Relazioni non Simmetriche tra Corpora

Maria Gabriella Grassia, Michelangelo Misuraca, Germana Scepi

Dipartimento di Matematica e Statistica – Università Federico II – Napoli – Italia
michelangelo.misuraca@unina.it

Abstract

In this paper the language used by firms for searching new employers by web is studied. Particularly, we are interesting in evaluating the dependence between two *corpora*, e.g. one defined by the forms used for describing the skills of the candidates for jobs and the other defined by the forms used by firms for describing their mission. The method used is textual data analysis, more precisely, a non symmetrical correspondence analysis on a peculiar lexical table forms/forms with an *ad hoc* weighting system on the explanatory variables. Furthermore, the main results of an application on a sample of firms are showed in terms of friendly readable graphical representations.

Riassunto

In questo lavoro si presenta uno studio sul linguaggio utilizzato dalle aziende alla ricerca di candidati da assumere per differenti mansioni. L'obiettivo è quello di valutare la dipendenza tra due *corpora*, con riferimento particolare alla relazione tra le forme usate per definire le caratteristiche dei candidati richiesti e le forme utilizzate dalle stesse aziende per descrivere la propria *mission* aziendale. Il metodo considerato è l'analisi dei dati testuali e, precisamente, l'analisi non simmetrica delle corrispondenze applicata però, ad una particolare tabella lessicale del tipo forme/forme, con l'introduzione, inoltre, di un sistema di pesi *ad hoc* sulle variabili esplicative. Nel lavoro vengono presentati i principali risultati, in termini soprattutto di rappresentazioni grafiche, conseguiti dall'applicazione su un campione di 167 aziende di diversa dimensione e settore di attività, distribuite su tutto il territorio nazionale.

Keywords: textual data analysis, non symmetrical correspondence analysis, term frequency.

1. Introduzione

Nel presente lavoro l'attenzione è rivolta allo studio del linguaggio utilizzato dalle aziende per assumere nuovo personale. In particolare, si vuole analizzare ed esplicitare la possibile relazione tra il linguaggio che ciascuna azienda usa per descrivere la propria *mission* (che da ora in poi battezzeremo con l'allocuzione "chi siamo") e quello utilizzato per descrivere i profili dei candidati all'assunzione (che battezzeremo con "chi cerchiamo"). L'ipotesi di partenza è che tale relazione non sia di tipo simmetrico bensì che si possa ipotizzare la dipendenza di un linguaggio dall'altro, ossia che il "chi siamo" influenzi il linguaggio utilizzato per descrivere il "chi cerchiamo".

Lo strumento prescelto è l'analisi dei dati testuali vista come estensione di metodi statistici proposti in origine per l'analisi delle relazioni tra variabili numeriche. A differenza della classica matrice di contingenza, del tipo documenti/forme, generalmente analizzata da tale tecnica, nel presente lavoro si definisce una matrice del tipo forme/forme. In particolare, partendo dai due differenti *corpora* rilevati sulle stesse unità statistiche, si costruisce una matrice di *co-presenza* che ha come termine generico il numero di volte in cui le forme dei due *corpora* si presentano contemporaneamente.

L'ipotesi di partenza ci induce ad analizzare tale matrice con una tecnica di analisi dei dati di tipo non simmetrico, e, per la precisione, con l'Analisi non Simmetrica delle Corrispondenze (ANSC, Lauro e D'Ambra, 1984). L'ANSC per tabelle lessicali è stata proposta da Balbi (1995) in alternativa all'uso dell'Analisi delle Corrispondenze (Lebart *et al.*, 1991), laddove le variabili interessate non sembrano essere in una relazione di tipo simmetrico. L'ANSC, inoltre, essendo basata su una metrica euclidea non ponderata, risulta particolarmente adatta per l'analisi di tabelle lessicali ricche di zeri, dove, invece, l'utilizzo di una metrica χ^2 finisce con l'attribuire un'importanza eccessiva alle modalità rare.

Rispetto al metodo originario, nel lavoro si introduce un sistema aggiuntivo di pesi sulle forme del vocabolario del *corpus* del "chi siamo", supposte esplicative rispetto alle forme del vocabolario del "chi cerchiamo". Tale sistema consente di introdurre ulteriori informazioni sulla frequenza di utilizzo delle forme e di migliorare i risultati dell'analisi, sia in termini di leggibilità delle rappresentazioni grafiche che di aumento dell'indice di predittività (τ di Goodman e Kruskal, 1954).

