Chi Cerca Cosa.
Esplorare le competenze richieste dalle imprese mediante tecniche di *mining*

Simona Balbi, Dario Bruzzese, Maria Gabriella Grassia[1]
Dipartimento di Matematica e Statistica, Università di Napoli "Federico II"

Riassunto: Obiettivo del presente lavoro è quello di proporre la costruzione di regole di associazione non simmetriche, in cui gli antecedenti sono rappresentati da modalità di variabili di tipo numerico, mentre i conseguenti sono espressi da variabili di tipo testuale. La non simmetria delle regole, unita alla particolare natura dei due insiemi di variabili, suggerisce inoltre una originale ridefinizione delle consuete misure di "supporto" e "confidenza" delle regole. Il lavoro si inserisce nell'ambito di un più ampio progetto volto ad approfondire le caratteristiche e le competenze richieste dalle aziende nei propri annunci di lavoro, in relazione alle caratteristiche strutturali dell'azienda.
Parole chiave: *Text mining*; Regole di associazione; Annunci di lavoro.

1. Introduzione

In questo lavoro si presenta uno studio del linguaggio che le imprese utilizzano sui siti Internet dedicati alla ricerca di nuovo personale. L'obiettivo è quello di identificare, attraverso lo studio dell'espressione verbale, le professionalità e le competenze richieste dalle imprese, in relazione alle loro caratteristiche strutturali (forma giuridica, localizzazione, settore di attività economica).

L'interesse è qui concentrato sulle aziende che scelgono come canale di reclutamento i siti Internet dedicati a domanda e offerta di lavoro (Balbi e Di Meglio, 2004; Grassia *et al.*, 2004) e sulla domanda di laureati. Lo strumento di analisi cui si

[1] Il presente lavoro è stato finanziato nell'ambito del progetto "Transizioni Università-Lavoro e valorizzazione delle competenze professionali dei laureati: modelli e metodi di analisi multidimensionali delle determinanti", cof. dal MIUR. Coordinatore nazionale è L. Fabbris, coordinatore del Gruppo di Napoli è S. Balbi. La nota è stata redatta da: S. Balbi per il Par. 2 , da D. Bruzzese per il Par. 3 e da M.G. Grassia per il Par. 4. I Paragrafi 1 e 5 sono frutto di elaborazioni comuni di tutti gli autori.

ricorre è tipico del *data mining*: le regole di associazione, ossia la costruzione, sulla base di ingenti quantità di dati disponibili, di regole logiche articolate nella forma di antecedenza/conseguenza, al fine di comprendere i legami esistenti all'interno della base di dati sottoposta ad analisi.

Il problema conoscitivo che si vuole affrontare presenta alcune peculiarità dal punto di vista metodologico. Prima di tutto, esiste una naturale gerarchia all'interno delle informazione di cui noi disponiamo sulle aziende. Si ipotizza che le parole utilizzate dalle aziende nei propri annunci di lavoro dipendano, in qualche modo, dalle caratteristiche strutturali dell'impresa. In questo senso si propone il ricorso a regole "non-simmetriche", ossia prodotte identificando *a priori* gli spazi degli antecedenti e dei conseguenti della regole stesse. Conseguenza della natura del fenomeno oggetto di studio è che lo spazio degli antecedenti abbia natura numerica, mentre quello dei conseguenti abbia natura testuale.

Nella nostra proposta originale si rivede, alla luce di queste peculiarità, anche la definizione delle tradizionali misure di "supporto" (riferito agli antecedenti numerici) e "confidenza" (riferito ai conseguenti testuali)[2].

Il lavoro si articola in un richiamo delle basi metodologiche di riferimento, che consente di porre il problema all'interno degli specifici contesti (Par. 2), nella formulazione della proposta metodologica (Par. 3) e nella sua implementazione al problema di analisi degli annunci di lavoro *on-line* (Par. 4). Le prospettive offerte da questa proposta concludono il lavoro (Par. 5).

2. Regole di associazione e pre-trattamento dei dati testuali

Le regole di associazione sono uno strumento di analisi il cui obiettivo è l'identificazione di relazioni presenti in basi di dati di grande dimensione, attraverso la formulazione di relazioni di dipendenza logica, del tipo *if-then*.

