Università degli Studi di Napoli
Federico II

La scelta delle unità e dei relativi pesi
nel processo di estrazione dell'informazione da dati testuali

Tesi di Dottorato in
Statistica

XIX ciclo

Candidato
Giorgio Infante

Coordinatore
Natale Carlo Lauro

La scelta delle unità e dei relativi pesi nel processo di estrazione dell'informazione da dati testuali

Napoli, novembre 2007

Ringraziamenti

Grazie alla Prof.ssa Simona Balbi per la costante disponibilità e cortesia, sia umana che professionale, avute nei miei confronti; particolarmente preziose sono risultate le sue indicazioni e le aperture di ricerca, con le quali sono stato costantemente guidato nell'elaborazione di questa tesi.

Grazie al prof. Carlo Lauro, coordinatore del dottorato, per essere il punto di riferimento mio e di tutti gli studenti del dipartimento di matematica e statistica.

Grazie ad Emilio di Meglio per avermi ben consigliato e confortato all'inizio di questo percorso.

Grazie a Michelangelo Misuraca e Pino Giordano i cui suggerimenti scientifici e pratici mi sono stati indispensabili per tutto il dottorato e per la conclusione di questo lavoro.

Grazie ad Antonio, Elvira, Rosaria, Ida, Valerio e Giorgio che si sono dimostrati amici prima che colleghi.

Grazie a papà, mamma, Stefano, Fulvio ed Enrico che ci sono sempre quando ne ho bisogno.

Grazie a Lalla che è sempre stata al mio fianco. Senza d lei tutto questo sarebbe stato impossibile.

INDICE

Indice delle figure

Indice delle tabelle

Introduzione

Da tempo sono stati introdotti algoritmi e tecnologie per estrarre conoscenza da dati numerici, mentre una sempre maggior importanza assume l'estrazione di conoscenza da dati testuali.

Obiettivo del Text Mining (TM) è proprio l'estrazione di informazione, rilevante per il ricercatore, da dati non strutturati che risiedono in file di testo.

Si pensi alla necessità di estrarre conoscenza da mail, pagine internet, forum di discussione, call center, risposte a domande aperte, verbali, brevetti, diagnosi, ecc.; l'analisi di tali testi richiede strumenti efficienti ed automatici, in mancanza la maggior parte delle comunicazioni non strutturate non è nemmeno letta.

Utilizzare approcci manuali non è pratico, richiede molto tempo, può essere molto costoso e i risultati possono essere inconsistenti.

Lo sviluppo tecnologico degli ultimi anni ha consentito una forte evoluzione delle tecniche e delle metodologie utilizzate in questo campo in quanto ha permesso l'implementazione di algoritmi statistici e di machine learning a forte componente computazionale.

D'altronde le tecniche statistiche su dati qualitativi, basandosi su concetti di distanza e dissimilarità, mal si adattano all'analisi di dati non strutturati. Diviene quindi essenziale la fase di codifica del dato stesso. Ma se nella maggior parte dei casi questo avviene calcolando le frequenze delle parole nei documenti, è utile considerare altre opzioni che maggiormente collimano con gli obiettivi della ricerca.

Il ruolo della codifica del testo assume pertanto una posizione sempre più rilevante nel text mining, dove per codifica si intende tutto il processo che va dal documento alla matrice dei dati da analizzare.

In sostanza la codifica dei dati può estrinsecarsi in due fasi: la scelta delle unità di analisi e il sistema di pesi da adottare. Infatti, una volta attuate la scelte, i dati testuali diventano dati numerici e come tali possono essere sottoposti a qualsiasi tipo di analisi statistica.

Tali scelte però non sono assimilabili alla cosiddetta fase di pulizia dei dati tipica di ogni ricerca statistica su dati qualitativi. La definizione di unità e pesi nell'ambito del TM è da considerarsi parte integrante dell'analisi statistica. Da esse deriveranno i risultati dell'analisi, la loro interpretabilità e soprattutto la loro inerenza con gli obiettivi prefissati. È chiaro quindi che la loro determinazione deve avvenire immediatamente dopo la definizione degli obiettivi della ricerca ed insieme alla scelta delle tecniche statistiche da utilizzare.

In tale ottica assume pertanto rilevanza il ruolo dello statistico: la sua presenza, in collaborazione con gli altri "esperti"

(prevalentemente linguisti e informatici), diventa ovviamente necessaria anche nella fase di codifica del testo.

STRUTTURA DELLA TESI

Il primo capitolo della tesi descrive il processo di text mining con particolare attenzione alla fase di codifica. Sono, in particolare, illustrate le principali tecniche statistiche tradizionalmente applicate ai dati testuali.

Gli altri capitoli della tesi sono collegati a contributi originali, metodologici o applicativi, già proposti in convegni nazionali o internazionali. In ognuno di essi viene evidenziato come le scelte effettuate in tema di unità e sistema di pesi dipendano dagli obiettivi dell'analisi e dalle fonti analizzate.

Il secondo capitolo illustra come le tecniche tipicamente utilizzate nell'analisi dei dati testuali possano essere indicate per obiettivi di rassegna di un determinato argomento.

Il terzo capitolo pone il problema della text classification (TC). La TC ha l'obiettivo di associare documenti in linguaggio naturale ad un insieme di categorie predefinite. Dopo una breve rassegna sui metodi più conosciuti in letteratura è proposta una strategia di TC basata sulle regole di associazione con l'uso di un particolare sistema di pesi: il TF-IDF. La validità della strategia è mostrata con un'applicazione sulle competenze offerte dai corsi di laurea delle università italiane.

Il quarto capitolo si pone il problema delle relazioni tra due tabelle lessicali che non presentano né gli stessi documenti (righe), né

gli stessi termini (colonne). La proposta metodologica è quella di codificare i dati costruendo oggetti simbolici espressi in forma modale, ovvero definiti dalle distribuzioni di frequenza dei termini utilizzati. Il calcolo di misure di dissimilarità sugli oggetti simbolici consente il confronto tra *corpora*. Il capitolo si chiude con un'applicazione sulle competenze offerte e domandate nel terzo settore.

Il quinto capitolo tratta infine del ruolo che possono avere i dati testuali in ausilio a dati numerici. Nella fattispecie viene proposto l'utilizzo di dati testuali come informazione esterna in un'analisi non simmetrica delle preferenze. In questo caso l'informazione testuale, proveniente da domande a risposta aperta, è codificata in presenza/assenza in considerazione della brevità dei documenti ed arricchisce le informazioni ottenute con la sola tecnica quantitativa. Un'applicazione al mercato degli orologi evidenzia le potenzialità della strategia adottata.

Le applicazioni del terzo e quarto capitolo si inseriscono nell'ambito del progetto di ricerca a interesse nazionale (PRIN 2005), coordinato dal Prof. Luigi Fabbris, *"Il mercato delle competenze: metodi statistici per il confronto e l'analisi multidimensionale delle figure professionali offerte e domandate nel terzo settore"* finalizzato alla valutazione della domanda e offerta di competenze del mercato del lavoro in Italia.

La tesi si conclude indicando nuove prospettive di ricerca, nella convinzione che il text mining assumerà un ruolo sempre più rilevante

nel trattamento automatico dell'informazione e, di conseguenza, la necessità di definire opportune strategie di strutturazione delle basi di dati documentarie (di "codifica" in termini statistici) sarà sempre più sentita.

CAPITOLO 1

1 Il ruolo della statistica nel text mining

1.1 L'analisi dei dati testuali

Il linguaggio naturale è la facoltà, esclusiva del genere umano, di esprimere sensazioni, sentimenti, riflessioni, giudizi; di narrare fatti o situazioni; di descrivere aspetti della realtà mediante un medium che sia espressione di un determinato livello comunicativo.

La lingua è lo strumento di comunicazione maggiormente utilizzato ed è costituita da un complesso sistema di segni organizzati all'interno di una struttura.