Nel paragrafo successivo (paragrafo 2) verranno introdotti i principi generali dell'ANSC, fornendo regole per l'interpretazione delle rappresentazioni grafiche da essa ottenute. L'interesse e la facilità di lettura di tali rappresentazioni le rendono decisamente uno dei motivi principali dell'utilizzo di tale tecnica nell'ambito dei dati testuali. Nel paragrafo 3 si presenterà la struttura dei dati analizzata e la strategia di analisi adottata. Infine, nel paragrafo 4 verranno mostrati i principali risultati ottenuti dall'applicazione della strategia ad un campione di 167 aziende.

2. Alcuni richiami all'Analisi Non Simmetrica delle Corrispondenze

Consideriamo due variabili qualitative, z_i (con i=1,...,*I*) ed y_j (con j=1,...,*J*), osservate sulle stesse *n* unità e aventi rispettivamente *I* e *J* categorie di risposta. Si classifichino le variabili nella matrice **F**(*I*,*J*) che ha come elemento generico la frequenza relativa congiunta f_{ij}.

Si indichi con **r** il vettore contenente i marginali di riga $f_i = \sum_j f_{ij}$, **c** il vettore dei marginali di colonna $f_j = \sum_i f_{ij}$, $\mathbf{D}_I$=diag(**r**) e $\mathbf{D}_J$=diag(**c**).

Per valutare l'influenza delle *J* categorie di *y* sulle *I* categorie della variabile *z*, l'ANSC trasforma la matrice **F** nella matrice $\tilde{\mathbf{F}}$ centrata rispetto all'ipotesi di indipendenza e di elemento generico:

$$\tilde{f}_{ij} = f_{ij} - f_{i.}f_{j.} \qquad (1)$$

L'attenzione viene, dunque, spostata sulla matrice centrata dei profili colonna, $\tilde{\mathbf{F}}\mathbf{D}_j^{-1}$, che considera le distribuzioni condizionate di *I* rispetto a *J*.

Geometricamente, l'ANSC ha come obiettivo la rappresentazione delle *I* categorie della variabile di risposta in un sottospazio di R^J, assumendo una metrica euclidea e un sistema di pesi pari a $\mathbf{D}_j$. Analogamente le *J* categorie della variabile esplicativa vengono rappresentate in un sottospazio di R^I, assumendo una metrica euclidea ponderata e un sistema di pesi unitario. L'ANSC, in particolare, cerca di visualizzare la dipendenza di *I* da *J* in un sottospazio di dimensioni ridotte R^m, con $m^* < m = [\min(I, J) - 1]$.

Volendo effettuare un paragone con l'Analisi delle Corrispondenze (AC), si può vedere l'ANSC come un'Analisi in Componenti Principali (ACP) sulla tripletta $\tilde{\mathbf{F}}\mathbf{D}_j^{-1}, \mathbf{I}, \mathbf{D}_j$, dove **I** è la matrice identità, mentre è noto che l'AC è un'Analisi in Componenti Principali sulla

tripletta $\tilde{\mathbf{F}}\mathbf{D}_j^{-1}, \mathbf{D}_I^{-1}, \mathbf{D}_j$. Entrambe le analisi sono, dunque, delle ACP sulla stessa matrice di partenza, con lo stesso sistema di pesi, ma in una differente metrica.

Dal punto di vista matematico, la differenza è nella decomposizione in valori singolari (SVD) che nell'ANSC viene effettuata imponendo vincoli differenti sugli autovettori di sinistra. Infatti, nell'ANSC si effettua la SVD di:

$$\tilde{\mathbf{F}}\mathbf{D}_j^{-1} = \mathbf{U}\mathbf{L}\mathbf{U}, \tag{2}$$

con i vincoli di ortonormalizzazione $\mathbf{U'U} = \mathbf{V'D}_j\mathbf{V} = \mathbf{I}$. Λ è la matrice diagonale che ha come elemento generico la radice quadrata degli autovalori, λ_α (con α =1,..,m), della matrice $\mathbf{A}=\tilde{\mathbf{F}}\mathbf{D}_j^{-1}\tilde{\mathbf{F}}'$.