Nascono in un contesto applicativo molto specifico, quello della *market basket analysis* (Brijs *et al.*,1999) ma sono ormai diventate uno strumento largamente utilizzato in numerosi campi applicativi, dal *credit scoring* all'analisi testuale, dalla ricerca sociale alla medicina.

Sia $I = i_1, \ldots, i_j, \ldots, i_p$ un insieme di p differenti *item* e $T = t_1, \ldots, t_k, \ldots, t_N$ un insieme di N differenti transazioni tali che $t_k \subseteq I$ con k=1, …, N. Nel contesto originale della *market basket analysis* gli *item* rappresentano i diversi prodotti che possono

[2] Le espressioni "supporto" e "confidenza" derivano da una traduzione letterale dei corrispondenti termini di lingua anglosassone "*support*" e "*confidence*". Sebbene tale traduzione non esprima appieno il significato originariamente associato a tali misure, essa è ampiamente diffusa nella comunità di *data mining* e verrà quindi adoperata anche in questo lavoro.

essere acquistati e ciascuna transazione descrive l'insieme dei prodotti acquistati da un consumatore. Le transazioni sono codificate in una matrice *booleana* (*N, p*) in cui 1 indica la presenta del generico *item* i_j nella generica transazione t_k, mentre 0 indica la sua assenza.

Una regola di associazione (Agrawal *et al.*, 1993) *r* è allora definita dall'implicazione:

$$A \rightarrow C \text{ con } A, C \subseteq I, \quad A \cap C = 0$$

Il lato sinistro dell'associazione è detto "antecedente", o "corpo della regola", il lato destro è chiamato "conseguente", o "testa della regola". A ciascuna regola *r* è associata una coppia di misure, chiamate S_r (*supporto*) e C_r (*confidenza*), che ne specificano la qualità:

$$S_r = \frac{N_{A \cup C}}{N}$$

$$C_r = \frac{N_{A \cup C}}{N_A}$$

dove $N_{A \cup C}$ indica il numero di transazioni in cui sono contenuti contemporaneamente gli *item* della premessa e della conseguenza e N_A indica il numero di unità in cui sono contenuti gli *item* della premessa[3].

Gli algoritmi di estrazione delle regole sono numerosi; sono, in genere, di natura esaustiva, ossia ricercano tutte le possibili associazioni con un livello di supporto e di confidenza superiore ai valori minimi fissati dall'utente. Questo implica che ciascun *item* si può trovare indifferentemente nel corpo o nella testa di una regola, anche se nelle applicazioni reali spesso si dispone di informazioni esterne che rendono interpretabili solo relazioni che abbiano una determinata struttura di antecedenza/conseguenza.

A monte di qualsiasi analisi statistica che operi su dati testuali sono necessarie delle operazioni di pre-trattamento e di codifica, che consentano la cosiddetta "numerizzazione" del testo, con la minor perdita di informazione. Nei termini propri del *data mining*, occorre trasformare una base di dati non strutturata (documentaria) in una strutturata, per poi procedere al cosiddetto *parsing*, che

[3] Supporto e Confidenza misurano la frequenza relativa con cui gli *item* della premessa e della conseguenza sono inclusi, rispettivamente, nell'insieme di tutte le transazioni e nel sotto insieme di quelle che contengono gli *item* in premessa. Qualora *N* risulti sufficientemente elevato, condizione naturale in un contesto di *Data Mining*, esse sono spesso interpretate in termini di probabilità congiunta e condizionata.

consiste nella suddivisione del testo in unità elementari di analisi. Questa operazione rende necessarie numerose scelte che vanno dalla stessa accezione da dare al singolo termine (*parola*), all'introduzione di regole di esclusione/inclusione, all'identificazione di situazioni critiche, quali la presenza di polirematiche, le omografie, le sinonimie.

Una volta effettuate queste scelte, si procede alla fase di codifica, nella quale i testi vengono trasformati in vettori numerici (codifica *bag-of-words* e varianti). È, infatti, possibile associare una indicazione booleana di presenza/assenza della parola, oppure la sua frequenza nel testo, o ancora, pesare l'importanza della sua presenza con indici del tipo *term frequency-inverse document frequency* (Salton e Buckley, 1988)[4].