Inizialmente gli studi sul linguaggio naturale erano territorio di ricerca di linguisti, sociologi e psicologi. È solo in seguito all'evoluzione tecnologica che anche informatici e statistici hanno iniziato ad interessarsi al linguaggio naturale fino a produrre l'analisi automatica dei testi e la statistica testuale(Lebart, et al., 1988)

Negli ultimi anni la crescita esponenziale della disponibilità di documenti in formato elettronico ha ulteriormente rivoluzionato criteri e tecniche in quest'ambito trasformandolo sempre di più in un'attività multidisciplinare.

Di fatto, come sostenuto da Bolasco (Bolasco, 1995), gli studi quantitativi su tali tipi di dati hanno subito una evoluzione spostandosi da una logica di tipo linguistico (Guiraud, 1954) (sviluppata fino agli anni sessanta del novecento), il cui obiettivo era la ricerca delle regolarità nella lingua, ad una di tipo lessicale (Muller, 1977) (intorno agli anni Settanta del secolo scorso) caratterizzata dai primi studi stilometrici sulle opere di autori per conoscere gli usi dei lemmi nei loro significati ed accezioni. In questo periodo vengono anche proposte statistiche elementari per confrontare testi di autori diversi.

Si arriva così con gli anni ottanta ad analisi di tipo testuale(Lebart, et al., 1988), basate su tecniche fattoriali su tabelle lessicali e agli anni novanta dove si diffonde la logica lessico-testuale, nella quale si riconosce l'ambiguità delle forme grafiche e si propongono soluzioni per integrare la fase lessicale con quella testuale.

Infine negli ultimissimi anni sono cambiati sia i soggetti che si occupano di ricerca in questo campo (alle università e ai centri di ricerca si aggiungono le aziende), sia le tecniche potenziate ed innovate dall’ausilio di strumenti e software sempre più potenti.

L’eccesso di informazione testuale disponibile sia nelle organizzazioni private che in quelle pubbliche, dovuta allo sviluppo della tecnologia e di Internet spinge a trovare metodi e tecniche, attraverso l’implementazione di algoritmi statistici e di machine learning a forte componente computazionale, per selezionare informazione rilevante da grandi basi di dati.

Obiettivo del Text Mining (TM) è proprio l’estrazione di informazione rilevante da dati non strutturati che risiedono in file di testo.

1.2 Il processo di text mining

Il processo di text mining si articola in tre fasi:

- la definizione degli obiettivi e l’acquisizione dei documenti;
- la codifica dei dati;
- l’estrazione dell’informazione.

Ma se la prima fase è per sua natura di competenza del soggetto interessato agli aspetti sostantivi e la terza fase riguarda

certamente lo statistico, in quanto concerne la scelta della tecnica più idonea per il raggiungimento degli obiettivi prefissati, è la seconda fase ad assumere una posizione sempre più rilevante dal punto di vista dello statistico, dove per codifica è qui intesa in senso lato , come l'intero processo che va dal documento alla matrice dei dati da analizzare.

In sostanza la codifica può estrinsecarsi, a sua volta, in due momenti:

- la scelta delle unità di analisi;
- il sistema di pesi da adottare.

Tali scelte però non sono assimilabili alla cosiddetta fase di pulizia dei dati tipica di ogni analisi statistica su dati di tipo numerico. La definizione di unità e pesi nell'ambito del TM è da considerarsi parte integrante dell'analisi stessa. Da esse infatti deriveranno i risultati, la loro interpretabilità e soprattutto la loro inerenza con gli obiettivi prefissati. È chiaro quindi che la determinazione delle unità di analisi e dei pesi deve avvenire immediatamente dopo la definizione degli obiettivi della ricerca ed insieme alla scelta delle tecniche statistiche da utilizzare.

In tale ottica assume pertanto rilevanza il ruolo dello statistico: la sua presenza, in collaborazione con gli altri "esperti" (prevalentemente linguisti e informatici) diventa ovviamente necessaria anche nella fase di codifica del testo.

1.3 La scelta delle unità

I metodi statistici si basano sempre su misure e conteggi realizzati su oggetti da confrontare. Nel caso di dati presenti in file di testo la rilevazione degli oggetti non è immediata. È quindi necessario definire regole che permettano di isolare dal corpus[1] le unità da analizzare. L'operazione che consente di individuare questi oggetti è detta segmentazione del testo.

In pratica vengono individuati i caratteri cosiddetti delimitatori (spazio, punteggiatura, ecc.) e non delimitatori (tutti gli altri). Una sequenza di caratteri non delimitatori alle cui due estremità sono presenti caratteri delimitatori è una occorrenza. Due sequenze identiche di caratteri non delimitatori costituiscono due occorrenze di una stessa forma. L'insieme di tutte le forme di un testo costituisce il suo vocabolario (*V*) mentre il numero totale delle occorrenze determina la dimensione o lunghezza del corpus (*T*).

Alla fase di segmentazione del testo succede la fase di numerizzazione dello stesso, attraverso la quale ad ogni occorrenza si associa un codice numerico. Se una stessa sequenza di caratteri ricorre più volte nel testo, ad essa si troverà associato sempre lo stesso

[1] Raccolta coerente di materiale testuale omogenea rispetto all'oggetto di interesse.

codice. Ad ogni codice, poi, viene associato l'insieme dei suoi indirizzi, cioè delle sue collocazioni nel testo.

1.3.1 Le forme grafiche

Una parola è, convenzionalmente, una forma grafica, ossia una sequenza di caratteri appartenenti ad un alfabeto predefinito, delimitata da due delimitatori (o separatori).

La forma grafica è l'unità elementare del linguaggio e come tale può essere considerata unità statistica.

Le forme grafiche non consentono però di individuare, ad esempio, la presenza di sinonimi o antonimi, in particolare questi ultimi qualora espressi per anteposizione alla forma di una particella con valore di negazione.

1.3.2 Le unità minimali di senso

Reinert (Reinert, 1988) suggerisce l'uso delle unità minimali di senso intese come sequenze di caratteri aventi significato autonomo. L'obiettivo è quello di ridurre le ambiguità tra forme omonime e poliformi minimizzando il "disturbo" dovuto a quelle forme che presentano un contenuto informativo ridotto o nullo. In sostanza

vengono scelte quali unità sia i segmenti di testo[2] (come ad esempio le polirematiche, gruppi di parole che hanno un significato unitario non desumibile dalle singole parole che la compongono, come ad esempio: "*dato di fatto*", "*carta di credito*") sia le forme grafiche con esclusione delle parole vuote[3].

1.3.3 Le forme lemmatizzate

Partendo dall'analisi delle forme grafiche presenti in un testo, la prima operazione di lemmatizzazione consiste nell'attribuire ad esse la propria categoria grammaticale, mentre la seconda operazione consiste nel riportare le singole parole al loro lemma. Questo secondo passo si traduce nel fondere di fatto tutte le forme flesse di uno stesso lemma, perdendo così le variazioni di linguaggio anche quando queste fossero semanticamente significative.

Bolasco propone "una lemmatizzazione ragionata", che operi cioè fusioni di termini solo quando questa operazione migliora le condizioni dell'analisi, senza toccare l'informazione contenuta nel corpus.

[2] disposizioni di 2, 3, …, q forme che si ripetono più volte nel corpus.
[3] Forme che, se portate fuori dal contesto, non hanno un significato autonomo, come ad esmepio articoli e preposizioni

1.3.4 Le forme testuali

Infine, in un'ottica di statistica testuale è possibile considerare quali unità le forme testuali, intese come unità minime significative del discorso, siano esse semplici o composte. In questa ottica l'interesse si sposta su unità che abbiano un significato proprio nel discorso. Si tratta pertanto sia di forme grafiche sia di espressioni qualora queste siano unità minimali in grado di catturare il giusto significato (Bolasco S., 2005), ossia dove il significato del segmento è diverso dalla somma dei significati delle forme componenti.

