Il *Text Mining* per l'individuazione dell'offerta universitaria di competenze nel terzo settore[1]

Simona Balbi[*], Giorgio Infante[*], Michelangelo Misuraca[**]
**Università di Napoli Federico II*
***Università della Calabria*

Riassunto. La riforma del sistema universitario Italiano prodotta dalla L. 509/99 è stata pensata allo scopo di riorganizzare e razionalizzare l'offerta formativa. Non tutte le potenzialità di tale riforma sono state però colte dalle università. Da un lato si è avuta una proliferazione eccessiva di corsi, dall'altro non sono stati spesso ben definiti obiettivi ed ambiti in fase di progettazione dei corsi stessi. Ciò è tanto più vero per quei segmenti del mercato del lavoro, nei quali gli studenti andranno ad operare, di ancora difficile e confusa definizione. Obiettivo di tale lavoro è individuare quali sono le competenze offerte dai programmi di studio che hanno come fine la preparazione alle professioni proprie del terzo settore, attraverso uno studio statistico del linguaggio utilizzato per descrivere gli obiettivi dei corsi di laurea attinenti.
Parole chiave: Analisi in componenti principali vincolata, Informazione esterna, Proiettori ortogonali, Dati testuali.

1 Introduzione

La riforma del sistema universitario seguita all'applicazione della L. 509/99 ha posto come uno dei principali punti qualificanti la maggiore flessibilità dell'offerta formativa, la valorizzazione dell'autonomia dei singoli atenei (anche in una logica di concorrenza e competizione), una minore auto-referenzialità del sistema universitario ed una maggiore attenzione alla domanda di formazione proveniente dall'esterno.

Questo si è tradotto nel superamento dell'organizzazione dei percorsi formativi in un numero ridotto di corsi, facenti capo a specifiche strutture didattiche,

[1] Il presente lavoro è stato finanziato nell'ambito del progetto "Modelli e metodi per abbinare profili formativi e bisogni di professionalità di comparti del terziario avanzato", cofinanziato dal MIUR e dall'Università di Padova. Coordinatore nazionale è L. Fabbris, coordinatore dell'Unità locale è S. Balbi.

le facoltà, attraverso la definizione di *Classi di corsi di laurea* che consentissero ai singoli atenei di caratterizzarsi rispetto ai propri specifici punti di forza, in questo valorizzando la loro autonomia, anche in termini di offerta formativa.

Lo scopo originario era indubbiamente "virtuoso", ma in diversi casi ha prodotto situazioni di confusione e ambiguità: proliferazione di corsi legati a interessi corporativistici, definizione di *curricula* sulla base di mode effimere, eccessivo frazionamento e specializzazione, anche nei corsi di primo livello[2].

Obiettivo del presente lavoro è analizzare come i singoli corsi di laurea siano stati progettati e presentati ai propri potenziali "clienti" (gli studenti degli ultimi anni delle scuole superiori e le loro famiglie), in termini di comunicazione delle proprie caratteristiche formative e degli sbocchi occupazionali possibili, ponendo particolare attenzione alle competenze che i corsi si propongono di fornire.

Con questo obiettivo si sono analizzate le descrizioni presenti sul sito del Ministero dell'Università (www.cercauniversità.it), relativi all'anno accademico 2005-2006, con strumenti propri dell'analisi dei dati testuali, al fine di far emergere l'immagine che l'università da di sé.

Considerata l'enorme ricchezza di offerta formativa, il raggiungimento di questo obiettivo ha posto il problema di estrarre informazione da una grande base di dati documentaria, ossia uno dei più tipici compiti del *text mining*. Nel seguito, dopo un inquadramento di carattere metodologico (par. 2), è illustrata la strategia adottata per la riduzione della complessità del fenomeno oggetto di analisi (par. 3), per concludere con la presentazione dei risultati raggiunti (par. 4).

2 Quadro metodologico di riferimento

Una delle maggiori caratteristiche di un'analisi statistica effettuata a partire da una base documentaria, e più in generale una delle maggiori differenze tra una *query* su una base di dati numerica e una effettuata, ad esempio, mediante un motore di ricerca, è legata alla necessità di un impegnativo pre-trattamento.