Quali che siano state le scelte di pre-trattamento e di codifica, la base documentaria è trasformata in una matrice [*documenti* x *parole*] pronta ad essere sottoposta a trattamenti statistici sviluppati per l'analisi di dati numerici. Occorre però non ignorare le scelte fatte in queste fasi preliminari e le loro implicazioni. Di questo si terrà conto nel seguito, nella riformulazione delle misure di supporto e confidenza, che tenga conto della natura di queste informazioni.

3. Regole di associazione non simmetriche

A fini applicativi, oltre alla matrice [*documenti x parole*] si utilizza una seconda fonte di informazione sulle caratteristiche strutturali delle aziende che offrono lavoro. Questa particolare configurazione ammette una rappresentazione più generale, estendibile ad altri contesti, in cui è possibile individuare la compresenza di due spazi differenti da analizzare congiuntamente: quello delle unità statistiche, su cui sono osservate un insieme di variabili numeriche che descrivono aspetti strutturali (spazio degli antecedenti), e quello delle *transazioni*[5] riferito invece a caratteristiche proprie dello specifico problema di analisi (spazio dei conseguenti). La relazione tra i due spazi è del tipo uno-a-molti dal momento che una singola unità statistica può dar luogo a più di una transazione.

In un contesto di *market basket analysis*, lo spazio degli individui è naturalmente descritto dalle informazioni socio-economiche degli acquirenti (informazioni disponibili grazie alle *fidelity cards*), mentre lo spazio delle transazioni

[4] Un'importante questione metodologica aperta riguarda la scelta del peso da attribuire alle parole all'interno delle procedure di *text mining* (per un esame critico delle diverse soluzioni proposte in letteratura, si veda Balbi, Misuraca, 2005).

[5] Il termine 'transazione', mutuato dall'originale contesto applicativo della *market basket analysis*, verrà per analogia esteso anche al contesto attuale.

è rappresentato dall'insieme dei prodotti acquistati da ciascun cliente nel corso del tempo. Nel problema conoscitivo che si sta affrontando, le transazioni corrispondono agli annunci di lavoro, gli *item* alle parole in essi contenuti e le unità statistiche si riferiscono alle aziende che offrono lavoro.

In un tale scenario, esiste una relazione di antecedenza logica tra i due spazi, dal momento che sono le caratteristiche strutturali delle unità a determinare il comportamento catturato dalle transazioni. Tuttavia, la soluzione spesso adottata consiste nel replicare le informazioni sulle unità statistiche, così da simulare una matrice rettangolare [*unità* x *variabili*] che contenga sia i dati strutturali che quelli 'transazionali', eliminando *ex post*, tutte quelle associazioni in cui la collocazione degli *item* nei due lati della regola non rispetta la struttura di antecedenza/conseguenza prevista.

Questo modo di procedere, oltre ad essere inefficiente dal punto di vista computazionale, non introduce una reale asimmetria nella metodologia di analisi che viene recuperata solo dopo che le associazioni sono state estratte. Inoltre l'approccio tradizionale non consente di associare a ciascun *item* la sua frequenza, cosa che, soprattutto in un contesto di analisi testuale, in cui gli *item* sono parole, risulta di primario interesse.

Allo scopo di definire un nuovo strumento, le *regole di associazione non simmetriche*, che assuma l'asimmetria come dimensione di analisi, verrà nel seguito introdotta la notazione in grado di catturare le peculiarità di ciascuno spazio.

Sia $X=x_1,\ldots,x_l,\ldots,x_n$ un insieme di n unità statistiche descritte da m variabili categoriche[6] e sia Ξ l'insieme delle modalità delle variabili osservate, di cardinalità M, di modo che $x_l \subset \Xi \ (l=1,..,n)$.

Inoltre, sia $Y= y_1,\ldots,y_h,\ldots,y_N$ un insieme di N transazioni definite a partire dall'insieme degli *item* $I= i_1,\ldots,i_j,\ldots,i_p$. A differenza di quanto accade nel tradizionale contesto delle regole di associazione, il generico elemento $y_{h,i}$ non esprime unicamente la presenza/assenza dell'item i_j ma indica il numero di volte in cui esso è presente nella transazione y_h.

Una regola di associazione non simmetrica $r_{X|Y}$ è allora definita dall'implicazione:

$$A \rightarrow c \qquad \text{con A} \subset \Xi \text{ , c} \subset \text{I.}$$

In questa sede saranno considerate solo regole *many-to-one*, regole cioè con un numero di modalità in premessa potenzialmente pari a m ma un solo *item* in conseguenza; questa scelta deriva da alcune difficoltà metodologiche nel considerare regole non simmetriche *many-to-many*, ma è coerente con una prassi ormai consolidata in un contesto di *data mining* .

