponibili

⁶ Nel presente lavoro i termini mercato e settore verranno utilizzati in modo intercambiabile. In passato i due termini sono stati utilizzati per distinguere le definizioni basate sugli utilizzatori, sui luoghi e a altri fattori legati al lato della domanda (mercato) da quelle basate su tecnologie e altri fattori dal lato dell'offerta (settori) Kay (1990).

⁷ Per un'analisi dettagliata dell'evoluzione del concetto di settore nella letteratura economica si può far riferimento a Barbarito (1999).

hanno numerosi vantaggi: sono prontamente e pubblicamente disponibili, relativamente poco costose da ottenere. (Gök, A et. Al, 2015) e la loro raccolta si allinea tra le tecniche non intrusive di indagine (Arora, S. K et. al., 2016), un aspetto rilevante in un periodo in cui i tassi di risposta ai questionari somministrati a scopo di ricerca sono in costante riduzione. Il risultato di tutto ciò è che, negli ultimi anni, l'uso di informazioni provenienti dal web ha trovato un crescente utilizzo nell'ambito della ricerca economica.

La metodologia viene applicata a titolo sperimentale all'analisi del settore della meccanica di Piemonte e Veneto utilizzando come dati di partenza le informazioni estratte dai siti web di un gruppo di imprese impegnate nella produzione di macchine utensili.

Nel primo paragrafo si fa il punto sui tentativi di definire da un punto di vista operativo i settori industriali, nel secondo si passa in rassegna una serie di esperienze che riguardano l'uso di tecniche di web scraping e text mining nell'analisi economica, nel terzo vengono descritte le tecniche di web scraping, text mining e document clustering utilizzate, nel quarto si discutono i principali risultati dell'analisi condotta su un gruppo di imprese metalmeccaniche del Piemonte e del Veneto.

1. Settore industriale: come rendere operativa una definizione chiara

Nella letteratura economica il concetto di settore industriale è stato declinato in diversi modi. Per un'analisi dell'evoluzione storica di tale concetto, che non è obiettivo di questo lavoro, si può far riferimento a Barbarito (1999). Allo stesso autore va il merito di aver individuato una definizione che, da un punto di vista teorico, identifica in maniera chiara il settore industriale come "...la porzione del sistema economico nella quale sono aggregate imprese simili, che producono beni simili e sono tra loro interdipendenti" (Barbarito, 1999). Una definizione che concentra l'attenzione su fattori legati al lato della domanda, dell'offerta ma anche all'interdipendenza tra imprese e quindi ha il merito di far sintesi di diversi approcci. La definizione di Barbarito tiene, infatti, in considerazione tre distinti criteri: la similitudine tra imprese, secondo la quale appartengono allo stesso settore imprese che utilizzano processi simili anche producendo output diversi; la similitudine tra prodotti, dove simili sono i prodotti che soddisfano uno stesso bisogno del consumatore; il requisito dell'interdipendenza tra imprese secondo il quale l'appartenenza ad un settore permette la rivalità e quindi anche la capacità di sottrarre domanda ad altri attori. Fanno parte dello stesso settore, quindi, le imprese tra di loro interdipendenti.

A fronte di un'elaborazione teorica ricca, nel momento in cui si affronta la questione della definizione di settore industriale da un punto di vista operativo le scelte appaiono, invece, limitate dalla carenza di informazioni. Soffrono di questo problema le tecniche utilizzate in ambito antitrust che richiederebbero la rilevazione puntuale di informazioni sui prezzi, ad esempio, e che quindi non si prestano a operazioni massive di definizione dei confini dei mercati, limiti che scontano molti dei tentativi che si propongono di lavorare sul concetto di interdipendenza tra imprese.

Sul fronte della similitudine (tra imprese e tra prodotti) va sottolineato il lavoro svolto nel corso degli anni dalle agenzie statistiche sia a livello nazionale che internazionale che si sono impegnate intensamente