Durante questa fase preliminare la trasformazione di una base documentaria non strutturata in una struttura di dati trattabile con tecniche di analisi statistica si traduce, necessariamente, in una riduzione della variabilità (linguistica). Si pensi a quello che accade quando si decide, ad esempio, di procedere ad una *lemmatizzazione*, o allo *stemming*. La riduzione di variabilità significa, in positivo, una riduzione del "rumore", cioè dell'informazione non significativa rispetto al

[2] Le modifiche introdotte dal D.M. 270/2004 sono di ancora troppo recente attuazione perché sia possibile, ad oggi, valutarne la portata correttiva.

fenomeno oggetto di analisi. Può significare però anche perdita di informazione, critica per la comprensione del fenomeno stesso. Un analogo problema si pone quando si decide di lavorare raggruppando i documenti in classi, ad esempio per realizzare successivamente un'analisi su una tabella lessicale aggregata. Anche in questo caso si ha di fatto un minor rumore/informazione.

Molto spesso si ha l'esigenza di recuperare o integrare, in qualche modo, almeno una parte dell'informazione perduta.

Esiste infatti molta altra informazione che si perde nella costruzione di una tabella lessicale, ed è quella legata al contesto all'interno del quale ciascun termine è utilizzato. Accanto a questo tipo di informazione, legata strettamente al *corpus*, può essere interessante introdurre nell'analisi altre informazioni, non individuabili direttamente dal *corpus*, relative al contesto in cui i documenti sono stati prodotti.

L'informazione esterna sul dato, ossia la *meta-informazione*, viene di solito recuperata, soggettivamente, in maniera informale, nell'interpretazione dei risultati. Si propone qui di affrontare il problema in maniera più formalizzata, incorporando questa informazione nell'analisi.

2.1 L'utilizzo dell'informazione esterna

Il tema dell'introduzione di *informazione esterna* nell'analisi esplorativa di strutture multivariate è stato ampiamente dibattuto in letteratura, a partire dall'analisi in componenti principali (ACP) con variabili strumentali proposta da Rao (1964). L'introduzione di una informazione esterna sia sugli individui che sulle variabili è stata proposta da Takane e Shibayama (1991), con un metodo che combina caratteristiche proprie del modello di regressione e dell'ACP. Questo metodo è stato successivamente e diffusamente sviluppato da Takane (1997), con la cosiddetta analisi in componenti principali vincolata, ed esteso ad una grande varietà di metodi di analisi multivariata.

Un approccio non dissimile, ma di natura prettamente descrittiva e focalizzato maggiormente sugli aspetti geometrici e sulla visualizzazione, è quello proposto da D'Ambra e Lauro (1982), esteso al caso dell'analisi di tabelle di contingenza in cui si può assumere una struttura di dipendenza fra le due variabili nominali oggetto di analisi (Lauro, D'Ambra, 1984). L'analisi in componenti principali in un sottopazio di riferimento rappresenta il punto di partenza metodologico del presente lavoro, e in particolare è presentata una sua estensione che include l'informazione esterna riferita alle due vie di una tabella lessicale del tipo (*termini*, *documenti*).

2.2 Alcuni richiami metodologici

2.2.1 Analisi in Componenti Principali Vincolata

La struttura dei dati nell'Analisi in Componenti Principali Vincolata (ACPV) è costituita da una matrice **Z** (*individui*, *variabili*) e da due matrici, **G** ed **H** contenenti, rispettivamente, informazione esterna sugli individui e sulle variabili.

Takane riconduce a questo schema numerose analisi statistiche multivariate, compresa l'analisi delle corrispondenze e la sua variante non simmetrica (Takane, 2008), senza limitazioni circa la distribuzione delle variabili, il pre-trattamento o la metrica, scelte in accordo allo specifico interesse applicativo del ricercatore.

L'ACPV si articola su due passi: nel primo, la cosiddetta *analisi esterna*, **Z** è proiettata ortogonalmente negli spazi generati da **G** e **H** con l'obiettivo di scomporre l'influenza delle variabili esterne nella somma di quattro termini: il comportamento di **Z** che può essere spiegato congiuntamente da **G** e **H**, quello spiegato solo da **G**, solo da **H** e una componente residuale. Questa soluzione è ottenuta nell'ottica dei minimi quadrati, attraverso la minimizzazione della matrice dei residui. Nel secondo passo si esegue la cosiddetta *analisi interna*, che consiste in una o più ACP effettuate sulle matrici ottenute nel primo passo dalla scomposizione di **Z**.