1.4 La disambiguazione

Nel paragrafo precedente sono state illustrate le principali unità di analisi adottate nell'ambito dell'analisi statistica dei dati testuali. Ad eccezione delle forme grafiche, se si escludono le omografie, notevoli problemi di ambiguità si possono incontrare nella individuazione di queste unità. In particolare le ambiguità possono essere di natura lessicale, come ad esempio le forme flesse di lemmi differenti, o semantica , perché riferite a più concetti differenti.

Negli ultimi anni lo sviluppo tecnologico, permettendo da un lato la raccolta di grandi masse di testi in formato elettronico e dall'altro lo sviluppo di algoritmi statistici volti all'identificazione di

regolarità e peculiarità nella distribuzione dei dati, ha consentito notevoli sviluppi nel campo della disambiguazione automatica anche attraverso lo sviluppo di appositi software.

1.5 Il sistema dei pesi

Per la corretta interpretazione del fenomeno linguistico è fondamentale considerare l'esatta importanza di ogni unità coinvolta. Per analizzare automaticamente un insieme di documenti si ricorre alla loro trasformazione in vettori. Tenendo conto quindi dell'importanza dei termini *j* un generico documento sotto forma vettoriale può così essere rappresentato:

$$d_i = \{w_{i1}, w_{i2}, \dots w_{ij} \dots w_{iq}\}$$

Dove w_{ij} è l'importanza della forma j-esima nel documento i-esimo.

1.5.1 Il peso booleano

Lo schema di ponderazione di presenza/assenza o Booleano è sicuramente il più semplice da adottare. Basato sulla presenza o l'assenza di una determinata forma testuale all'interno di un

documento: questo sistema assegna valore $w_{ij}=1$ qualora la forma *j*-esima è presente nel documento *i*-esimo, altrimenti la forma avrà importanza $w_{ij}=0$.

Il vantaggio della semplicità di applicazione di tale sistema è pagato con la distorsione dell'importanza di ogni forma, in quanto essa è misurata allo stesso modo tanto nei documenti fortemente caratterizzati da una determinata forma quanto nei documenti in cui la stessa forma è priva di contenuto informativo caratterizzante.

1.5.2 Le frequenze

Questo sistema è il più diffuso in alternativa al sistema basato sullo schema Booleano. Esso si basa sullo schema bag-of-word (BOW) assegnando al peso w_{ij} la frequenza statistica (numero di occorrenze) della forma *j*-esima nel documento *i*-esimo.

1.5.3 Il TF-IDF

Con lo sviluppo di Internet si sono ampliate le fonti di raccolta dei dati e dei testi; in alcuni casi, dipendenti ovviamente dall'obiettivo dell'indagine, è preferibile utilizzare sistemi di ponderazione complessi, in particolar modo per le tecniche di trattamento del linguaggio naturale connesse all'Information Retrieval. Questi sistemi

tengono conto dell'importanza di ogni forma sia rispetto ad ogni specifico documento che alla totalità dei documenti contenuti nel corpus.

L'indice che è alla base dei più robusti e computazionalmente semplici sistemi di ponderazione "complessi" delle forme testuali è il Term Frequency / Inverse Document Frequency (Salton, et al., 1988).

Il TF-IDF fonda le sue radici su due idee:

- Le forme principali (o parole piene) che occorrono con una frequenza maggiore all'interno di un documento, sono generalmente indicative del suo contenuto. Nasce quindi il cosiddetto Term Frequency normalizzando la frequenza del termine *j*-esimo con la frequenza della forma con occorrenza maggiore, all'interno del documento i-esimo:

$$tf_{ij} = 0{,}5 + 0{,}5\frac{f_{ij}}{maxf_i}$$

  In corrispondenza di livelli più alti dell'indice, considerando un campo di variazione compreso tra 0,5 e 1, si individuano le forme con un contributo informativo maggiore;
- Le forme principali possono essere presenti all'interno di più documenti con importanza differente. Bisogna quindi valutare il livello di discriminazione delle forme all'interno del corpus; si può quindi considerare l'Inverse Document Frequency:

$$idf_j = log\left(\frac{n}{df_j}\right)$$

dove n è il numero totale di documenti e df_j il numero di documenti in cui appare la forma *j*-esima. L'uso del logaritmo è giustificato dalla necessità di "compensare" l'effetto del TF.

Combinando in vario modo i due indici si ottiene il TF-IDF (Term Frequency / Inverse Document Frequency); non esiste infatti uno schema ideale, ma è l'esperienza del ricercatore a consigliare la combinazione più appropriata.

Oren (Oren, 2002) ha proposto il ricorso alla programmazione genetica per ottenere automaticamente nuovi schemi basati sul TF-IDF.

Il best fully weighted system è una delle formulazioni del TF-IDF più utilizzate in quanto consente di comparare i risultati ottenuti da corpora differenti. Questo perché la quantità al denominatore normalizza l'indice considerando la lunghezza del documento in esame. Essa è data da:

$$tf - idf_{ij} = \frac{f_{ij} log\left(\frac{n}{df_j}\right)}{\sqrt{\sum_j \left(f_{ij} log\left(\frac{n}{df_j}\right)\right)^2}}$$

1.6 L’analisi

Le analisi quantitative svolte sul linguaggio naturale sono di due tipi: linguistiche e statistiche: le prime individuano una serie di misure atte a misurare la ricchezza lessicale del vocabolario utilizzato e le seconde permettono l’interpretazione e la visualizzazione del fenomeno.

I metodi statistici più comunemente utilizzati nell’analisi statistica dei dati testuali rientrano nell’ambito delle tecniche di Analisi multidimensionale dei dati.

Tali metodi hanno origine dall’analisi multivariata, i cui primi studi risalgono all’inizio del secolo scorso con i lavori di Spearman (Spearman, 1904) e Pearson (Pearson, 1901) e sono poi formalizzati dagli studi di Hotelling (Hotelling, 1933); ma è solo a partite dagli anni ’60 che l’analisi multidimensionale dei dati si sviluppa e diffonde per opera della scuola francese di Benzécri (Benzécri, 1973).

Questi metodi vengono preferiti ad altri, nell’approccio all’analisi di dati testuali (Benzécri, 1982), perché il fenomeno oggetto di studio (il linguaggio naturale) è di tipo osservazionale e non si adatterebbe a riferimenti probabilistici sul tipo di distribuzione delle variabili considerate.

Lo scopo principale dell'analisi multidimensionale è di evidenziare la struttura latente sottostante al testo in esame tramite una riduzione di dimensionalità dello spazio di rappresentazione delle variabili linguistiche o dei documenti.

1.6.1 L'analisi delle corrispondenze

L'analisi delle corrispondenze (AC) costituisce uno dei più noti ed efficaci metodi per il trattamento multidimensionale di dati qualitativi.

L'AC, che probabilmente ha avuto origini dai lavori di Fisher (Fisher, 1940) sulle tabelle di contingenza, era rivolta allo studio delle relazioni esistenti tra gli elementi di due insiemi rappresentati dalle modalità di due caratteri riportate sulle righe e sulle colonne di una tabella, appunto, di contingenza; esso si è poi esteso al caso di più variabili qualitative (analisi delle corrispondenze multiple) e quindi ad un approccio non simmetrico, cioè all'esistenza di una relazione di antecedenza e conseguenza delle variabili (analisi non simmetrica delle corrispondenze).

Obiettivo dell'AC è descrivere, sia da un punto di vista geometrico che da un punto di vista algebrico, le relazioni tra distribuzioni, espresse in forma matriciale, delle modalità di due o più caratteri in un insieme di unità statistiche.

Sia **N** (*p*,*q*) una tabella di contingenza in cui si incrociano due variabili qualitative *x*, con *p* modalità, ed *y*, con *q* modalità, il cui valore generico n_{ij} è la frequenza con cui si presentano contemporaneamente la *i*-esima modalità di *x* e la *j*-esima modalità di *y*.

L'obiettivo è quello di calcolare una serie di fattori, ciascuno dei quali rappresenta un aspetto latente del tipo di associazione presente nei dati, procedendo a delle procedure di decomposizione.