nell'attività di classificazione dei settori industriali, dando vita a un sistema integrato delle classificazioni che oggi permette la comparabilità delle statistiche prodotte in diversi campi. Rientrano in questo sistema integrato: la Classificazione internazionale per industrie di tutti i rami dell'attività economica delle Nazioni unite (Isic), la Classificazione centrale dei prodotti delle Nazioni unite (Cpc), il Sistema armonizzato di designazione e di codificazione delle merci (Hs) gestito dall'Organizzazione mondiale delle dogane, la Classificazione europea dei prodotti (Cpa), la classificazione dei prodotti usata per le statistiche sulla produzione industriale nell'Ue (Prodcom), la nomenclatura combinata, ossia la classificazione europea delle merci utilizzata per le statistiche sul commercio estero (Nc). Molte di queste classificazioni sono caratterizzate, tra le altre cose, dall'utilizzo di categorie mutuamente esclusive (ogni elemento deve essere classificato in una sola categoria) e dall'uso di principi metodologici che consentono un'allocazione coerente degli elementi nelle varie categorie della classificazione. L'Ateco 2007, versione nazionale della classificazione (Nace Rev.2) definita in ambito europeo che, a sua volta, deriva da quella definita a livello Onu (Isic Rev. 4) presenta le varie attività economiche raggruppate in sezioni, divisioni, gruppi, classi, categorie e sottocategorie. Tale classificazione utilizza un approccio pragmatico, che considera, nella classificazione, sia la similitudine tra imprese (nei processi produttivi utilizzati) che quella tra prodotti.

“I criteri sono applicati diversamente a seconda del livello della classificazione: i criteri per i livelli più dettagliati prendono in considerazione le similitudini nel processo produttivo attuale, mentre tale aspetto risulta poco rilevante ai livelli più aggregati della classificazione.” Istat (2009).

Uno tra i limiti di questa tipologia di classificazione si evidenzia nel momento in cui ci si trova di fronte all'attribuzione di un codice Ateco a un'impresa che, a causa di un alto livello di integrazione verticale o orizzontale, svolge contemporaneamente attività che ricadrebbero in più categorie Ateco. In casi come questo viene applicato il metodo top-down: il processo inizia identificando la posizione più rilevante al livello più alto e scende attraversando i vari livelli della classificazione. La rilevanza è legata al valore aggiunto prodotto da ogni attività. Il metodo top-down può portare ad attribuire ad un'impresa l'appartenenza a una classe che non ha la percentuale di valore aggiunto più alta, un risultato che nasconde agli utilizzatori informazioni importanti sul range di attività svolte ma anche sul modello di integrazione (verticale o orizzontale) di un'azienda. A questo si deve aggiungere che la lentezza di aggiornamento delle classificazioni e la loro rigidità mal si adattano sia alla velocità di cambiamento che alle diverse necessità di chi le deve utilizzare. Queste e altre motivazioni, per una rassegna delle critiche che le tassonomie ufficiali hanno attirato si veda Lind (2005), rendono le classificazioni ufficiali uno strumento importante ma con alcuni pesanti limiti quando utilizzate nello studio dei settori industriali.

2. Web crawling, web scraping e text analysis nell'analisi economica

L'analisi del contenuto testuale è un approccio innovativo nell'ambito delle scienze sociali che ha trovato, negli ultimi anni, applicazioni interessanti anche nella ricerca economica, si veda, a titolo di esempio, il lavoro di Cortelazzo & Gambarotto (2013) che utilizzano i discorsi dei Presidenti di Confindustria dal 1945 al 2011 per indagare il ruolo giocato dalla Confederazione Italiana degli Industriali nel processo di cambiamento socio-economico del nostro paese dal dopoguerra fino alla prima decade degli anni 2000. Oggi il web si aggiunge come fonte di contenuto testuale mettendo a disposizione volumi sempre più importanti di informazioni che sono ormai diventate una preziosa fonte per le ricerche. Prima di passare in rassegna alcune delle applicazioni che utilizzano dati raccolti dai siti web è necessario sottolineare limiti e vantaggi di tali metodologie.

I vantaggi principali nell'uso di informazioni raccolte dal web sono legati al fatto che:

- i siti web aziendali funzionano da vere e proprie vetrine dei propri prodotti e servizi. E quindi vi è (o vi dovrebbe essere) l'interesse da parte dell'azienda a spiegare quanto più dettagliatamente la natura di questi e più in generale di fornire una presentazione accurata dell'impresa. Al-Hassan et. al. (2013), tenendo in considerazione i processi di internazionalizzazione, sottolineano anche il fatto che i contenuti di un sito Web aziendale dovrebbero essere forniti per soddisfare le esigenze dei clienti di tutto il mondo la cui ricerca dovrebbe essere facile da eseguire e facilitata da informazioni precise e facili da comprendere. I dati ricavabili dai siti web aziendali dovrebbero, quindi, permettere una raccolta ricca e dettagliata di informazioni sull'impresa stessa, sui suoi prodotti e sulle tecnologie adottate,
- le informazioni provenienti dai siti web sono pubblicamente disponibili e aggiornate dato l'interesse da parte dell'impresa a rendere pubbliche le informazioni potenzialmente interessanti per il proprio mercato (nuovi prodotti, nuovi mercati, nuove tecnologie adottate...),
- sono relativamente poco costose da ottenere. (Gök, A et. Al, 2015),
- la raccolta di informazioni dal web si allinea tra le tecniche non intrusive di indagine (Arora, S. K et. al., 2016), un aspetto rilevante in un periodo in cui i tassi di risposta ai questionari somministrati a scopo di ricerca sono in costante riduzione. Oggi Internet può essere considerato come una fonte di dati da sfruttare in sostituzione o in combinazione con i dati raccolti mediante gli strumenti tradizionali di un'indagine campionaria (Barcaroli et. al., 2014 e ten Bosch et. al., 2018).

I principali limiti si possono individuare nel fatto che:

- i dati raccolti dai siti web aziendali sono, di fatto, delle “auto-dichiarazioni”. Come evidenziato da Arora et. al (2013), a margine di un’analisi condotta su piccole e medie imprese *technology-based*, le aziende possono decidere di modificare nel tempo le informazioni pubblicate: potrebbero quindi esserci momenti in cui le aziende sono interessate a pubblicare un gran numero di informazioni che possono riguardare temi quali lo sviluppo della tecnologia e le strategie di business (quando ad esempio cercano visibilità verso potenziali clienti o finanziatori, o a favore della tecnologia e/o dei prodotti che hanno sviluppato); al contrario, se un’azienda desidera essere più cauta riguardo alla diffusione di informazioni che riguardano la propria capacità tecnologica, le linee di ricerca o le partnership, il sito Web potrebbe “oscurarsi” e fornire poche o nulle informazioni rispetto a questi temi,
- le informazioni che si ricavano dal web non sono standardizzate e hanno quindi bisogno di un accurata attività di trattamento per essere utilizzate (Arora et. al., 2016 e Kinne, & Resch, 2018).

Nel campo della ricerca economica sono diversi gli utilizzi che sono stati fatti delle tecniche di raccolta e analisi dei dati dal web: nell’ambito del *numeric data mining* a partire dal 2014 l’Office for National Statistics (ONS) ha cominciato a integrare i dati tradizionalmente utilizzati per il calcolo dell’indice dei prezzi al consumo (CPI) con informazioni sui prezzi provenienti dal web. I vantaggi dell’utilizzo di tali tecniche appaiono evidenti: in un periodo di 13 mesi sono stati raccolti giornalmente circa 6.500 quotazioni di prezzo per 35 diversi articoli (Breton et. al., 2015). Sul fronte del *textual data mining*, un filone interessante di applicazioni è rappresentato dalle ricerche sulle attività di innovazione e R&D che usano le informazioni provenienti dai siti web delle imprese per integrare quelle che si ricavano da fonti ufficiali. Arora et. al (2013) hanno usato metodi di estrazione delle informazioni dal web per sondare le strategie di piccole e medie imprese *technology-based* nella commercializzazione di tecnologie emergenti. Gök, A et. Al, (2015) utilizzano il web mining per esplorare le attività di ricerca e sviluppo di 296 piccole e medie imprese con sede nel Regno Unito. Più recentemente Kinne, & Resch (2018) propongono una metodologia per il web scraping di siti aziendali che utilizza metodi di apprendimento automatico basato sull’uso di reti neurali artificiali, finalizzata alla creazione di dati spaziali per l’analisi delle attività di innovazione delle imprese. L’utilizzo di strumenti come la *Wayback Machine* ha permesso di estendere le analisi basate sul web data mining offrendo una preziosa fonte di dati per analizzare le informazioni web nel tempo. Arora, S. K., et. al (2016).