Le coordinate fattoriali sull'α–esimo asse fattoriale in R^I sono così calcolate:

$$\psi_\alpha = \sqrt{\lambda_\alpha u_\alpha} \tag{3}$$

mentre in R^J sono date da:

$$\varphi_\alpha = \sqrt{\lambda_\alpha \mathrm{D}_j^{-1/2} v_\alpha} \tag{4}$$

La traccia della matrice **A**, diagonalizzata nell'analisi, rappresenta il numeratore dell'indice di predittività, caratteristico di quest'analisi, che è il τ di Goodman e Kruskal (1954):

$$\tau = \frac{\sum_\alpha \lambda_\alpha = \left(\sum_i \sum_j f_y^2 / f_i - \sum_i f_i^2 \right)}{\left(1 - \sum_i f_i^2 \right)} \tag{5}$$

Tale indice che presenta al denominatore la misura di eterogeneità del Gini per le distribuzioni condizionate (Light e Margolin, 1971), interpreta la bontà dell'analisi in termini di predittività della variabile dipendente dalla variabile esplicativa. È evidente, che in situazioni in cui il numero di righe e di colonne della matrice di partenza è elevato, caso piuttosto frequente quando si lavora con tabelle lessicali, tale indice risulta piuttosto basso.

Le rappresentazioni grafiche

Uno dei motivi principali della diffusione dell'analisi delle corrispondenze è la capacità di esprimere i risultati sotto forma di rappresentazioni grafiche di facile comprensione.

Esistono alcune regole, in parte specifiche per l'analisi non simmetrica, che è importante richiamare per poter interpretare i grafici proposti dal nostro caso studio:

a) la dispersione della nube dei punti attorno all'origine in R^J visualizza la forza del legame di dipendenza di *I* rispetto a *J*;

b) le due nubi hanno la stessa origine, grazie all'operazione di centratura ed al sistema di pesi adottato;

c) la distanza di un punto dall'origine, in R^J, mostra come la *i*-esima categoria sia influenzata dalle J categorie della variabile esplicativa, così come la distanza di un punto dall'origine, in R^I, visualizza la sua influenza sull'insieme delle *I* categorie della variabile dipendente;

d) se due punti sono vicini, in R^J, vuol dire che sono influenzati dalla stessa categoria della variabile esplicativa; se due punti sono vicini, in R^I, vuol dire che influenzano nello stesso modo le categorie della variabile di risposta;

e) la posizione relativa di una categoria della variabile di risposta e di una categoria della variabile esplicativa, può essere valutata solo in termini di coseni fra l'angolo formato dai vettori che congiungono i due punti dall'origine: un coseno grande mostra una forte influenza, viceversa un coseno piccolo mostra una bassa influenza.

3. La struttura dei dati e la strategia di analisi

Con l'obiettivo di voler studiare il linguaggio che le aziende utilizzano sui siti internet per cercare nuovo personale, sono state raccolte le informazioni disponibili nel sito www.carrierain.it. Tale portale è stato creato dall'Associazione Mercurius di Torino per agevolare i neo-laureati e i diplomati in cerca di prima occupazione nella fase di contatto con le aziende ed, inoltre, per indirizzare coloro che, già occupati, siano alla ricerca di nuovo lavoro. Il Network Mercurius è costituito complessivamente da cinque portali tematici che riguardano la formazione, il lavoro interinale, la ricerca di lavoro.

Nella sezione "Profili aziendali" del sito sono contenuti i profili di 282 aziende (incluse le società di selezione e le società di lavoro interinale). Per il caso studio proposto, si è scelto di non considerare le società di selezione e quelle di lavoro interinale, poiché hanno una funzione mediatrice tra azienda e lavoratore ed utilizzano quindi un linguaggio molto standardizzato. Sono state, dunque, considerate 167 aziende, di diversa dimensione e settore di attività, distribuite su tutto il territorio nazionale.

Scopo preciso dell'analisi è evidenziare come il linguaggio utilizzato dalle aziende per descrivere se stesse e la loro *mission*, influenzi il linguaggio da loro stesse adoperato per descrivere le diverse posizioni lavorative richieste. In una prima fase, quindi, sono stati creati due *corpora* distinti, uno ("chi siamo") contenente le informazioni generali riguardanti l'azienda (storia, struttura, attività) e un altro ("chi cerchiamo") contenente le informazioni specifiche relative alle posizioni lavorative (formazione, competenze, attitudini). I *corpora* così ottenuti sono stati poi normalizzati e lessicalizzati con procedure automatiche, al fine di rendere più facilmente confrontabili le forme costituenti i vocabolari ed evidenziare la presenza di poliformi e polirematiche di interesse per l'analisi.