2.2.2 Analisi in Componenti Principali su un Sottospazio di Riferimento

La struttura dei dati dell'Analisi in Componenti Principali su un Sottospazio di Riferimento (ACPR) è data da due matrici (*individui*,*variabili*), **Z** and **X**.

L'ACPR si pone l'obiettivo di visualizzare la dipendenza di **Z** da **X**. Operativamente ricerca le componenti principali della proiezione ortogonale di **Z** sullo spazio generato dalle colonne di **X**. Da questo punto di vista può essere letta come un caso particolare dell'ACPV, prendendo in considerazione soltanto il termine della scomposizione relativo all'influenza di **X** su **Z** (centrando e, di solito, standardizzando le variabili in **Z** come, in una classica ACP).

I vantaggi dell'ACPR sono strettamente connessi alle sue rappresentazioni grafiche e, conseguentemente, alla facilità di interpretazione dei risultati: le mappe fattoriali mostrano sia le correlazioni fra le variabili appartenenti alla stessa matrice (**X** o **Z**) sia quelle fra le variabili di **X** e **Z**.

2.2.3 Analisi Doppiamente Proiettata

Balbi e Misuraca (2009) propongono un'analisi doppiamente proiettata, al fine di introdurre informazione esterna sia sugli individui che sulle variabili, nell'ottica specifica dell'Analisi dei Dati Testuali (ADT).

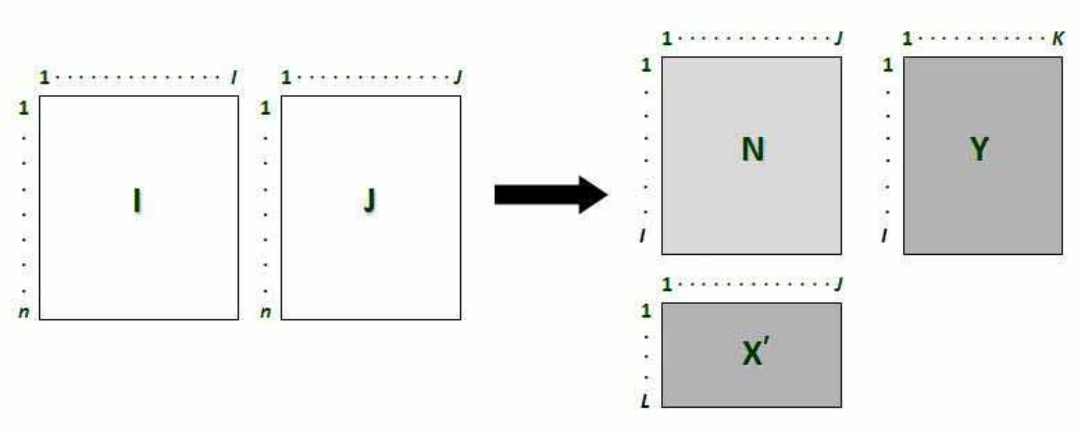
La proposta ha come obiettivo la comprensione dell'uso di determinate *keyword*, sotto condizioni date. Ciò consente, ad esempio, di comprendere le peculiarità d'utilizzo di termini presenti in singoli documenti senza far riferimento, però, all'interpretazione globale del fenomeno linguistico oggetto d'analisi.

Si supponga di avere due matrici indicatrici **I** (*n*, *I*) e **J** (*n*, *J*), rappresentative di due variabili nominali osservate sullo stesso collettivo di individui, e sia **N** (*I*, *J*) la matrice che incrocia le due variabili in **I** e **J**. Nell'ambito dell'ADT, **N** è una tabella lessicale che ha in riga *I* documenti e in colonna *J* parole rappresentative del vocabolario della collezione di documenti analizzata. Un'Analisi delle Corrispondenze Lessicali (ACL) è tipicamente eseguita su questa matrice, per analizzare e rappresentare graficamente le relazioni latenti fra documenti e termini.