L’analisi più simile a quella presentate in questo lavoro è quella di Al-Hassan et. al. (2013) nella quale si analizza la relazione tra i contenuti del sito web di un’impresa e la sua classificazione

NAICS. Gli autori partono dall'ipotesi che il materiale presente nel sito di un'organizzazione riveli l'appartenenza ad un settore industriale così come dichiarata dalle statistiche ufficiali. L'analisi viene condotta estraendo informazioni dai siti delle 500 imprese presenti nella classifica *US Fortune 500* che riguardano esclusivamente i *legal attachments statements (Privacy and Term of Use Statements)*. Pur condividendo l'ambito di analisi (la relazione tra i contenuti del sito web di un'impresa e l'identificazione del settore di appartenenza) il lavoro di Al-Hassan et. al. (2013) si differenzia rispetto al nostro per gli obiettivi e per:

- il campione di partenza - Al-Hassan et. al. (2013) utilizzano un database di imprese che appartengono a settori diversi, in questo lavoro viene utilizzato un database di imprese che condividono la medesima classificazione ufficiale,
- i dati estratti dai siti web - *legal statements* nel lavoro sulle imprese *US Fortune 500*, informazioni su prodotti e processi produttivi in questo lavoro,
- le tecniche di text analysis e successiva clusterizzazione.

I risultati del lavoro di Al-Hassan et. al. (2013) non confermano l'ipotesi di partenza.

3. La metodologia utilizzata – Web crawling, web scraping e text analysis

La metodologia che è stata seguita si articola in sette distinte fasi ed è basata in gran parte sull'uso di Qiba (Quantitas Intelligent Business Analyzer), un programma che supporta le fasi di *web crawling*, *web scraping*, pre-processing e pre-analisi del testo sviluppato in Python da Quantitas Srl, utilizzando librerie Open Source.

1. La prima fase del processo riguarda l'individuazione dell'insieme di imprese delle quali analizzare il sito web. Il set minimo di informazioni per alimentare la fase di Web Crawling di Qiba è composto da ragione sociale e partita IVA dell'impresa.
2. La seconda fase del processo è quella di Web Crawling. Un Web Crawler è un sistema automatizzato che "scandaglia" la rete alla ricerca di siti e/o pagine web, generalmente a scopo di indicizzazione (un'attività tipicamente svolta dai motori di ricerca); ai fini del presente lavoro il web crawling serve a determinare se una data azienda ha un proprio sito web. Immettendo come input una serie di identificativi di aziende {Partita IVA, Ragione Sociale} Qiba effettua una serie di ricerche sui principali motori e assegna ai risultati un punteggio in base ad un algoritmo proprietario. Viene quindi selezionato, per ogni identificativo, il sito che presenta il punteggio più alto e che si distacca sufficientemente da quello degli altri risultati. L'obiettivo di questa fase di analisi è di individuare con precisione il sito dell'impresa, scartando eventuali pagine che

riportano ragione sociale e partita IVA ma non appartengono al sito web dell'impresa (si pensi alle pagine relative alla partecipazione a fiere o ad altri cataloghi online che non fanno parte del sito aziendale).

3. La fase successiva effettua il *Web Scraping*, ovvero un processo di estrazione automatica di informazioni dal WWW. Tra le diverse forme di web scraping (Miner et al., 2012) viene utilizzato il web *content mining*. Partendo dalla homepage di ogni sito rilevato si va alla ricerca, all'interno del sito web aziendale, delle pagine che possano caratterizzare i prodotti e i servizi delle aziende considerate. In questa fase si usa una funzione euristica, ovvero si va alla ricerca di link interni a pagine con nomi che si ritengono comunemente utilizzati per descrivere l'azienda (es. "prodotti", "servizi", "cosa facciamo", ...). Qiba è stato sviluppato in modo da permettere una personalizzazione di questa fase di ricerca al fine di aumentarne la flessibilità rispetto alle due domande che guidano l'attività di *web scraping* ossia che tipo di informazioni presenti nelle pagine Web devono essere sottoposte a scansione? Quali pagine sul sito di un'organizzazione devono essere incluse nella raccolta? (Arora et. al., 2016).
4. Ottenute le pagine potenzialmente "interessanti", viene effettuato un *pre-processing* del testo: eliminazione di punteggiatura, riduzione del testo in minuscolo e rimozione di tutte le parole considerate comuni nel linguaggio italiano (preposizioni, congiunzioni, nomi propri, articoli, pronomi...) al fine di identificare le parole di contenuto.
5. In considerazione del fatto che l'osservazione delle parole che assumono un significato diverso se considerate congiuntamente alle parole adiacenti contribuisce ad aumentare la portata informativa delle stesse (Cemin e Tuzzi, 2013), è stata sviluppata un'interfaccia grafica che dà la possibilità di scegliere delle parole di partenza tra le keywords già estratte (es. "lastre") per costruire sequenze (e.g. "lastre di acciaio", "lastre di alluminio", ...) interessanti per l'analisi. In questa fase è necessario l'intervento umano.
6. L'ultima fase trasforma i contenuti estratti dai siti Web (che sono essenzialmente dati non strutturati) in variabili utilizzabili dalle scienze sociali. Il problema, in questa fase, diventa come "filtrare le informazioni dal rumore" (Arora et. al., 2016 e Kinne, & Resch, 2018). Mutuando le tecniche maggiormente utilizzate nell'ambito dell'analisi testuale Qiba produce una serie di output che possono fornire le prime indicazioni sui contenuti più presenti nel corpus, come, ad esempio, elenchi che mettono in ordine le parole a partire da quelle più frequenti fino ad arrivare agli hapax. I termini più frequenti non sono, però, necessariamente i più informativi. Al contrario, i termini che appaiono frequentemente in un piccolo numero di documenti ma raramente negli altri documenti tendono ad essere più pertinenti e specifici per quel particolare gruppo di documenti, e quindi più utili per trovare documenti (siti web) simili (Huang, A., 2008). Per passare dalle parole più frequenti a quelle più rilevanti ai fini dell'analisi del contenuto, Qiba permette di calcolare un indice della famiglia TFIDF (*term frequency – inverse document*