---

[6] Qualora le variabili siano di tipo continuo è necessario procedere ad una fase di categorizzazione.

Dal momento che la premessa e la conseguenza di una regola non simmetrica sono definite a partire da due spazi differenti (quello delle unità e quello delle transazioni) si rende necessario specificare i valori di supporto e di confidenza in modo nuovo rispetto a quanto tradizionalmente viene fatto per le regole di associazione.

Indicando con X_S ($X_S \subseteq X$) l'insieme delle unità statistiche le cui transazioni contengono almeno una volta gli *item* in S ($S \subseteq I$), il supporto di una regola non simmetrica $r_{X|Y}$ ($S_{R_{X|Y}}$) è ottenuto come frequenza relativa delle unità statistiche caratterizzate dalle modalità contenute in A e che contemporaneamente appartengono all'insieme X_c:

$$S_{R_{X|Y}} = \frac{\#\{x_l \mid A \subseteq x_l \cap x_l \in X_C\}}{n}.$$

Indicando con Y_R ($Y_R \subseteq Y$) l'insieme delle transazioni riferite alle unità statistiche descritte dalle modalità in R ($R \subset \Xi$), la confidenza della regola $r_{X|Y}$ ($C_{R_{X|Y}}$) è ottenuta rapportando la frequenza *dell'item c* nelle transazioni che appartengono a Y_A alla frequenza complessiva dell'item *c*, secondo l'espressione:

$$C_{R_{X|Y}} = \frac{\sum_{h=1}^{N} y_{h,c} \mid y_h \in Y_A}{\sum_{h=1}^{N} y_{h,c}}.$$

Le misure introdotte presentano alcuni elementi di novità su cui è opportuno soffermarsi. In primo luogo, il supporto misura il peso della regola in termini di unità statistiche che la soddisfano e non in termini di transazioni, come invece accade nelle regole tradizionali. In questo modo risultava impossibile verificare se la regola agiva effettivamente su molte unità statistiche o su poche unità cui però erano associate numerose transazioni. In un contesto di *market basket analysis*, *e quindi di grande distribuzione,* questo valutazione è invece di primaria importanza dal momento che non risulta conveniente adottare specifiche azioni di *marketing* laddove il *target* potenziale è rappresentato da un esiguo numero di consumatori, qualunque sia il numero di transazioni coinvolte.

Analogamente nella caratterizzazione delle offerte di lavoro è indispensabile valutare se l'uso di particolari termini risulta essere specifico di poche aziende con caratteristiche simili, che ne fanno però un uso frequente, oppure se l'associazione è realmente presente in un numero di aziende elevato.

E' però nella definizione di $C_{R_{X|Y}}$ che si evidenzia la maggiore differenza rispetto al contesto tradizionale. Innanzitutto, essa tiene conto della frequenza dell'*item* e non solo della sua presenza/assenza. Inoltre il fattore di normalizzazione che compare a denominatore non è riferito agli elementi della premessa dal momento che questi sono definiti a partire da spazi differenti e non rapportabili. Essa non ammette, quindi, una interpretazione in termini di probabilità condizionata, come invece accade nelle regole di associazione, ma esprime in quale misura l'*item* in conseguenza caratterizza le unità statistiche descritte dalla premessa.

L'utilizzo congiunto delle due misure così definite consente di ottenere una informazione articolata sulla natura dell'associazione: una regola $r_{X|Y}$, con supporto pari ad *s* e confidenza pari a *c*, indica che quella particolare associazione tra caratteristiche strutturali e 'transazionali' è presente in una frazione pari ad *s* del totale delle unità statistiche, mentre, in una proporzione pari a *c*, l'unità statistica è caratterizzata dalle modalità in premessa sul numero complessivo di volte in cui, in una transazione, è presente *l'item* della conseguenza.

4. Un'analisi del mercato del lavoro *on line*

Le regole di associazione non simmetriche sono utilizzate come strumento di analisi per comprendere le caratteristiche e le competenze richieste dalle aziende negli annunci di lavoro, in relazione alle proprie caratteristiche strutturali.