Innanzitutto si costruiscono la matrice delle frequenze relative **F** (con $f_{ij}=n_{ij}/n_{..}$) e i vettori delle frequenze marginali di riga **r** e colonna **c** (con $f_{i.}=n_{i.}/n_{..}$ e $f_{.j}=n_{.j}/n_{..}$); si ricavano dapprima le matrici diagonali delle distribuzioni marginali di riga $\mathbf{D}_v \equiv diag(\mathbf{r})$ e di colonna $\mathbf{D}_q \equiv diag(\mathbf{c})$ e successivamente le matrici delle distribuzioni condizionate:

$$\tilde{\mathbf{R}} \equiv \mathbf{D}_p^{-1}\mathbf{F} \equiv \begin{bmatrix} \tilde{\mathbf{r}}'_1 \\ \vdots \\ \tilde{\mathbf{r}}'_v \end{bmatrix}$$

$$\tilde{\mathbf{C}} \equiv \mathbf{F}\mathbf{D}_q^{-1} \equiv \begin{bmatrix} \tilde{\mathbf{c}}'_1 \\ \vdots \\ \tilde{\mathbf{c}}'_q \end{bmatrix}$$

Dove $\tilde{\mathbf{R}}$ rappresenta la matrice dei profili riga e $\tilde{\mathbf{C}}$ la matrice dei profili colonna, ossia delle distribuzioni condizionate.

Si rappresentano così la nube dei profili riga $\tilde{\mathbf{r}}_1, \dots, \tilde{\mathbf{r}}_v$ in uno spazio $\mathbb{R}^{q-1}$ con pesi $\mathbf{D}_p$ e metrica $\mathbf{D}_q^{-1}$ e la nube dei profili colonna $\tilde{\mathbf{c}}_1, \dots, \tilde{\mathbf{c}}_n$ in uno spazio $\mathbb{R}^{v-1}$ con pesi $\mathbf{D}_p$ e metrica $\mathbf{D}_v^{-1}$. La Figura 1.1 illustra le matrici dell'AC.

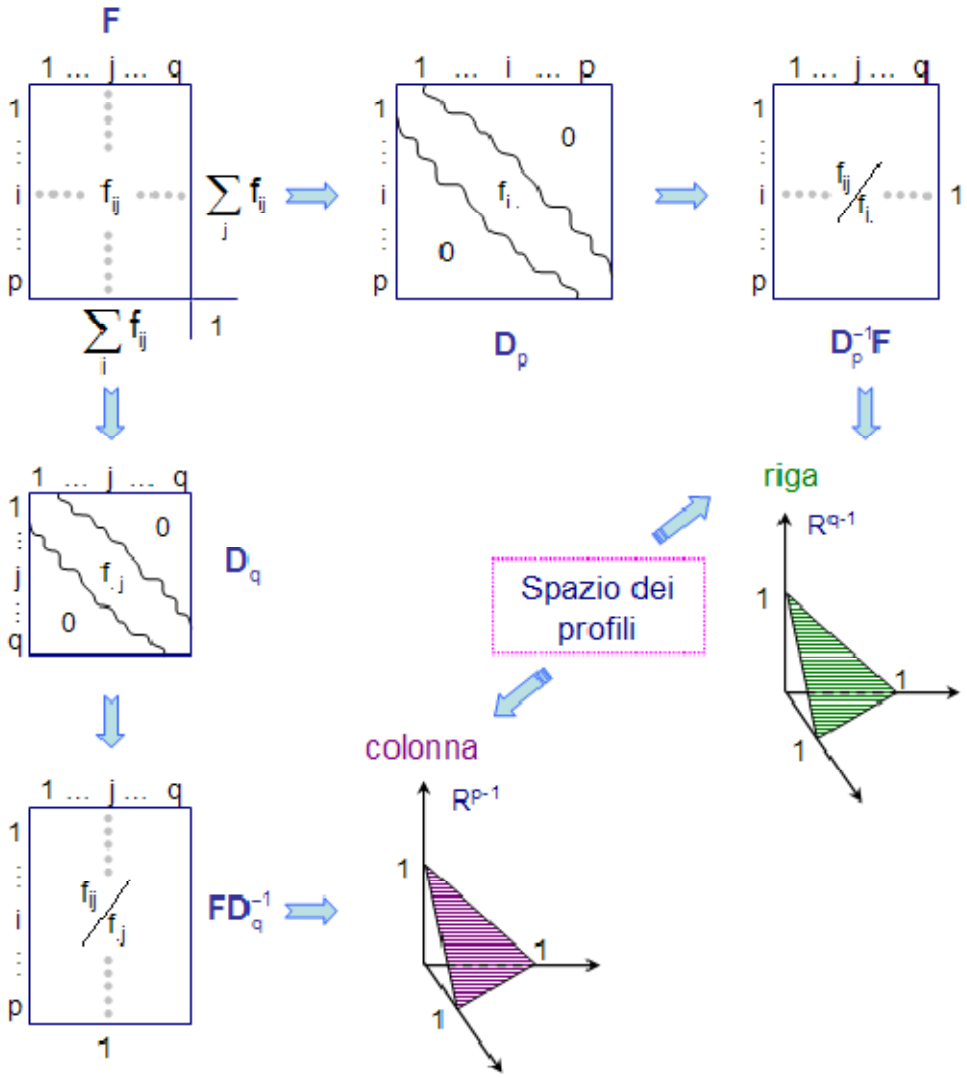

Figura 1.1 - Le matrici dell'analisi delle corrispondenze

La decomposizione in valori singolari generalizzata (Greenacre, 1984) della matrice **F** permette di individuare i sottospazi *n*-dimensionali che meglio approssimano in termini di minimi quadrati le proiezioni dei profili riga e colonna.

Gli *n* vettori singolari destri e sinistri definiscono gli assi principali degli spazi in cui sono rappresentate le nubi dei profili riga e colonna:

$$\boldsymbol{F} = \boldsymbol{U}\boldsymbol{D}_u\boldsymbol{V}'$$

$$\boldsymbol{con}\ \boldsymbol{U}'\boldsymbol{D}_p^{-1}\boldsymbol{U} = \boldsymbol{V}'\boldsymbol{D}_q^{-1}\boldsymbol{V} = \mathbf{I}$$

dove $\mathbf{D}_u$ è la matrice dei valori singolari, e le colonne delle matrici $\boldsymbol{D}_v^{-1/2}\boldsymbol{U}$ e $\boldsymbol{D}_q^{-1/2}\boldsymbol{V}$ rappresentano gli assi principali degli spazi dei profili colonna e dei profili riga.

Le coordinate dell' *i*-esimo punto riga e del *j*-esimo punto colonna sull'α-esimo asse sono ottenute dalle relazioni:

$$\psi_{\alpha i} = \sqrt{\lambda_\alpha} f_{i.}^{-1/2} u_{\alpha i}$$

$$\varphi_{\alpha j} = \sqrt{\lambda_\alpha} f_{.j}^{-1/2} v_{\alpha i}$$

La bontà dell'approssimazione è valutata in termini di variabilità del fenomeno spiegata dai sottospazi individuati con l'analisi fattoriale ricorrendo al rapporto tra il quadrato dell'α-esimo valore singolare e la somma dei quadrati di tutti i valori singolari ottenuti con la SVD.

Il contributo assoluto del punto *i*, ovvero del punto *j*, è la quota di variabilità dell'asse spiegata dal punto *i*, ovvero *j*. Esso è dato da:

$$ca_\alpha(i) = f_{i.}\psi_{\alpha i}^2$$

$$ca_\alpha(j) = f_{.j}\varphi_{\alpha j}^2$$

I risultati di un'AC possono essere rappresentati su grafici piani ottenuti utilizzando coppie di assi principali; a differenza di altri metodi fattoriali come l'analisi in componenti principali, nell'AC è possibile la rappresentazione congiunta delle modalità di un carattere con le modalità dell'altro perché recupera una simmetria complessiva come conseguenza di una ideale sovrapposizione dei risultati di due analisi asimmetriche.