frequency) che misura la forza discriminante di una parola. Dato un corpus composto da M testi il valore $TFIDF_i$ di una parola i viene calcolato come:

$$TFIDF_i = f_i \log \frac{M}{m_i}$$

L'indice è calcolato considerando la frequenza f_i della parola i (tanto più una parola è frequente tanto più è rilevante) per il logaritmo del rapporto tra il numero M di testi presenti nel corpus e il numero m_i di documenti che contengono la parola i (risulta maggiormente rilevante la parola che è presente in certi testi e assente in altri).

7. I dati ricavati da Qiba possono essere utilizzati come input per le diverse tipologie di esplorazione del dato utilizzabili nell'ambito dell'analisi del testo. Ai fini del presente lavoro sono state utilizzate strategie di *document clustering* basate sulla modellazione del testo come “*bag-of-words*” (Tuzzi, A. (2010). Fra i diversi algoritmi di clustering considerando che quelli partizionali sembrano più adatti alla gestione di set di dati di documenti di grandi dimensioni rispetto a quelli gerarchici (Huang A., 2008) si è scelto di utilizzare il *k-means*, usando come misura di distanza la distanza euclidea.

4. Qiba: un'applicazione all'analisi del settore della meccanica in Veneto e Piemonte

La definizione di settore industriale ha una grossa rilevanza per gli economisti regionali in quanto le analisi sulle caratteristiche economiche di un territorio, e in particolare sulle condizioni di tipo strutturale, si basano spesso sulla comparazione di sistemi economici suddivisi in settori industriali. Partendo da questa considerazione la scelta del settore produttivo da utilizzare per testare la metodologia proposta è stata fatta calcolando un indice di specializzazione relativa per le regioni di Piemonte e Veneto utilizzando i dati relativi agli addetti per settore (classificato con codice Ateco) ricavati dal censimento 2011,

$$IS_{REG} = \frac{A_i^{REG}}{A^{REG}} / \frac{A_i^{ITA}}{A^{ITA}}$$

dove A_i^{REG} e A_i^{ITA} rappresentano il numero di addetti impiegati nel settore i -esimo rispettivamente in una regione (REG) ed in Italia (ITA), mentre A^{REG} e A^{ITA} il numero complessivo di addetti nella regione ed in Italia. Tra i diversi settori in cui Piemonte e Veneto evidenziano una specializzazione relativa rispetto all'Italia (i valori dell'indice di specializzazione sono per le due regioni maggiori di 1 evidenziando, quindi, una specializzazione in quel particolare settore rispetto al contesto nazionale) è stato scelto quello della “fabbricazione di altre macchine utensili”, con codice Ateco 28.49.09. Una scelta su cui ha pesato la ricca articolazione di cui gode il comparto e che quindi ben si presta a operazioni di “riclassificazione”.