L'interesse dell'applicazione è concentrato sulle competenze richieste dalle aziende che scelgono come canale di reclutamento i siti Internet dedicati a domanda e offerta di lavoro. A tal fine, sono state rilevati 164 annunci di lavoro disponibili sul sito *www.carrierain.it*[7] nella sezione "Ricerca di Lavoro", associando ad essi le informazioni relative all'assetto societario dell'azienda richiedente, al settore di attività, alla localizzazione geografica presenti, nella sezione "Profili aziendali" di tale sito.

Il *corpus* testuale, costituito dalla base dei dati non strutturati degli annunci, è stato in primo luogo sottoposto ad una procedura di *tagging* "semantico" manuale, per individuare i contesti locali delle forme ambigue e quindi procedere ad una loro etichettatura (ad esempio la parola chimica è stata etichettata con il *tag_lau* se relativa alla laurea richiesta, altrimenti con il *tag_set* se riferita al settore di attività).

[7] Tale portale è stato creato dall'Associazione Mercurius di Torino per agevolare l'incontro tra la domanda e l'offerta di lavoro. Il Network Mercurius è costituito complessivamente da cinque portali tematici che riguardano la formazione, il lavoro interinale, la ricerca di lavoro.

Successivamente il *corpus* è stato normalizzato e lessicalizzato con procedure automatiche, al fine di evidenziare la presenza di poliformi e poliriematiche di interesse per l'analisi. 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 ulteriore lemmatizzazione interna (Lebart *et al*., 1998), necessaria a fondere le forme con differente significante e medesimo significato.

Dopo la fase di pre-trattamento e codifica, si è costruita la tabella lessicale [*annuncixparole*] in cui le righe sono i 164 annunci, ognuno pubblicato da una corrispondente azienda, e le colonne sono costituite da 530 forme testuali e il cui termine generico è la frequenza di una forma in un annuncio.

Anche le informazioni strutturali delle aziende sono state ricodificate, prendendo spunto dalla classificazione ATECO 2002 (ISTAT, 2002) utilizzata dalle C.C.I.A.A, per la classificazione delle aziende nel "Registro delle imprese". La tabella ottenuta è una matrice *individuixvariabili* in codifica ridotta (164,3), in cui le righe sono le aziende e le colonne sono le tre variabili considerate (tipologia societaria, sede, attività), rispettivamente con le seguenti modalità:

- tipologia societaria (s.p.a., s.r.l., Soc. Persone ed altro);
- sede (Nord Est, Nord Ovest, Centro, Sud ed Isole);
- attività (Attività culturali, Attività professionali, Commercio ingrosso e dettaglio, Comunicazione e servizi connessi, Industrie pesanti, Industrie alimentari e chimiche, Informatica ed attività connesse, Trasporti, logistica e telecomunicazioni).

È da considerare che tra gli annunci e le aziende vi è una corrispondenza biunivoca.

Figura 1. *Le matrici utilizzate ai fini dell'analisi*

S (164,3)

	Tipologia	Sede	Attività
Azienda1	***1***	***3***	***7***
Azienda2	***2***	***3***	***4***
…	**….**	**….**	**…**
Azienda164	***3***	***1***	***1***

T (164,530)

	Abilità	Access	……	Vendere	Windows
Annuncio1	***4***	***1***	**…**	***5***	***1***
Annuncio2	***0***	***1***	**…**	***2***	***1***
…	**….**		**…**	**…**	**…**
Annuncio164	***3***	***0***	**…**	***0***	***0***

Le regole sono state costruite estraendo l'antecedente dalla matrice S di sinistra (*item* sulla struttura delle imprese) e il conseguente dalla matrice T di destra (*item* testuali).

Per l'estrazione delle regole si è utilizzato una variante dell'algoritmo proposto in Balbi *et al.* (2003), modificato in modo da introdurre nel calcolo della confidenza un sistema di ponderazione basato sulla frequenza delle parole in un annuncio.

Complessivamente, fissando un valore minimo di confidenza pari a 0,5 ed un valore minimo di supporto pari a 5 in valore assoluto e 0,03 in valore relativo, si ottengono 635 regole.

Prima di analizzarle, è necessario premettere che il campione è costituito per il 53,7% da aziende appartenenti all'Informatica e alla produzione di *software*. La particolare struttura del campione e l'obiettivo di semplificazione espositiva, ci induce, nel seguito, a commentare in dettaglio proprio le regole estratte in questo specifico settore.