In una fase successiva è stato costruito nuovamente il vocabolario dei due *corpora* con le forme normalizzate, le polirematiche e i poliformi. Si è scelto, inoltre, di introdurre un filtro sulle forme (numero di occorrenze maggiore di 2 e numero di caratteri maggiore di 4) ed effettuare quindi una lemmatizzazione interna (Lebart *et al.*, 1998), mantenendo separate solo quelle forme che si prestavano ad una interpretazione ambigua, per non perdere informazioni interessanti.

Le tabelle lessicali ottenute si presentano come due matrici di intensità: una (**Y**) in cui sono raccolte 375 forme lemmatizzate del "chi siamo" per le 167 aziende e l'altra (**Z**) in cui sono raccolte 530 forme lemmatizzate del "chi cerchiamo".

A partire da queste due matrici, si vuole costruire una matrice **F** che abbia in colonna le *J* categorie della variabile esplicativa (in **Y**), in riga le *I* della variabile dipendente (in **Z**), e il cui elemento generico consista nel numero di volte in cui ciascuna forma *i*-esima e ciascuna forma *j*-esima siano state utilizzate simultaneamente nel collettivo. La tabella lessicale **F**, quindi, che si vuole ottenere, è del tipo forme/forme e differisce evidentemente dalla classica matrice di contingenza introdotta nel paragrafo precedente. A tal fine è necessario trasformare le matrici **Y** e **Z** in matrici booleane di presenza(1)/assenza(0) di ciascuna forma nei due *corpora*. Si effettua, quindi, il loro prodotto interno **F=Z'Y**.

La matrice **F** così ottenuta è, dunque, una particolare tabella di contingenza e, per la precisione, è una *matrice di co-presenza*. Si noti che i marginali di riga e di colonna di **F** indicano il numero di volte che ciascuna forma di un *corpus* è utilizzata in combinazione con tutte le altre forme appartenenti all'altro *corpus*.

Sulla matrice **F** di *co-presenza* si propone di applicare l'ANSC, introdotta nel paragrafo precedente, con l'obiettivo di studiare il legame di dipendenza del *corpus* del "chi cerchiamo" da quello del "chi siamo". La trasformazione effettuata per passare dalle singole matrici **Z** ed **Y** alla matrice **F** ha causato però la perdita di un'informazione che nell'analisi delle tabelle lessicali risulta particolarmente importante, e cioè il numero di volte che ciascuna forma è stata usata nei due *corpora*, indicazione che non ritroviamo più sui marginali della **F**.

Per tale motivo, si introduce nell'analisi un ulteriore sistema di pesi che riguarda la variabile esplicativa e, quindi, le forme del vocabolario del *corpus* contenente le descrizioni delle aziende. Tale sistema di pesi viene definito a partire dal calcolo del *term frequency* (TF, Salton e Buckley, 1988) per ogni forma della tabella **Y**:

$$\mathrm{TF} = 0,5 + 0,5\frac{nterm_j}{\max nterm} \qquad (6)$$

dove $nterm_j$ rappresenta la frequenza all'interno del *corpus* della forma j-*esima*, mentre max *nterm* è la frequenza della forma che nel *corpus* occorre maggiormente.

In generale, in corrispondenza di livelli più alti del TF si individuano le forme con un contributo informativo maggiore in relazione alla descrizione del testo. Nell'analisi proposta si è scelto di ponderare i dati con il reciproco del TF, in modo da dare un'importanza maggiore alle forme che occorrono di meno e una minore alle forme con occorrenza più alta e, inoltre, vista l'ipotesi di relazione asimmetrica delle variabili, si è deciso di introdurre tale peso solo per la variabile esplicativa.

Nella formula (2) si introduce la matrice $\mathbf{D}_Y^{-1}$ che ha come elemento generico i pesi calcolati (6) per le forme del corpus del "chi siamo". Le formule degli autovettori e delle coordinate sono quindi, modificate tenendo conto di questo nuovo sistema di pesi.

4. I principali risultati

La matrice di co-presenza **F** analizzata ha 530 righe (le parole del *corpus* del "chi cerchiamo") e 375 colonne (le parole del *corpus* del "chi siamo ").