L'AC infatti gode della proprietà baricentrica per cui è possibile calcolare la posizione dei profili riga (ovvero dei profili colonna) come media ponderata dei vertici definiti dai profili colonna (ovvero dei profili riga), dove i pesi sono pari proprio agli elementi del

profilo riga (ovvero profilo colonna) stesso. La rappresentazione così ottenuta è detta β-baricentrica per sottolineare la diversità dalla situazione ideale, ma impossibile, in cui ogni riga è baricentro delle colonne e viceversa.

1.6.2 L’analisi delle corrispondenze lessicali

Nel caso specifico dell’analisi delle corrispondenze lessicali (ACL), si considera una tabella lessicale **T** di dimensioni (*V,n*) il cui elemento generico t_{ij} rappresenta il numero di occorrenze della *i*-esima forma nel *j*-esimo documento. E’ chiaro che la scelta delle unità di analisi deve essere già stata compiuta.

La tabella **T** è però una matrice sparsa: presenta molte celle vuote ed è di grandi dimensioni. Spesso si considera la cosiddetta tabella lessicale aggregata ottenuta da una matrice **Q** di dimensioni (*n*,*q*) in codifica disgiuntiva completa, le cui *q* colonne sono modalità di una variabile qualitativa relativa ai documenti. Si ottiene così una tabella $\mathbf{T_q}$ = **TQ** che ha in riga le *v* forme ed in colonna le *q* classi di documenti.

Per la lettura degli assi generati da una ACL è comunque opportuno sottolineare che:

- la dispersione dei punti intorno all'origine degli assi principali mostra la forza dell'associazione nella tabella lessicale;
- se due forme sono vicine sono utilizzate in maniera simile;
- se due modalità della variabile di partizione sono vicine, allora vuol dire che le corrispondenti categorie di documenti utilizzano un vocabolario simile;
- non si può leggere la prossimità di una forma ad una categoria (o viceversa), ma valutare la sua posizione rispetto all'intera nube delle categorie (o delle forme), secondo la logica della rappresentazione β-baricentrica;
- l'importanza di un punto rispetto alla spiegazione di un asse principale è letta in termini di contributi assoluti.

1.6.3 Il Latent Semantic Indexing

Il latent semnatic indexing (LSI) è una tecnica proposta da Dumais et al. (ins citazione) che cerca di "estrarre i concetti" presenti nei documenti, e quindi di superare il limite proprio di ogni tecnica di information retrieval che si basa sulla presenza o meno di termini per stabilire la rilevanza di un documento rispetto a una query.

Partendo da una rappresentazione vettoriale dei documenti, in cui ogni coordinata corrisponde a un termine, LSI cerca di "proiettare" i vettori dei documenti in un "sotto-spazio semantico latente", a dimensionalità ridotta, in cui le coordinate sono i "concetti".

Un concetto è visto come un insieme di termini che occorrono (frequentemente) insieme negli stessi documenti.

Alla base del LSI vi è una SVD della matrice **T** di dimensioni (*V,n*) il cui elemento generico t_{ij} rappresenta il numero di occorrenze della *i*-esima forma nel *j*-esimo documento che permette una riduzione della dimensionalità.

L'idea è quindi che le nuove dimensioni sono la vera rappresentazione oscurata dai processi compositivi dei singoli autori che esprimono una particolare dimensione mediante un differente insieme di parole all'interno di un altro documento. Il LSI riporta alla luce l'originale struttura semantica dello spazio e le sue dimensioni originali.(Dulli, et al., 2004).

1.7 Una possibile procedura (taltac)

In questo paragrafo si vuole illustrare una strategia di base per l'analisi di un corpus, con particolare riferimento alla fase di codifica utile per qualsiasi tipo di indagine. Lo sviluppo di software dedicati

permette, infatti, di individuare una filiera di operazioni da svolgere. Ad ogni passaggio sarà compito del ricercatore valutare ed individuare le giuste opzioni che meglio collimano con gli obiettivi dell'analisi. In questa trattazione si farà riferimento al software TALTAC[4] sviluppato da Bolasco, Morrone e Baiocchi (www.taltac.it).

1.7.1 Acquisizione e normalizzazione

La prima operazione da compiere è detta *parsing*; questa procedura consente l'individuazione dei caratteri dell'alfabeto comprese tra i caratteri separatori. Come descritto nel paragrafo 1.3 questa fase permette l'individuazione di tutte le forme grafiche.

L'operazione immediatamente successiva è la *normalizzazione* del testo, che può essere leggera o basata su liste(Giuliano, 2004). La normalizzazione leggera consente alcune modifiche al corpus necessarie per la corretta individuazione delle forme grafiche. Si tratta sostanzialmente di apportare modifiche dovute alle diverse digitazioni dei corpus: la riduzione degli spazi multipli e dei doppi apici in virgolette, l'aggiunta dello spazio dopo l'apostrofo e la trasformazione degli apostrofi in accento.

[4] Trattamento Automatico Lessico-Testuale per l'Analisi del contenuto di un Corpus.

Queste fasi sono strettamente attinenti all'individuazione delle fome grafiche presenti nel corpus e quindi necessarie qualunque sia l'obiettivo dell'analisi e la scelta del tipo di unità da analizzare.

La normalizzazione basata su liste permette di categorizzare le forme delle quali si vuole conservare la specificità. Essa si basa sul riconoscimento di locuzioni grammaticali, polirematiche, nomi propri, celebrità, ecc. presenti nelle liste dei software. La scelta di attuare questa fase, invece, per quanto fortemente consigliata, dipende dalle decisioni del ricercatore.

1.7.2 La lessicalizzazione

L'operazione susseguente riguarda la scelta dei segmenti ripetuti da lessicalizzare. È chiaro che questa fase interesserà esclusivamente i ricercatori interessati alle forme testuali quali unità di analisi.

I segmenti sono disposizioni di 2, 3, …, g forme che si ripetono più volte nel corpus. Per individuarli è necessario, nei casi di corpora di medie dimensioni o più, fissare due soglie di frequenza: la soglia minima delle forme appartenenti al segmento e la soglia minima di frequenza del segmento stesso.

Morrone (Morrone, 1993) ha proposto costruito un indice di rilevanza del segmento:

$$IS = \left(\sum_{i=1}^{L} \frac{f_{segm}}{f_{fg_i}} \right) P$$

Dove L è la lunghezza del segmento[5], f_{segm} è la frequenza della forma nel segmento, f_{fg_i} è la frequenza della forma nel corpus e P è il numero di di parole piene presenti nel segmento. L'indice si annulla quando il segmento è composto solo da parole vuote, e ha un massimo pari a L^2. È possibile pertanto costruire un l'indice IS relativo:

$$IS_{rel} = \frac{IS}{L^2}$$

In funzione di questo indice è possibile individuare i segmenti ripetuti da lessicalizzare.

1.7.3 La lemmatizzazione

L'ultima operazione da eseguire necessaria per l'individuazione delle unità è il tagging grammaticale. Attraverso questa operazione è possibile associare, e *"marcare"*, ogni forma alla sua categoria grammaticale di appartenenza, ossia la "parte del

[5] Numero di forme componenti il segmento

discorso" (in inglese POS, *Part Of Speech*). Una volta riconosciuta la POS di appartenenza è possibile ricondurre ogni forma grafica al lemma di appartenenza.

Il lemma è la "forma canonica" con cui una data voce è presente in un dizionario: si considera quindi l'infinito per i verbi, il singolare per i sostantivi, il singolare maschile per gli aggettivi. Il principio alla base di tale strategia è che le varianti morfologiche più comuni di una forma hanno significato simile e sono usati in contesti simili.

Questa fase presenta comunque delle difficoltà relative alla necessità di disambiguazione.

L'operazione comunque consente il raggiungimento di tre obiettivi diversi:

- l'individuazione delle forme lemmatizzate per chi decide di utilizzarle quali unità di analisi;
- l'assegnazione delle POS consente l'individuazione delle forme vuote, qualora si vogliano eliminare dalle analisi successive;
- il calcolo degli indicatori sulle frequenze d'uso delle POS nei corpora.