Il commento delle regole avverrà partendo dal gruppo di regole con numero maggiore di *item* nell'antecedenza, facendo seguire i gruppi che a parità di *item* presentano valori di confidenza più elevati. Nelle tabelle seguenti viene indicato il supporto (*Sup*) e la confidenza (*Conf*) di ciascun gruppo di regole.

Il primo gruppo di regole ha, dunque, tre *item* nell' antecedenza: la modalità s.*r.l.* per la tipologia societaria; la modalità *Nord Ovest* per la localizzazione geografica e il settore dell'*informatica* per l'attività svolta. In relazioni a queste caratteristiche aziendali, l'annuncio sembra contenere le seguenti informazioni (Tab. 1):

- presentazione della propria azienda come organizzazione sul territorio nazionale;
- richiesta specifica al candidato di realizzare un applicativo;
- richiesta della conoscenza della lingua inglese;
- offerta di un'attività legata ad un periodo determinato.

Tabella 1: *Le regole con tre modalità nell'antecedenza*

If	**Srl**	*and*	**Nord Ovest**	*and*	**Informatica ed attività connesse**	*then*	**Sup**	**Conf**
						organizzazione	5	0,50
						territorio_nazionale	5	0,63
						realizzare	6	0,86
						applicativo	5	0,50
						inglese	6	0,50
						periodo	6	0,50

Il secondo gruppo di regole ha come antecedenti due *item*: la modalità s.*r.l.* per tipologia societaria ed il settore *dell'informatica* per l'attività svolta (Tab.2).

Le regole mostrano come, indipendentemente dalla localizzazione geografica, una società a responsabilità limitata operante nel settore dell'informatica abbia idee molto chiare sulla persona da assumere: il candidato deve essere *neolaureato* in *matematica* o *fisica*, deve conoscere i linguaggi di programmazione, in particolare

Oracle, deve conoscere la lingua *inglese*, deve essere flessibile. Al candidato si offre un inquadramento come *collaboratore*, oppure uno *stage* e *training on the job*.

Tabella 2: *Le regole con due modalità nell'antecedenza (s.r.l., Settore Informatica)*

If	**Srl**	*and*	**Informatica ed attività connesse**	*then*	**Sup**	**Conf**
				matematica	6	0,75
				fisica	8	0,73
				linguaggi_di_programmazione	5	0,71
				oracle	8	0,53
				inglese	7	0,58
				neolaureato	9	0,53
				flessibilità	6	0,55
				inquadramento	5	0,56
				collaboratore	5	0,54
				corso	13	0,68
				training	5	0,55
				on_the_job	6	0,75
				stage	8	0,53

Tabella 3: *Le regole con due modalità nell'antecedenza (s.p.a., Settore Informatica)*

If	**Spa**	*and*	**Informatica ed attività connesse**	*then*	**Sup**	**Conf**
				obiettivo	7	0,69
				possibilità	8	0,62
				inserire	9	0,72
				lavoro	9	0,64
				ambiente	6	0,58
				stimolante	5	0,68
				crescita_professionale	6	0,61
				giovani	9	0,61
				laureato	9	0,57
				tecnico	5	0,54
				commerciale	5	0,57
				aziendale	7	0,56

Il terzo gruppo di regole ha sempre due *item* come antecedenza, presentando la modalità *s.p.a.* per tipologia societaria e sempre il settore *dell'informatica* per l'attività svolta (Tab.3).

In questo caso le regole, con confidenza mediamente inferiore al gruppo precedente, mostrano una minore attenzione alle caratteristiche del candidato che deve essere *giovane*, *laureato* ma anche *diplomato* con conoscenze tecniche, ed una maggiore enfasi sul tipo di lavoro (*commerciale*, *aziendale*) e sull'*ambiente stimolante* (Tab.3).

Le differenze emerse attraverso l'utilizzo delle regole sono avvalorate da una analisi più approfondita sulla tipologia di settore. Tale analisi mostra come le aziende con forma societaria s.r.l. siano prevalentemente *Software House*, giustificando, in tal senso, il loro maggior interesse circa le competenze possedute dai candidati. Le aziende con forma societaria s.p.a. sono, invece, prevalentemente aziende produttrici o distributrici di *hardware* (anche multinazionali con sede in Italia) e per tale motivo mostrano maggiore interesse verso figure non altamente specializzate da inserire nella propria rete commerciale.