L'ANSC è sta effettuata considerando il sistema aggiuntivo dei pesi ($\mathbf{D}_Y^{-1}$). L'indice di predittività ottenuto è pari a 0,0041. Se non avessimo utilizzato alcun sistema di pesi aggiuntivo, il τ sarebbe stato pari a 0,0022, mentre se avessimo considerato come pesi la frequenza di ogni singola forma sul totale di forme utilizzate ($nterm_j / \sum_j nterm_j$) il τ sarebbe stato pari a 0,0016. Il valore dell'indice di predittività della nostra analisi è un primo segnale che il sistema di pesi introdotto migliora i risultati dell'analisi.

Nella tavola 1 è riportata la percentuale spiegata della struttura di dipendenza tra il "chi siamo" e il "chi cerchiamo", dai primi 10 autovalori. La percentuale del τ di Goodman-Kruskal spiegata dal primo piano fattoriale è pari al 13,4%; tale valore, sebbene apparentemente basso, è in realtà significativo considerando che ogni forma singolarmente spiega solo lo 0,22%.

autovalore	% spiegata	% cumulata
λ_1	8,3	8,3
λ_2	5,1	13,4
λ_3	4,5	17,9
λ_4	4,3	22,1
λ_5	3,7	25,8
λ_6	3,0	28,8
λ_7	2,8	31,6
λ_8	2,7	34,3
λ_9	2,7	37,0
λ_{10}	2,4	39,4

Tavola 1. Percentuale della struttura di dipendenza spiegata dai primi 10 autovalori

Per interpretare i risultati delle rappresentazioni grafiche ottenute, è opportuno ricordare che, nell'ANSC, occorre mantenere separate le rappresentazioni nei due spazi e valutarle congiuntamente alla luce della regola *e* del paragrafo 2.

Sul piano fattoriale dove sono rappresentate le forme del "chi siamo" (Fig. 1a), troviamo alla sinistra del primo asse quelle forme che, caratterizzate da coordinate negative e contributo assoluto alto (vedi Tav. 2), individuano aziende con attività tradizionali (*costruzione, raffinazione, elettrica, petrolifera, commercializzazione, etc.*). Diversamente, a destra si evidenziano forme che caratterizzano aziende operanti nei settori innovativi dell'informatica e dei servizi web (*hardware, software, networking, etc.*).