Poniamo ora di disporre di informazioni relative ad una classificazione interessante dei documenti e al contesto di utilizzo dei termini. Si consideri una matrice indicatrice **Y** (*I*, *K*) che assegna ogni documento alla categoria *k* (*k* = 1, …, *K*). È possibile effettuare allora una ACL sulla tabella lessicale aggregata **T** (*K*, *J*) ottenuta dal prodotto di **Y** e **N**, per una migliore lettura delle relazioni fra gruppi di documenti e termini (Lebart et al., 1997). L'introduzione di informazione esterna sui termini era già stata proposta in precedenti lavori (Balbi e Giordano, 2000; Balbi et al., 2002).

L'analisi doppiamente proiettata porta l'attenzione sull'analisi interna della ACPV e sulle caratteristiche geometriche proprie dell'ACPR. In altri termini, attraverso gli operatori di proiezione ortogonale si visualizza sui piani fattoriali la struttura di associazione in **N** relativa alla informazione esterna. Oltre ad introdurre la matrice **Y** (informazione sui documenti) ed **N** (la tabella lessicale), si considera una matrice **X** (*J*, *L*) relativa alla informazione sul vocabolario (Fig. 1).

Figura 1. *La struttura dei dati dell'analisi doppiamente proiettata*



Nell'ottica della proposta di Takane è possibile in tal modo studiare una delle matrici risultanti dalla scomposizione: l'influenza della categorizzazione dei documenti sull'analisi testuale, l'influenza della classificazione dei termini sull'analisi, l'influenza congiunta, o la parte residua, che non dipende né dall'informazione esterna sui documenti né da quella sui termini.

3 L'offerta universitaria per il terzo settore

Non è semplice né banale definire il terzo settore. Nella letteratura di riferimento sono presenti innumerevoli tentativi, in accordo alle diverse scuole o alle diverse teorie sociologiche. Nel seguito, si è scelta la seguente definizione di riferimento:
"*il Terzo Settore costituisce quell'area che si è andata formando tra stato e mercato, nella quale si offrono servizi, si scambiano beni relazionali, si forniscono risposte a bisogni personali o a categorie deboli secondo approcci che non sono originariamente connotati da strumentalità (come nel mercato), né da puro assistenzialismo (come nello stato)*"[3].

Alla luce di ciò sono stati selezionati 139 corsi di laurea triennale tra i 3082 offerti dalle Università Italiane nell'anno accademico 2005/2006, che nella loro descrizione sulla base di dati ministeriale fanno esplicitamente riferimento al terzo settore (Tab. 1).

[3] G. Di Gennaro, www.terzosettorenapoli.it

Tabella 1. *Corsi di Laurea Triennali di interesse per il Terzo settore*

Classe di Laurea	Denominazione	N° di corsi selezionati
6	*Sc. del servizio sociale*	47
15	*Sc. politiche e delle relazioni internazionali*	1
18	*Sc. dell'educazione e della formazione*	66
28	*Sc. Economiche*	6
35	*Sc. sociali per la cooperazione, lo sviluppo e la pace*	19