1. Una volta individuato il settore sono state estratte le informazioni relative a Ragione Sociale e Partita IVA per le società di capitale presenti nella banca dati Aida BvD. Sono state individuate 140 imprese, di cui 97 in Veneto e 43 in Piemonte.
2. Il crawler, partendo dall'elenco di 140 aziende ha rilevato 92 siti; di questi, 2 non è stato possibile analizzarli poiché veniva impedito l'accesso programmatico. Il risultato della fase di *crawling* appare sostanzialmente in linea con le statistiche di fonte Eurostat sulla diffusione di siti web tra le imprese in Italia. Nelle imprese con 10 e più addetti solamente il 71% ha un sito web (Tabella 1).
3. Considerando i fini dell'analisi la fase di *web scraping* è stata indirizzata alla ricerca di link interni a pagine con nomi che si ritengono comunemente utilizzati per descrivere l'azienda (es. "prodotti", "servizi", "cosa facciamo", ...). Nove aziende sono state scartate perché non presentavano sezioni del sito finalizzate alla presentazione di prodotti e servizi. In questo caso le statistiche di fonte Eurostat non sono perfettamente sovrapponibili alla ricerca svolta dato che considerando non solo la presenza di un sito web con la descrizione di prodotti o servizi, ma anche di un listino prezzi. In ogni caso la caduta della percentuale di imprese con tali informazioni è sensibile.

Tabella 1 – Imprese con sito web per paese e classe dimensionale, 2018 (val. %).

All enterprises, without financial sector

	10 persons employed or more	Small enterprises (10-49 persons employed),	Medium enterprises (50-249 persons employed)	Large enterprises (250 persons employed or more)
European Union - 28 countries	77	74	89	94
Germany	87	86	94	97
Spain	76	73	87	96
France	69	66	87	95
Italy	71	70	82	90

Fonte: ns. elaborazioni su dati Eurostat – [isoc_ciweb]

Tabella 2: Imprese in cui il sito web ha fornito una descrizione di prodotti o servizi, listini prezzi per paese, 2018 (val. %).

	All enterprises (10 persons employed or more)	Small enterprises (10-49 persons employed)	Medium enterprises (50-249 persons employed)	Large enterprises (250 persons employed or more)
European Union - 28 countries	56	54	67	72
Germany	74	72	80	85
Spain	37	35	45	52
France	58	55	71	74
Italy	32	31	41	46

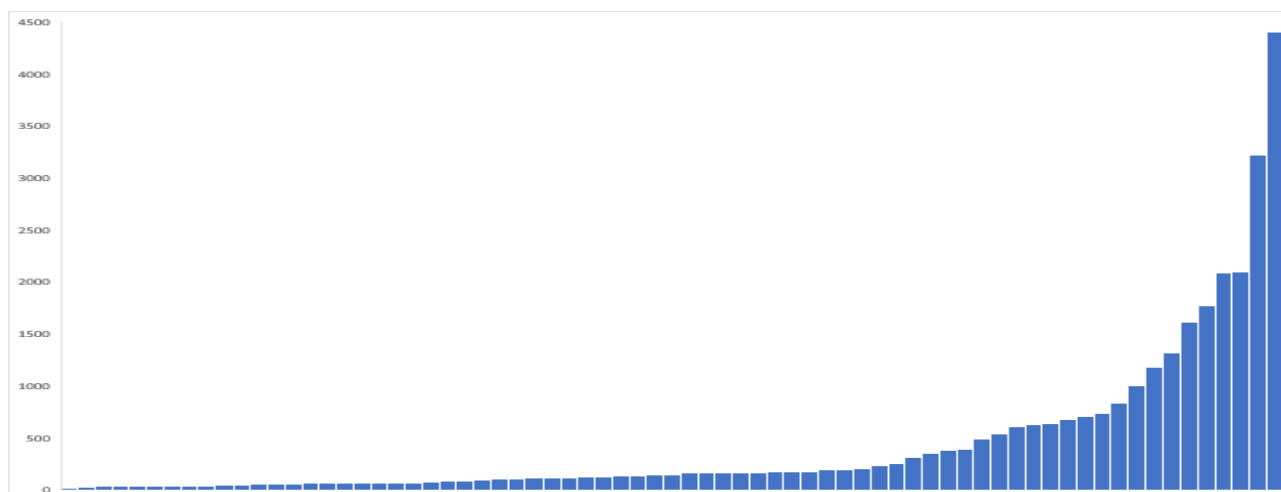
Fonte: ns. elaborazioni su dati Eurostat – [isoc_ciweb]

4. Ottenute le pagine potenzialmente "interessanti", è effettuato un *pre-processing* automatico del testo al fine di identificare le parole di contenuto. In questa fase sono state scartate le informazioni provenienti da 10 siti web perché esclusivamente in lingua inglese.