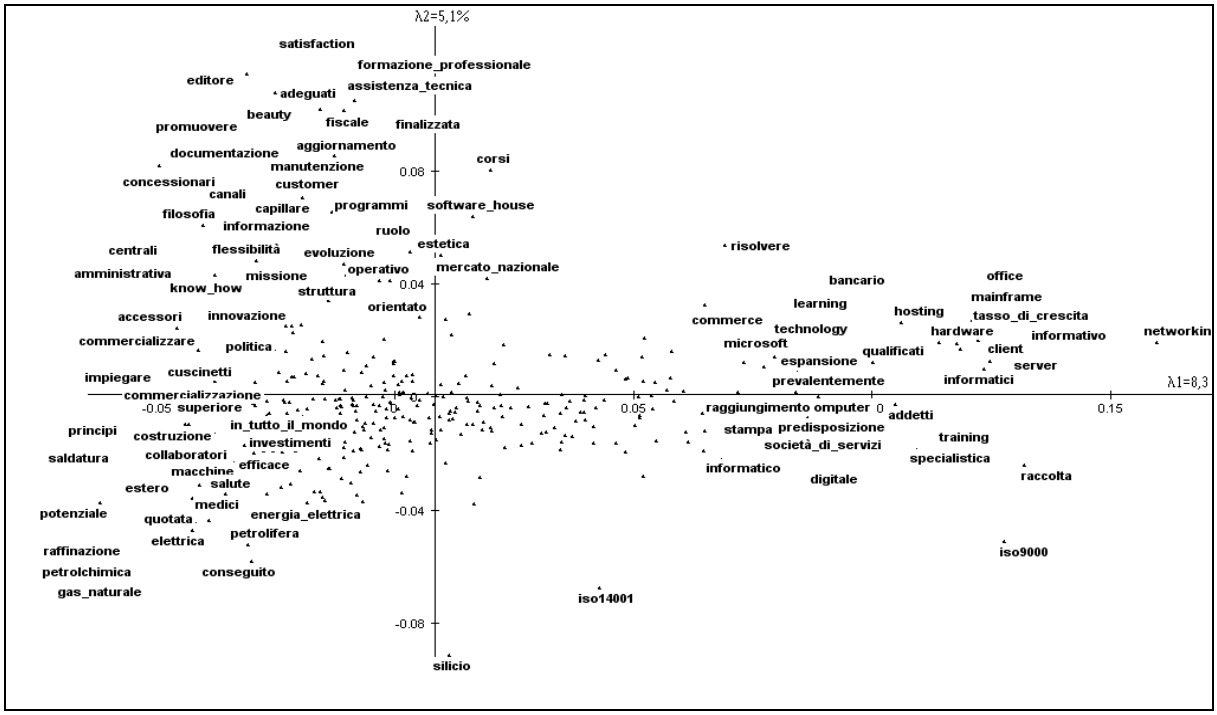

Figura 1a. ANSC primo piano fattoriale della rappresentazione delle parole del "chi siamo"

	forme	coordinata	contributo assoluto		**forme**	coordinata	contributo assoluto
1	potenziale	-0,062059434	4,04E-06	1	networking	0,159636049	1,08E-05
2	gas_naturale	-0,060309706	1,85E-06	2	raccolta	0,131902066	2,28E-06
3	raffinazione	-0,058871858	3,08E-06	3	iso9000	0,127858346	3,15E-06
4	petrolchimica	-0,058304691	3,02E-06	4	client	0,124662377	1,07E-05
5	impiegare	-0,053683353	6,63E-06	5	server	0,123559133	9,43E-06
6	documentazione	-0,049543869	3,58E-06	6	office	0,123492745	6,59E-06
7	accessori	-0,045803329	1,64E-06	7	tasso_di_crescita	0,122193708	3,11E-06
8	principi	-0,044076608	4,03E-06	8	mainframe	0,121007841	5,31E-06
9	collaboratori	-0,04398157	2,11E-06	9	informatici	0,118569509	1,22E-05
10	costruzione	-0,043271691	3,09E-06	10	informativo	0,117866419	8,90E-06
11	quotata	-0,042630399	2,36E-06	11	training	0,115401966	3,39E-06
12	elettrica	-0,042575964	1,92E-06	12	hardware	0,114178771	1,46E-05
13	petrolifera	-0,042111233	2,90E-06	13	specialistica	0,109334775	3,51E-06
14	politica	-0,041782517	3,23E-06	14	hosting	0,106078927	5,04E-06
15	commercializzare	-0,041414147	2,66E-06	15	addetti	0,104814214	2,29E-06
16	Estero	-0,040995457	2,34E-06	16	qualificati	0,100246884	5,51E-06
17	superiore	-0,040982496	2,17E-06	17	computer	0,088871937	7,68E-06
18	Filosofia	-0,040400157	3,30E-06	18	bancario	0,087550783	3,67E-06
19	Energia_elettrica	-0,03904156	1,92E-06	19	società_di_servizi	0,086850543	3,55E-06
20	saldatura	-0,039004015	2,70E-06	20	stampa	0,086635692	1,97E-06
21	in_tutto_il_mondo	-0,03816545	3,98E-06	21	learning	0,084430836	3,14E-06

Tavola 2. Coordinate e contributi sul primo asse fattoriale

Il secondo asse fattoriale (vedi Tav. 3 per le coordinate) contrappone dal basso verso l'alto le forme riferite ad aziende manifatturiere di grandi dimensioni con mercato internazionale (*quotata*, *estero*, *in tutto il mondo*, etc.) a quelle relative ad aziende di servizi (*fiscale*, *manutenzione*, *amministrative*, *concessionari*) che si collocano sul mercato nazionale (*nazionale*) e hanno una strategia di mercato orientata al cliente (*customer satisfaction)*.