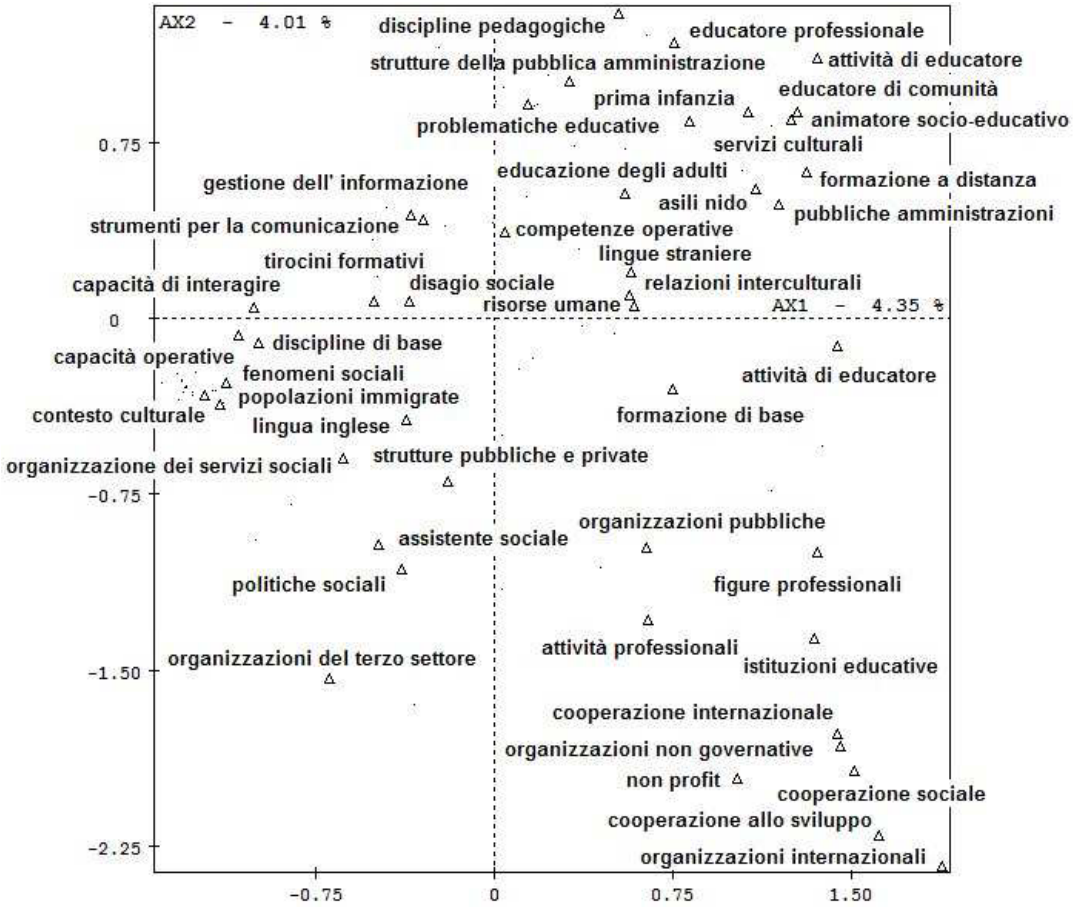
La base documentale analizzata è costituita dalle declaratorie ministeriali dei corsi di laurea. Ciascuna declaratoria è strutturata in tre sezioni, comprendenti gli obiettivi formativi, gli ambiti occupazionali e le conoscenze pregresse richieste.

La prima parte del trattamento necessario per analizzare le declaratorie da un punto di vista statistico è quella solitamente utilizzata nell'ADT. I documenti sono stati normalizzati, lessicalizzati e quindi lemmatizzati, allo scopo di ridurre la variabilità linguistica ed eliminare le forme strumentali prive di significato. Al fine di estrarre le *keyword* inerenti le competenze è stato ricostruito un *corpus* costituito dai soli obiettivi formativi.

Sulla tabella lessicale ottenuta da tale *corpus* è stata effettuata un'ACL, di cui per brevità si riporta solo la rappresentazione grafica (Fig. 2), al fine di acquisire conoscenza sul fenomeno d'interesse.

Si può notare sul primo asse fattoriale (da sinistra a destra) una contrapposizione tra le professionalità del sociale in ambito pubblico e/o privato contraddistinte dal fatto di fornire servizi alle persone in quanto individualità, e le professionalità proprie della formazione e della cooperazione internazionale, nelle quali si ha una maggior attenzione ai servizi per la comunità. Sul secondo asse fattoriale risulta evidente una contrapposizione tra i corsi propri delle scienze della formazione (in alto) e quelli invece più attinenti alle scienze sociali (in basso).

Figura 2. *ACL sui corsi di laurea: primo piano fattoriale*



A valle di tale analisi è stata effettuata una categorizzazione semantica delle competenze considerate, ottenendo così 5 differenti classi che considerano la natura della competenza, l'ambito di utilizzo e le attitudini complementari richieste.