CAPITOLO 2

2 Un’analisi classica per obiettivi di rassegna

Questo capitolo propone di utilizzare le capacità di sintesi e di rappresentazione grafica delle tecniche di analisi multidimensionale dei dati, applicate a basi di dati testuali, nell’intento di far emergere concordanze e ricorrenze d’uso di vocaboli in rapporto a determinati ambiti tematici (temporali, spaziali, settoriali).

L’obiettivo è quello di mostrare come le tecniche proprie del TM possano permettere la realizzazione di una rassegna di pubblicazioni scientifiche.

Anche in questo caso è l'obiettivo dell'analisi ad influenzare la fase di codifica del dato testuale.

2.1 Introduzione

La disponibilità sempre crescente di risorse digitali on-line nel campo della ricerca di materiale bibliografico se da un lato offre nuove potenzialità in termini di quantità, qualità e immediatezza del reperimento, dall'altro richiede alti gradi di standardizzazione e organizzazione di grandi basi di dati. La disponibilità, pressoché in tempo reale, di pubblicazioni in formato digitale non può prescindere dai requisiti fondamentali di pertinenza, omogeneità e, possibilmente, di esaustività rispetto ad un determinato ambito di ricerca. La necessità di creare strutture di controllo per tali caratteristiche è strettamente legata alla possibilità di organizzare il materiale disponibile, attraverso opportune operazioni di codifica di dati e metadati, in database operativi.

I passaggi metodologici sono illustrati con riferimento ad una rassegna svolta sulla Conjoint analysis, una tecnica di analisi multivariata molto applicata nell'analisi di mercato.[6]

Le fasi della ricerca sono le seguenti:

[6] Per una descrizione più approfondita del metodo si veda il par. 4.1.1

- creazione del database delle pubblicazioni riferite all'argomento di interesse;
- codifica dei dati;
- costruzione di mappe fattoriali attraverso la tecnica dell'Analisi delle Corrispondenze su una tabella di frequenza lessicale.

2.2 La creazione del database

La ricerca è stata svolta con riferimento a tutte le pubblicazioni redatte su rivista scientifica in lingua inglese con contenuti sia metodologici che applicativi dal 1960 al 2004.

Le risorse digitali consultate sono state le emeroteche virtuali delle Università *Federico II* di Napoli e dell'Università di Salerno[7]. La prima utilizza una piattaforma integrata *MetaLib/SFX*, il *link server* di *Ex Libris* basato sull'*OpenURL*, mentre la seconda oltre a far parte del consorzio *Caspur* gestito dall'Università *La Sapienza* di Roma offre il servizio di aggregazione *SwetsWise* con il quale è possibile consultare gli indici dei fascicoli, partendo dal titolo dell'articolo, oppure effettuare una ricerca per autore, titolo e parole chiave. Anche in

[7] - Le emeroteche virtuali dei rispettivi atenei sono reperibili ai seguenti URL:
http://www.biblio.unina.it/risorsedigit.html, (Napoli);
http://www.csab.unisa.it/periodici.asp, (Salerno).
http://periodici.caspur.it, (Roma)

questo caso l'accesso alle risorse avviene tramite riconoscimento dell'indirizzo del protocollo Internet (*IP address*) e solo se il titolo è in abbonamento si può accedere al *full-text*.

L'indagine è iniziata con la ricerca di tutte le pubblicazioni contenenti (in: titolo, parola chiave, abstract, corpo) la forma "conjoint analysis". In tal modo si è provveduto ad un primo screening delle pubblicazioni per soddisfare il requisito di pertinenza col tema cercato.

Al termine di questa fase di selezione il numero di pubblicazioni utili è risultato essere di circa 1300 titolazioni. Si è resa necessaria la scelta di limitare il tipo di pubblicazioni alle sole riviste scientifiche eliminando le pubblicazioni provenienti da atti di convegno (*proceedings*), tesi di laurea, tesi di dottorato (*Ph.D.*) e dissertazioni in genere (master, rapporti di ricerca, ecc.). Tanto allo scopo di ottenere una maggiore omogeneità dei lavori e rendere più facilmente raggiungibile il traguardo che in prospettiva futura mira all'esaustività.

La raccolta effettuata è stata, dunque, organizzata in un data base che ha visto la registrazione degli Autori, del Titolo del lavoro, della Rivista Scientifica e dell'Anno di pubblicazione.

L'arco temporale riguarda, dunque, il periodo intercorso dalle prime pubblicazioni ufficialmente riconosciute sulla Conjoint Analysis (databili agli inizi degli anni '70), fino ai giorni nostri (primo semestre del 2004). Una curiosità, a tal proposito, è costituita da alcune pubblicazioni apparse nel 1969: una prima, a cura del *L.L. Thurstone*

Psychometric Laboratory, dal titolo: *"Polynomial Conjoint Analysis of Similarities: a Model for Constructing Polynomial Conjoint Measurement Algorithms"* ad opera di Forrest Young. In tale lavoro l'Autore fornisce un modello computazionale prima ancora che la Conjoint Analysis sia definita come metodologia. L'ambito metodologico di riferimento sembra piuttosto essere, in questo caso, il *Multidimensional Scaling*[8].

Invece risulta tecnicamente impossibile da rilevare sia il manoscritto, sia il contributo di J. D. Carroll di quell'anno: "*Categorial Conjoint Measurement*", presentato al *Meeting of Mathematical Psychology*, Ann Arbor, 1969.

Tali riferimenti, vista la loro datazione, dovrebbero rappresentare una sorta di opera prima per la Conjoint Analysis, ma in realtà essi raramente vengono citati in tal senso.

Aldilà di quest'eventi, ancora isolati, l'anno ufficiale del battesimo scientifico della Conjoint Analysis appare essere il 1971 con la pubblicazione di "*Conjoint measurement for quantifying judgement data"* di P.E. Green. & V.R. Rao sul *Journal of Marketing Research.* A cui seguono due pubblicazioni nel 1972 e altre due nel 1973 sempre di Paul Green con altri studiosi che ancora avrebbero, in futuro, scritto

[8] - Il contributo apparirà nel 1972 nella raccolta a cura di R.N. Shepard, A.K. Romney, S.B. Nerlove: *"Multidimensional Scaling: Theory and Application in the Behavioral Sciences, Vol.1: Theory"*. New York: Seminar Press.

insieme la storia della Conjoint Analysis: Frank J. Carmone e Yoram Wind.

Occorrerà comunque attendere fino al 1974 per assistere ad una più ampia divulgazione della tecnica statistica della Conjoint Analysis anche in settori applicativi diversi dal marketing e dal dibattito teorico sulle scale di misura di interesse in particolare per psicologici e psicometrici. Si registra, infatti, in quell'anno il primo contributo in ambito economico dal punto di vista applicativo e nell'ambito della Ricerca Operativa, dal punto di vista teorico e computazionale.

Preliminarmente le pubblicazioni sono state attribuite a settori disciplinari individuati sulla base della citazione che per ogni rivista è possibile rilevare sulla pagina web dell'editore. I 30 settori sono riportati nella tabella.

1 . agriculture	11 .food	21 . social
2 . consumer	12 .forest	22 . software
3 . decision science	13 .geography	23 . statistics
4 . ecology	14 .health	24 . teaching
5 . economics	15 .Information	25 . tourism
6 . educational	16 . library	26 . transports
7 . engineers	17 . management	27 . urban
8 . environments	18 . marketing	28 . media
9 . ergonomics	19 . operational research	29 . military
10 .finance	20 . psychology	30 . nutrition

TABELLA 2.1 - I SETTORI DISCIPLINARI DELLA CONJOINT ANALYSIS

Naturalmente non mancano casi di sovrapposizione e di gerarchie dei temi trattati; in tutti i casi di ambiguità si è scelto di attribuire i lavori al settore più ristretto e logicamente più interno possibile alla gerarchia. Ad esempio: {Consumer $\subset$ Economics}, {Agriculture $\subset$ Food}, e così via.