5. Conclusioni

In questo lavoro si è proposto l'uso di regole non simmetriche con l'obiettivo di evidenziare associazioni privilegiate tra le informazioni strutturali delle aziende che offrono lavoro e le forme testuali utilizzate nei loro annunci.

La proposta metodologica sfrutta i vantaggi tipici delle regole di associazione, e contemporaneamente consente di affrontare le principali questioni che caratterizzano lo specifico problema conoscitivo: conoscenza a priori di una struttura di antecedenza/conseguenza tra i due spazi di analisi (quello delle aziende e quello degli annunci); l'introduzione del peso di ciascuna forma testuale all'interno dello strumento di analisi attraverso la ridefinizione della confidenza in termini di frequenza e non di presenza/assenza.

L'analisi dei risultati ha mostrato come le regole non simmetriche siano efficaci nell'estrarre da ingenti quantità di dati l'informazione strutturale in essi contenuta.

L'applicazione ha mostrato gli aspetti metodologici innovativi, estendibili sia a basi di dati più ampie di quella analizzata, sia a dati provenienti da contesti differenti. Tali innovazioni, particolarmente rilevanti in un contesto testuale, riguardano tre aspetti fondamentali:

- la differente natura degli spazi dell'antecedenza e della conseguenza, numerico il primo e testuale il secondo;
- il tipo di matrice utilizzata per lo spazio della conseguenza, non più di presenza/assenza, ma di *frequenza* con termine generico pari al numero di volte in cui una forma risulta essere presente in un annuncio;

- la ridefinizione delle tradizionali misure di "supporto" e "confidenza".

Il limite principale al loro utilizzo risiede tuttavia nell'impossibilità di generare regole *many-to-many*; l'uso della frequenza nella definizione di confidenza, infatti, non consente di associare a regole con più di un *item* in conseguenza, un valore consistente. A questa problematica saranno rivolti gli ulteriori sviluppi di ricerca nella direzione di generare regole caratterizzate da una disgiunzione logica, piuttosto che da una congiunzione, all'interno della conseguenza.

Riferimenti bibliografici

AGRAWAL R., IMIELINSKI T., SWAMI A. (1993) Mining Associations between Sets of Items in Massive Databases, *Proc. of the ACM-SIGMOD* 1993 Int'l Conference on Management of Data, Washington D.C.: 207-216.

BALBI S., BRUZZESE D., SCEPI G. (2003) Analyze the E-commerce impact on business to consumer transactions through Non Symmetrical Association rules, *Statistica Applicata*, **15(1)**: 131-148

BALBI S., DI MEGLIO E. (2004) Una strategia di Text Mining basata su regole di associazione. In: AURELI CUTILLO E., BOLASCO S. (a cura di), *Applicazioni di analisi statistica dei dati testuali*, Casa Editrice Università, Roma: 29-40

BALBI S., MISURACA M. (2005) Weighting systems for text mining, Abstracts from *GfKl2005 Data and information analysis to knowledge engineering*, Univ. Magdeburg (D), 61

BRIJS T., SWINNEN G., VANHOOF K., WETS G. (1999) The Use of Association Rules for Product Assortment Decisions: a Case Study, *Proc. of the Fifth International Conference on Knowledge Discovery and Data Mining, San Diego (USA), August 15-18*: 254-260

GRASSIA G., MISURACA M., SCEPI G. (2004) Relazioni non simmetriche tra corpora. In: PURNELLE G. *et al.* (eds.): *Le pois des mots. Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles. Vol. 1*, UCL Presses, Louvain (B): 524–532

LEBART L., SALEM A., BERRY L. (1998) *Exploring Textual Data*, Kluwer Academic Publisher

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

Who is Looking for What? Exploring Competences Required by Companies using Mining Techniques

Summary. *In this paper we propose the building and use of non symmetrical association rules, when antecedents are modalities of numerical variables and consequents are textual variables. The asymmetry of rules and the peculiar nature of the two variable sets suggest a new definition of the "support" and "confidence" usual measures for rules. The paper is part of a wider project aimed at studying characteristics and competences required by firms in their job advertisements, related with their structural characteristics.*
Keywords*. Text mining, Association rules , Job advertisements*