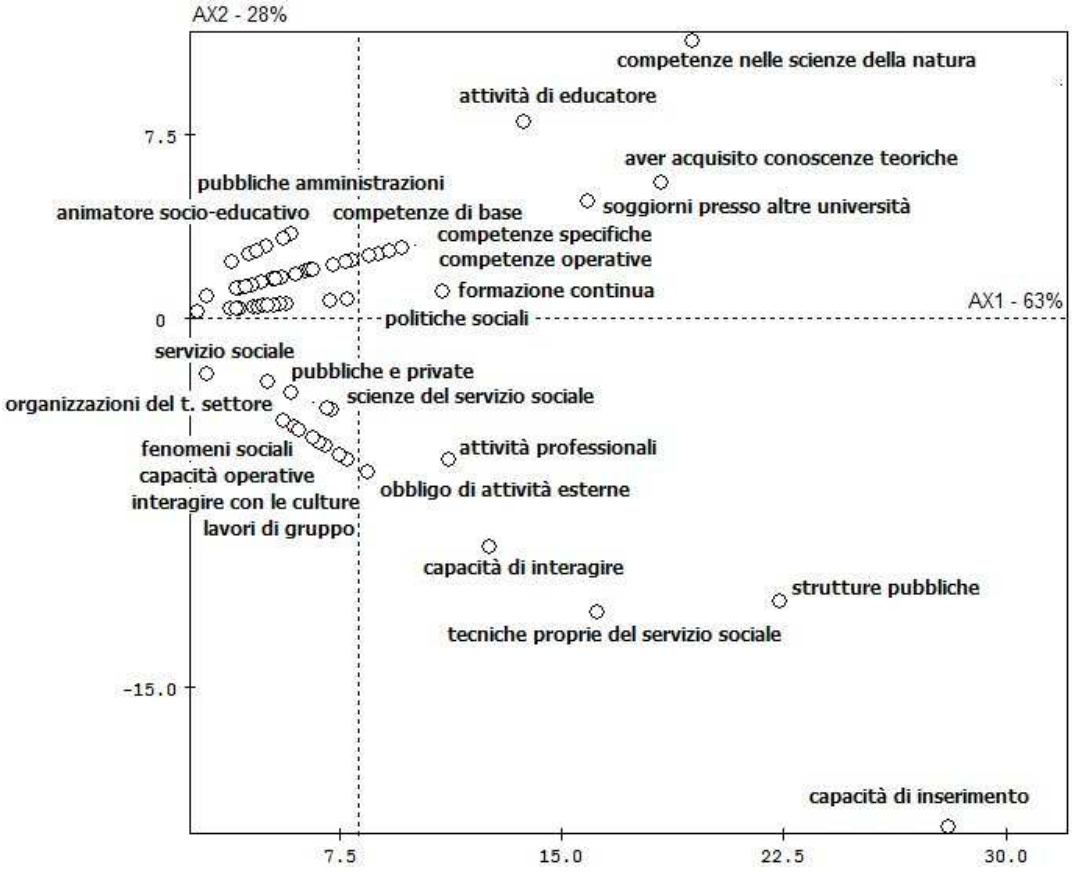
4 Analisi delle competenze

I dati a disposizione sono stati organizzati in tre diverse matrici: una tabella lessicale (*corsi di laurea*,*competenze*), una tabella d'informazione esterna sui corsi di laurea (*classi di laurea*) e una tabella d'informazione esterna sulle competenze (*classi di competenze*).

L'analisi interna sulla tabella lessicale (*corsi di laurea,competenze*), nel caso in cui si considerino come informazione esterna due matrici identità, e come metrica e sistema di pesi nei due spazi di rappresentazione di righe e colonne le distribuzioni marginali, corrisponde all'ACL presentata in Fig. 2. Per brevità non vengono riportati i risultati dell'analisi con metrica e sistemi di pesi basati sull'informazione esterna perché non di interesse per l'obiettivo del presente lavoro.

Dell'analisi esterna sulle quattro tabelle, secondo la decomposizione proposta da Takane, sono di seguito presentati i risultati relativi alla proiezione delle *keyword* dei corsi di laurea nello spazio delle classi di competenze, senza tener conto quindi delle diverse classi di laurea, e sull'uso residuale delle *keyword* rispetto alle classi di laurea e alle classi di competenze.