In alcuni casi si è preferito non effettuare alcuna attribuzione e lasciare il settore della rivista inalterato. È questo il caso di tre pubblicazioni che sono state rinvenute nella nostra ricerca: i settori di appartenenza (*media, military, nutrition*) ci sembrano interessanti per la singolarità e per la circostanza che il periodo di pubblicazione è molto recente (2001, 2002).

La Tabella 2.2 consente di avere un'immediata percezione della distribuzione dei lavori classificati per anno di pubblicazioni e per settore scientifico.

Titoli/Anni	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	TOTALE
agriculture	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	1	0	1	0	0	0	3	1	2	2	1	1	5	6	4	10	2	16	9	9	8	84
consumer	0	0	0	0	0	0	1	0	1	3	1	0	2	2	0	1	0	0	1	1	1	1	0	1	2	0	0	0	1	3	3	0	3	1	5	2	36
decision science	0	0	0	0	0	1	0	0	0	0	0	4	4	0	0	1	0	1	0	2	1	0	1	0	1	0	0	0	0	2	0	0	2	2	1	3	26
ecology	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	3	1	0	6
economics	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	1	0	3	4	3	5	1	4	1	3	1	4	1	3	41
educational	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	0	1	2	1	0	3	1	1	13
engineers	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	1	1	1	0	0	0	0	3	0	3	2	0	14
environments	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	3	0	2	3	4	2	2	0	20
ergonomics	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	4	0	7
finance	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	0	1	0	0	1	0	1	1	0	0	8
food	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	3	2	4	2	7	3	4	4	7	6	5	3	52
forest	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	3	0	1	3	1	0	2	13
geography	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	2	2	0	0	0	0	0	1	2	0	0	0	0	1	0	1	0	11
health	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	3	3	0	2	1	0	4	3	2	5	6	6	6	8	10	13	22	18	16	23	20	176
Information	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	0	0	1	1	2	1	3	1	5	5	2	3	3	1	32
library	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	5
management	0	0	0	0	0	1	1	0	0	3	0	5	4	4	3	3	2	6	3	4	4	7	6	14	6	12	5	10	6	15	10	8	14	7	4	167
marketing	0	0	1	2	2	2	3	0	1	2	2	4	6	9	6	6	3	11	11	19	6	5	14	11	6	11	15	12	11	13	6	10	9	17	17	6	259
operational research	0	0	0	0	0	1	0	1	0	0	0	0	0	2	0	0	0	0	1	0	2	1	0	1	0	0	0	0	1	1	0	0	3	3	0	1	18
psychology	2	1	0	0	1	0	0	0	0	1	0	0	0	0	2	0	0	1	1	0	2	2	0	0	2	1	0	1	0	0	1	0	3	0	4	3	28
socio	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	2	1	0	0	0	2	2	1	3	1	0	5	0	2	0	5	2	3	3	1	2	37
software	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	4
statistics	0	0	0	0	0	0	0	0	0	0	0	3	0	1	0	0	1	0	1	1	0	1	1	0	0	0	0	1	0	1	1	2	1	3	3	1	22
teaching	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	2
tourism	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0	0	1	0	3	2	0	1	1	3	1	18
transports	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	1	2	0	0	7
urban	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	2	1	0	7
media	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
military	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
nutrition	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
TOTALE	2	1	1	2	3	6	5	1	2	6	6	14	21	20	19	18	14	17	27	35	25	29	34	29	45	35	51	47	54	56	71	72	90	103	94	61	1 116

TABELLA 2.2 - LA DISTRIBUZIONE DEI LAVORI SULLA CONJOINT ANALYSIS

Questa tabella mette in evidenza alcuni aspetti degni di rilievo:

- i settori con più attribuzioni sono nell'ordine: *marketing*, *health*, *management*;
- seguono come numero di attribuzioni i settori: *agriculture* e *food* che superano anche le attribuzioni dei settori *economics* e *consumer*;
- le pubblicazioni in settori come *health* e *agriculture* hanno caratterizzato soprattutto gli ultimi cinque anni nonostante facciano la loro apparizione sin dall'inizio degli anni '80.

Da tali considerazioni sembrano emergere dei tratti distintivi ben delineati che possiamo dunque assumere come linee guida nella ricerca:

- La Conjoint Analysis è nata dal dibattito teorico sulle scale di misura che veniva affrontato a metà degli anni ’60 dagli psicometrici.
- L’interesse dal punto di vista statistico è legato allo studio dei dati di prossimità o similarità e viene affrontato principalmente nel contesto delle tecniche di *Multidimensional Scaling* per la costruzione di mappe di posizionamento.
- Gli studi di Marketing rappresentano storicamente il primo ed il più importante ambito applicativo della Conjoint Analysis;
- Nuovi ambiti applicativi vanno affermandosi in tempi più recenti: la Conjoint Analysis viene scoperta dagli studiosi di mercati atipici ed affluisce negli studi di settore nel campo della salute pubblica, delle cure mediche (*health*), dei mercati alimentari (anche non tradizionali, come quelli orientali), dell’acquacultura e degli allevamenti (*food*, *agriculture*).

L’analisi del vocabolario utilizzato nei titoli dei lavori ha mostrato diversi aspetti, anche inattesi, facendo prefigurare una

varietà di accezioni teoriche ed applicative contraddistinguenti la Conjoint Analysis.

Il passo successivo dell'analisi ha consentito l'approfondimento di questi aspetti attraverso l'esplorazione del vocabolario utilizzato nei titoli dei lavori.

L'idea è che dalle forme testuali utilizzate dagli Autori nei titoli dei propri lavori potevano emergere tipizzazioni in grado di confermare le linee guida già emerse o a disegnarne di nuove.

2.3 La codifica dei dati

Una volta definiti i caratteri delimitatori ***, " ' / . ; : ( ) ! ? [ ],*** il successivo passo consiste nella definizione delle unità di analisi.

L'insieme delle forme grafiche utilizzate dagli Autori (2412) è stata sottoposta ad un'opera di lemmatizzazione che è consistita nella riduzione delle forme dei modi verbali, di numero e di genere; nel riconoscimento di radici comuni (ad esempio parole come "*health*", "*healthy*", "*healthcare*", sono ricondotte ad un'unica voce); nella disambiguazione di forme grafiche (ad esempio "new" ricorre con significato diverso in segmenti testuali del tipo "*new product*" e "*New Zealand*", ma qui il caso più ricorrente è consistito proprio nell'attenzione alla forma "*Conjoint Analysis*" per il particolare ruolo che essa svolge).

È stata inoltre scelta una soglia di frequenza per le forme pari a tre, tanto al fine di poter inserire nell'analisi tutte le forme che potevano rappresentare delle particolari tendenze nell'utilizzo della Conjoint Analysis. Sono state inoltre eliminate tutte le parole "vuote" (preposizioni, articoli, congiunzioni, ecc.).

Il vocabolario così definito è stato incrociato con i sub testi (i titoli delle pubblicazioni). Tale incrocio ha permesso di definire la tabella lessicale intera (vocaboli $\times$ lavori). Successivamente i lavori sono stati riaggregati per settore scientifico (classificazione già illustrata precedentemente) costituendo così la matrice di base per l'Analisi delle Corrispondenze (vocaboli $\times$ settori).

2.4 L'analisi delle corrispondenze

L'analisi delle corrispondenze applicata ad una tabella di dati testuali consente di individuare e di rappresentare in un sottospazio di ridotte dimensioni la dipendenza statistica tra le unità testuali ed i frammenti di testo che li contengono.

Una caratteristica tipica di tale analisi è il ruolo svolto dalle modalità rare. In genere tali modalità finiscono col condizionare fortemente i risultati dell'analisi. Si è ritenuto opportuno porre tali modalità come "illustrative" conducendo l'analisi sulle restanti

modalità a condizione che superino un requisito di soglia minima di frequenza.

Un'altra soluzione sarebbe potuta essere quella di applicare l'analisi non simmetrica delle corrispondenze (Balbi, 1995) in cui l'utilizzo di una metrica euclidea non ponderata permette l'analisi di matrici sparse.