5. Sono stati individuati tra i “segmenti ripetuti” quelli interessanti ai fini dell’analisi. Al termine dell’elaborazione il database contava 3.152 parole, 31.104 occorrenze. Il rapporto *type-token* tra il numero di parole diverse presenti nel corpus (3.152) e parole totali (31.104) è pari al 10,1% che corrisponde ad una frequenza media per parola di 10 occorrenze, dati che evidenziano l’omogeneità dei testi che esprimono, quindi, un lessico circoscritto e ripetitivo. 233 parole compaiono una sola volta nel testo. Il numero di occorrenze per sito è particolarmente variabile. Si va un minimo di 14 ad un massimo di 4.409 – Grafico 1.

Partendo dal corpus è stato individuato un elenco coerente di parole (lista di tematizzazione) capaci di rappresentare l’oggetto della ricerca (prodotti, servizi, tecnologie adottate).

Grafico 1 – Numero di occorrenze per sito



Fonte: ns. elaborazione

6. Utilizzando Qiba è stato possibile calcolare il TFIDF per le parole che compongono la lista di tematizzazione, per poter modellare il testo come “*bag-of-words*”. Attraverso questa operazione è possibile riconoscere le parole più rilevanti per l’analisi del contenuto. Considerando l’estrema variabilità che caratterizza il numero di occorrenze dei siti analizzati e per evitare l’effetto distorsivo f_i viene calcolato come sommatoria dei rapporti calcolati per ogni sito web tra il numero di occorrenze del termine i nel documento (sito web, nel nostro caso) e il numero di occorrenze totale del sito web. La “*bag-of-words*” utilizzata nella fase di *document clustering* è composto dalle 60 parole con TFIDF più elevato.
7. Il *document clustering* è stato condotto utilizzando come algoritmo il *k-means*, e la distanza euclidea come misura di distanza. Si evidenziano tre gruppi, Il primo è composto da 22 imprese, il secondo da 44 e il terzo da 4. L’analisi delle parole più caratterizzanti per ognuno dei tre gruppi permette di proporre una classificazione che aggiunge informazioni rispetto al codice Ateco di partenza – “fabbricazione di altre macchine utensili”. Al primo cluster appartengono le imprese della meccanica classica, che rivolgono i propri prodotti al settore delle lavorazioni meccaniche.

Al secondo i fornitori di sistemi di automazione con mercati diversificati (lavorazioni del granito, del marmo e della meccanica), al terzo le imprese che con maggior frequenza segnalano la produzione di macchine che utilizzano la tecnologia laser. Le verifiche condotte direttamente sui siti delle imprese sembrano confermano l'attribuzione effettuata attraverso la *cluster analysis*.

I risultati evidenziano che i dati ottenuti attraverso la raccolta dal web di informazioni (*web scraping*) e la loro successiva analisi con tecniche di *text mining*, *text analysis* e *document clustering* offrono informazioni aggiuntive rispetto alla classificazione Ateco delle imprese stesse.

Conclusioni

La letteratura economica identifica in maniera chiara il concetto di settore ma l'operativizzazione dello stesso subisce una serie di limiti legati alla carenza di informazioni, alla lentezza di aggiornamento delle stesse e alla rigidità delle tassonomie impiegate. Partendo da tali considerazioni il presente lavoro si è posto l'obiettivo di verificare se sia possibile utilizzare strategie di analisi basate sulla raccolta dal web di informazioni (*web scraping*) e la loro successiva analisi con tecniche di *text mining*, *text analysis* e *document clustering* per arricchire le informazioni che possono essere utilizzate per attribuire un'impresa ad un settore.

L'utilizzo di tali tecniche nell'ambito dell'analisi economica trova ormai una discreta diffusione in applicazioni diverse, favorito da una serie di vantaggi che sono legati alla disponibilità dei dati dal web: ricchezza di informazioni aggiornate il cui reperimento ha un costo relativamente basso e non intrusività delle indagini, un aspetto rilevante in un periodo in cui i tassi di risposta ai questionari somministrati a scopo di ricerca sono in costante riduzione.