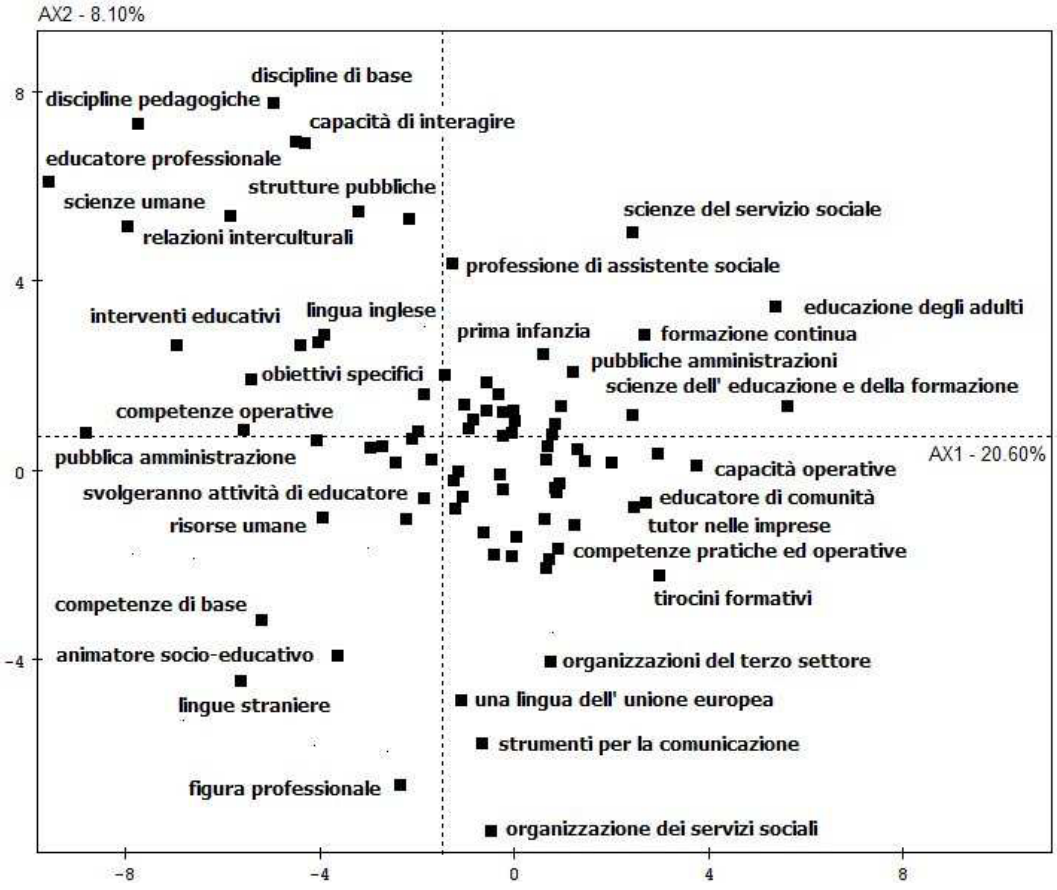
Figura 3. *Proiezione delle keyword nello spazio delle competenze*



Nella Fig. 3 è possibile visualizzare la descrizione delle competenze a prescindere dalla classe di laurea, quindi senza considerare il contesto d'uso. Sul

primo asse fattoriale si nota prevalentemente una opposizione fra descrizioni incentrate su quello che l'università s propone di offrire, in termini di competenze e di valorizzazione di capacità, sulla destra e discorsi più generici sulla sinistra. Sul secondo asse fattoriale è invece possibile visualizzare (dall'alto verso il basso) una contrapposizione tra le competenze che i diversi Corsi di Laurea si propongono di offrire (*nelle scienze della natura*, *specifiche*, *operative*) e le capacità (*di interagire*, *di inserimento*, *operative*) che gli studenti dovranno acquisire per poter lavorare nei diversi ambiti lavorativi.

Figura 4. *Usi residui delle keyword nello spazio delle competenze*



Nella Fig. 4 è possibile visualizzare invece l'utilizzo delle *keyword* utilizzate dai corsi di laurea, senza tener conto delle classi di laurea (contesto d'utilizzo) e delle classi di competenze. Appaiono, quindi, nei quattro quadranti le specifiche figure professionali che i diversi gruppi di corsi di laurea si propongono di offrire, in senso

orario: *educatore professionale* (in alto a sinistra), *assistente sociale* (in alto a destra), *tutor di impresa* (in basso a destra, con *educatore di comunità*) e in generale lavorare in *organizzazioni* (*dei servizi sociali, del terzo settore*) e *animatore socio-educativo* (in basso a sinistra).