	forme	coordinata	contributo assoluto		**forme**	coordinata	contributo assoluto
1	silicio	-0,091535061	1,55E-06	1	networking	0,159636049	1,08E-05
2	iso14001	-0,067615654	1,38E-06	2	raccolta	0,131902066	2,28E-06
3	gas_naturale	-0,059334586	1,79E-06	3	iso9000	0,127858346	3,15E-06
4	raffinazione	-0,055367507	2,72E-06	4	client	0,124662377	1,07E-05
5	petrolchimica	-0,054834102	2,67E-06	5	server	0,123559133	9,43E-06
6	iso9000	-0,051072138	5,03E-07	6	office	0,123492745	6,59E-06

	forme	coordinata	contributo assoluto		**forme**	coordinata	contributo assoluto
7	elettrica	-0,047323764	2,37E-06	7	tasso_di_crescita	0,122193708	3,11E-06
8	petrolifera	-0,04390463	3,15E-06	8	mainframe	0,121007841	5,31E-06
9	energia_elettrica	-0,043485566	2,38E-06	9	informatici	0,118569509	1,22E-05
10	dispositivi	-0,037331704	9,68E-07	10	informativo	0,117866419	8,90E-06
11	semiconduttori	-0,036716715	1,13E-06	11	training	0,115401966	3,39E-06
12	quotata	-0,035864847	1,67E-06	12	hardware	0,114178771	1,46E-05
13	elettronica	-0,033812638	1,93E-06	13	specialistica	0,109334775	3,51E-06
14	produttori	-0,033582915	1,64E-06	14	hosting	0,106078927	5,04E-06
15	ingegneria	-0,033459492	1,89E-06	15	addetti	0,104814214	2,29E-06
16	estero	-0,030920056	1,33E-06	16	qualificati	0,100246884	5,51E-06
17	acciaio	-0,026195629	6,14E-07	17	computer	0,088871937	7,68E-06
18	industria	-0,025275743	1,95E-06	18	bancario	0,087550783	3,67E-06
19	automobilistiche	-0,025060031	3,49E-07	19	società_di_servizi	0,086850543	3,55E-06

Tavola 3. Coordinate e contributi sul secondo asse fattoriale

Passando alla lettura della rappresentazione della variabile dipendente "chi cerchiamo" (Fig. 1b), è possibile leggere la dipendenza di tali forme da quelle precedenti, in termini di coseni di angoli. Come si evince dai risultati la relazione è chiaramente confermata.