La metodologia utilizzata, che sfrutta in molte fasi delle ricerca Qiba - *Quantitas Intelligent Business Analyzer* - un software sviluppato ad hoc per supportare le fasi di *web crawling*, *web scraping*, pre-processing e pre-analisi del testo sviluppato, viene applicata a titolo sperimentale all'analisi del settore della meccanica di Piemonte e Veneto utilizzando come dati di partenza le informazioni estratte dai siti web di un gruppo di imprese che condividono l'attribuzione al codice Ateco "28.49.09 - fabbricazione di altre macchine utensili". I risultati dell'attività di *document clustering* permettono di proporre una classificazione che aggiunge informazioni rispetto a quella di partenza.

I risultati evidenziano, quindi, che i dati ottenuti attraverso la raccolta dal web di informazioni (*web scraping*) e la loro successiva analisi con tecniche di *text mining*, *text analysis* e *document clustering* offrono informazioni aggiuntive rispetto alla classificazione Ateco delle imprese stesse.

Bibliografia

- Al-Hassan, A. A., Alshameri, F., & Sibley, E. H. (2013). A research case study: Difficulties and recommendations when using a textual data mining tool. *Information & Management*, 50(7), 540-552.
- Arora, S. K., Youtie, J., Shapira, P., Gao, L., & Ma, T. (2013). Entry strategies in an emerging technology: a pilot web-based study of graphene firms. *Scientometrics*, 95(3), 1189-1207.
- Arora, S. K., Li, Y., Youtie, J., & Shapira, P. (2016). Using the wayback machine to mine websites in the social sciences: a methodological resource. *Journal of the Association for Information Science and Technology*, 67(8), 1904-1915.
- Barbarito, L. (1999). *L'analisi di settore. Metodologia ed applicazioni* (Vol. 24). FrancoAngeli.
- Barcaroli, G., Nurra, A., Scarnò, M., & Summa, D. (2014, June). Use of web scraping and text mining techniques in the Istat survey on "Information and Communication Technology in enterprises". In *Proceedings of quality conference* (pp. 33-38).
- Breton, R., Clews, G., Metcalfe, L., Milliken, N., Payne, C., Winton, J., & Woods, A. (2015). Research indices using web scraped data. Office for National Statistics UK.
- Cemin, M., & Tuzzi, A. (2013). I discorsi dei presidenti di Confindustria: una lettura mediante l'analisi statistica dei dati testuali. in *Parole, economia, storia. I discorsi dei presidenti di Confindustria dal 1945 al 2011*, Marsilio Editore, Venezia.
- Cortelazzo, M. A., & Gambarotto, F. (a cura di) (2013). *Parole, economia, storia. I discorsi dei presidenti di Confindustria dal 1945 al 2011*. Marsilio Editore, Venezia.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37-37.
- Geroski, P. A. (1998). Thinking creatively about markets. *International Journal of Industrial Organization*, 16(6), 677-695.
- Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, 102(1), 653-671.
- Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008)*, Christchurch, New Zealand (Vol. 4, pp. 9-56).
- Istat (2009), *Classificazione delle attività economiche Ateco 2007 – derivate dalla Nace Rev. 2*, https://www.istat.it/it/files/2011/03/metenorme09_40classificazione_attivita_economiche_2007.pdf
- Kinne, J., & Resch, B. (2018). Generating Big Spatial Data on Firm Innovation Activity from Text-Mined Firm Websites. *GI_Forum* 2018,, 1, 82-89.
- Lind, J. (2005, January). Ubiquitous Convergence: market redefinitions generated by technological change and the Industry Life Cycle. In *DRUID Academy Winter 2005 Conference*.

- Miner, G., Elder IV, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). Practical text mining and statistical analysis for non-structured text data applications. Academic Press.
- Kay, J. A. (1990). Identifying the strategic market. *Business Strategy Review*, 1(1), 2-24.
- ten Bosch, O., Windmeijer, D., van Delden, A., & van den Heuvel, G. (2018, October). Web scraping meets survey design: Combining forces. In Big Data Meets Survey Science Conference, Barcelona, Spain.
- Tuzzi, A. (2010). What to put in the bag? Comparing and contrasting procedures for text clustering. *Italian Journal of Applied Statistics/Statistica Applicata*, 22(1), 77-94.