5 Conclusioni

La strategia proposta ha consentito di approfondire il linguaggio utilizzato dalle università che hanno scelto, nell'ambito della lagge 509/99, di attivare corsi di laurea triennali per operatori del terzo settore. In questo senso, la metodologia fornita ha consentito di tener conto uno degli elementi maggiormente rilevanti della descrizione degli obiettivi formativi: il tipo di competenze offerte. Ancora, riferendosi alla scompozione dell'informazione, proposta da Takane, si è cercato di fare emergere le specificità individuali, che sembrano far emergere la tendenza delle università a collegare gli obiettivi formativi alle professionalità che si intendono formare. In ogni caso, appare una notevole ripetitività dei termini e dei concetti ad essi collegati, che non sembra nascere da una reale volontà di concorrenza.

Potrà essere interessante comparare questi risultati (ad esempio con tecniche procustiane, quali quelle proposte da Balbi, Misuraca, 2005) con quelli relativi ai nuovi corsi di laurea introdotti dalla legge 270/2004, al fine di comprendere se l'obiettivo di informazione, a valenza anche promozionale, oltre che informativa, sia entrata più in profondità nel sistema universitario e nella sua progettazione.

Riferimenti bibliografici

BALBI S., BOLASCO S., VERDE R. (2002) Text Mining on Elementary Forms in Complex Lexical Structures, In: MORIN A., SÉBILLOT P. (eds.), *Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles (JADT'02)*, IRISA-INRIA, Paris: vol. 1: 89-100.

BALBI S., GIORDANO G. (2000) A Factorial Technique for Analysing Textual Data with External Information, in: BORRA S. et al. (eds.), *Advances in Classication and Data Analysis*, Springer, Heidelberg: 169-176.

BALBI S., MISURACA M. (2005), "Procustes techniques for Text Mining", in ZANI S., CERIOLI A. (eds.), *Classifications and Data Analysis 2005. Book of Short Papers (CLADAG05)*, M.U.P., Parma: 37-40

BALBI S., MISURACA M. (2009) A doubly projected analysis for lexical tables, in: SKIADAS C.H. (ed.) *Advances in Data Analysis*, Birkhäuser [in stampa]

D'AMBRA L., LAURO N.C. (1982) Analisi in Componenti Principali in Rapporto ad un Sottospazio di Riferimento, *Rivista di Statistica Applicata,* **1(15)**: 51-67.

LAURO N.C., D'AMBRA L. (1984) L'Analyse Non Symétrique Des Correspondances. In: DIDAY E. et al. (eds.), *Data Analysis and Informatics*, North Holland, Amsterdam: vol. III: 433-446.

LEBART L., SALEM A., BERRY L. (1997) *Exploring Textual Data*. Kluwer Academic Publishers, Boston.

RAO C.R. (1964) The use and interpretation of principal component analysis in applied research, *Sankhya*, series A, **26**: 329-358.

TAKANE Y. (1997) CPCA: A Comprehensive Theory. In: *Proceedings of the 1997 IEEE-SMC International Conference*: vol. 1: 35-40.

TAKANE Y. (2008), More on regularization and (generalized) ridge operators, in Shigemasu K., Okada A., Imaizumi T. and Hoshino T. (Eds.) *New Trends in Psychometrics*, University Academic Press, Tokio: 443-452.

TAKANE Y., SHIBAYAMA T. (1991) Principal Component Analysis with External Information on both Subjects and Variables, *Psychometrika*, **1(56)**: 97-120.

Text Mining for detecting the academic competences offered for the third sector

Summary. *The reform of the Italian university system lead by the L. 509/99 has been thought aiming at reorganizing and rationalizing the academic offer. Not all the potentialities of such reform have been however taken by the Universities. It has been an excessive proliferation of programs and, at the same time, the objectives have not often been well defined in a planning phase. This is more significant for that segments of the job market, in which the students will operate, of hard and confused definition. This paper aims at identifying which competences are offered by the academic programs that prepare to the Third Sector professionalisms, by a statistical study of the language used for describing the goals of the different proper programs.*

Keywords**:** *Constrained Principal Component Analysis, External Information, Orthogonal Projectors, Textual Data*