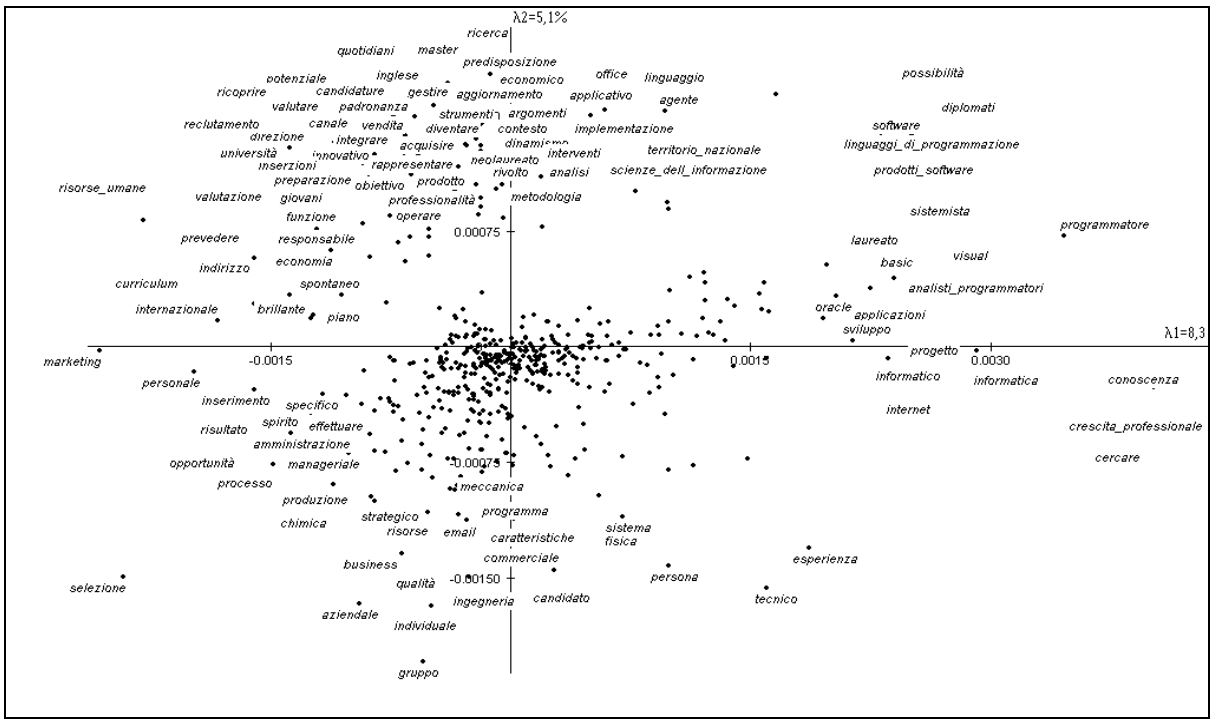

Figura 1b. ANSC primo piano fattoriale della rappresentazione delle parole del "chi cerchiamo"

In alto a sinistra dove, nel grafico precedente erano rappresentate le forme raffiguranti le imprese di servizi tradizionali, orientate al cliente, qui troviamo forme quali *master*, *inglese*,

aggiornamento, *preparazione*, che descrivono il profilo di un candidato che si adatta bene alla tipologia di azienda considerata. In basso a sinistra dove, nel grafico precedente trovavamo forme raffiguranti le multinazionali manifatturiere, qui troviamo forme quali *marketing*, *personale*, *amministrazione*, *produzione*, *commerciale*, che descrivono chiaramente la funzione aziendale che i candidati andranno a svolgere in quelle aziende.

In alto a destra dove, come visto, erano rappresentate le forme raffiguranti le imprese informatiche e le imprese di servizi innovativi, vi sono forme quali *programmatore*, *linguaggi di programmazione*, *sistemista*, *diplomato*, *laureato*, tipiche dello specifico profilo richiesto.

5. Conclusioni

In questo lavoro si è voluto proporre una strategia di analisi delle relazioni di dipendenza tra *corpora* testuali a partire dalla definizione di una particolare tabella lessicale del tipo forme/forme. L'utilizzo di tale matrice che trae origine da due *corpora* rilevati sulle stesse unità ma relativi a documenti distinti, consente evidentemente di generalizzare la strategia a casi studio piuttosto frequenti, come ad esempio quelli in cui i documenti sono relativi ad occasioni differenti. Si pensi, ad esempio, alle indagini ripetute nel tempo, dove evidentemente l'obiettivo può essere quello di valutare la relazione tra le risposte date dagli stessi soggetti in tempi differenti, sullo stesso argomento.

Un interessante sviluppo dell'analisi è quello di prevedere la possibilità di inserire anche le eventuali informazioni disponibili sulle unità che, invece, nell'approccio proposto non sono state considerate per considerare, invece la dipendenza del *corpus* del "chi cerchiamo" da quello del "chi siamo".

Si può, quindi, pensare di recuperare le informazioni sulle unità proiettandole in supplementare attraverso un operatore di proiezione che tenga conto del fatto che il sottospazio di riferimento è definito dalle distribuzioni condizionate delle forme testuali. Inoltre, si può pensare di sfruttare le potenzialità dell'analisi simbolica dei dati per rappresentare sia categorie di individui come oggetti simbolici definiti dalle modalità delle forme per ciascun *corpus*, sia le intensità della matrice di *co-presenza* come oggetti simbolici definiti dalle caratteristiche di ciascuna cella. Quest'ultimo obiettivo, evidentemente, richiede la definizione di particolari distribuzioni di probabilità delle variabili considerate, spostando l'attenzione sull'eventualità di introdurre una possibile modellizzazione dei dati.

Bibliografia

Balbi S. (1995). Non symmetrical correspondence analysis of textual data and confidence regions for graphical forms. In *JADT 1995*, vol (2) : 5-12

Goodman L.A. e Kruskal W.H. (1954). Measures of association for cross-classification. *J.A.S.A*, vol (49) : 732-764.

Lauro N. e D'Ambra L. (1984). L'analyse non symétrique des correspondances. In Diday *et al.* (Eds), *Data Analysis and Informatics*. NH : 433-446.

Lebart L., Salem A. e Berry L. (1991). Recent developments in the statistical processing of textual data. *Applied Stochastic Models and Data Analysis*, vol (7) : 47-62.

Lebart L., Salem A. e Berry L. (1998). *Exploring textual data*. Kluwer Academic Publisher.

Light R.J. e Margolin B.H. (1971). An analysis for variance for categorical data. *J.A.S.A.*, vol. (335/66) : 534-544.

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