In questo caso la presenza di un settore numericamente forte (il Marketing), tradizionalmente legato all'uso della Conjoint Analysis, gioca un ruolo di attrattore delle diverse forme testuali utilizzate: è lecito attendersi che i vocaboli *nuovi* e, quindi, meno frequenti finiscano col caratterizzare dei settori non tradizionali nell'ambito applicativo della tecnica oggetto di studio.

La tabella lessicale è costituita da un vocabolario di 556 vocaboli classificati nei 30 settori scientifici individuati.

L'Analisi delle Corrispondenze condotta su questa tabella ha consentito di derivare un numero ridotto di fattori in grado di spiegare la maggior parte dell'inerzia presente nelle osservazioni ed interpretabile come misura dall'allontanamento dalla condizione di indipendenza presente nella tabella di contingenza iniziale. Il modello d'indipendenza tra vocaboli e settori scientifici rappresenta una base di confronto ideale, utile per l'interpretazione della sintesi ottenuta.

Interpretazioni alternative sono disponibili, considerando l'Analisi delle Corrispondenze nell'approccio *dual scaling*, ad esempio, i settori scientifici rappresentano le categorie di un criterio di

classificazione in cui le unità statistiche tendono ad affluire in maniera indipendente e dove le rispettive coordinate sugli assi fattoriali rappresentano la quantificazione ideale che separa al meglio i baricentri delle categorie (settori) (Nishisato, 1980).

Le analisi sono state svolte con l’ausilio del software SPAD v.5.5. Nella Tabella 2.3 si nota la distribuzione dell’inerzia totale rispetto ai primi assi fattoriali. Le dimensioni considerevoli della tabella inducono misure piuttosto modeste per ciascun asse: i primi due assi raccolgono circa il 15% dell’informazione totale.

N.		% inerzia	% cum	
1	0.2556	8.09	8.09	**********
2	0.2210	7.00	15.09	**********
3	0.1807	5.72	20.81	**********
4	0.1611	5.10	25.91	**********
5	0.1489	4.71	30.62	**********
6	0.1429	4.52	35.15	**********
7	0.1353	4.28	39.43	**********

TABELLA 2.3 - LA DECOMPOSIZIONE DELLA TABELLA LESSICALE: SOMMA DEGLI AUTOVALORI

Nelle tabelle che seguono sono riportati, per i primi cinque fattori, le coordinate (Tabella 2.4), i contributi assoluti e relativi (coseni al quadrato) (Tabella 2.5) della proiezione dei Settori disciplinari su ciascun asse fattoriale. Sono individuati (evidenziati in tabella) alcuni settori che si contrappongono sui diversi assi e sembrano costituire

delle caratterizzazioni interessanti per la tipicità del vocabolario utilizzato.

La contrapposizione principale appare essere quella tra i settori tradizionali, il Marketing ed il Management, rispetto a quello Sanitario (*Health*) sul primo asse e rispetto a quello Agricolo-Alimentare (*Agricolture*, *Food*) sul secondo asse.

La caratterizzazione è approfondita dall'analisi grafica con la rappresentazione del primo piano fattoriale (vedi FIGURA 2.1).

Settore	**Coordinate**				
	1	**2**	**3**	**4**	**5**
agriculture	-0.04	-0.97	0.12	0.08	0.17
consumer	0.14	0.09	-0.36	0.07	0.10
decision science	0.09	0.40	-0.18	-0.23	0.14
ecology	-0.30	-1.30	2.79	-2.98	-1.67
economics	0.04	-0.04	0.07	-0.10	0.28
educational	0.01	0.17	0.25	0.25	0.07
engineers	0.28	0.30	0.30	0.21	0.28
environments	0.07	-0.20	1.32	0.87	0.59
ergonomics	0.38	0.34	0.42	0.50	0.61
finance	0.44	0.56	1.24	-1.35	1.74
food	0.05	-1.26	-0.67	0.06	0.00
forest	-0.02	-0.30	0.87	0.67	0.38
geography	0.48	0.20	0.21	0.57	-1.27
health	-1.02	0.21	-0.07	0.00	-0.03
information	0.42	0.38	0.14	0.32	0.13
library	-0.25	0.06	0.14	0.31	-0.06
management	0.34	0.05	-0.12	-0.03	0.02
marketing	0.32	0.22	-0.21	-0.22	0.03
operational research	0.70	0.36	0.16	0.07	0.06

psychology	0.02	0.48	−0.23	−0.22	−0.54
Socio	−0.01	0.24	0.33	0.51	−0.63
software	0.30	0.64	0.29	−0.20	0.10
statistics	0.40	0.27	0.11	−0.45	0.23
teaching	−0.53	−0.01	−0.26	0.49	−0.26
tourism	0.73	0.05	0.30	0.67	−1.54
transports	0.37	0.08	−0.12	0.18	−0.38
Urban	0.03	0.06	1.54	1.58	0.55
Media	−0.12	0.66	0.21	0.32	0.30
military	0.68	1.11	0.52	−0.77	1.70
nutrition	0.41	−0.16	−0.32	0.00	−0.08

TABELLA 2.4 - LE COORDINATE DELL'AC

Settore	**Contributi**					**Coseni al quadrato**				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
agriculture	0.1	34.8	0.6	0.3	1.6	0.00	0.54	0.01	0.00	0.02
consumer	0.2	0.1	2.1	0.1	0.2	0.01	0.00	0.04	0.00	0.00
decision science	0.1	1.7	0.4	0.8	0.3	0.00	0.05	0.01	0.02	0.01
ecology	0.2	5.3	30.0	38.4	13.0	0.00	0.07	0.33	0.38	0.12
economics	0.0	0.0	0.1	0.2	2.0	0.00	0.00	0.00	0.00	0.03
educational	0.0	0.2	0.5	0.5	0.0	0.00	0.00	0.01	0.01	0.00
engineers	0.4	0.5	0.7	0.4	0.7	0.01	0.01	0.01	0.01	0.01
environments	0.0	0.3	17.7	8.6	4.2	0.00	0.01	0.28	0.12	0.05
ergonomics	0.5	0.4	0.8	1.2	2.0	0.01	0.01	0.01	0.02	0.03
finance	0.6	1.2	6.8	9.1	16.5	0.01	0.02	0.09	0.10	0.17
food	0.1	38.1	13.1	0.1	0.0	0.00	0.54	0.15	0.00	0.00
forest	0.0	0.6	5.9	4.0	1.4	0.00	0.01	0.09	0.05	0.02
geography	1.0	0.2	0.3	2.3	12.3	0.03	0.01	0.01	0.05	0.23
health	72.1	3.6	0.5	0.0	0.1	0.92	0.04	0.00	0.00	0.00
information	1.9	1.8	0.3	1.8	0.3	0.04	0.04	0.00	0.03	0.00
library	0.1	0.0	0.0	0.2	0.0	0.00	0.00	0.00	0.00	0.00
management	6.1	0.2	1.1	0.1	0.0	0.15	0.00	0.02	0.00	0.00
marketing	8.5	4.8	5.1	6.5	0.1	0.20	0.10	0.08	0.10	0.00
operational	2.8	0.9	0.2	0.0	0.0	0.08	0.02	0.00	0.00	0.00

research										
psychology	0.0	2.4	0.7	0.7	4.5	0.00	0.04	0.01	0.01	0.05
Socio	0.0	0.9	2.2	5.7	9.3	0.00	0.02	0.04	0.08	0.12
software	0.1	0.7	0.2	0.1	0.0	0.00	0.01	0.00	0.00	0.00
statistics	1.1	0.6	0.1	2.2	0.6	0.03	0.01	0.00	0.04	0.01
teaching	0.1	0.0	0.0	0.1	0.0	0.00	0.00	0.00	0.00	0.00
tourism	3.5	0.0	0.8	4.8	26.7	0.07	0.00	0.01	0.06	0.33
transports	0.3	0.0	0.0	0.1	0.5	0.01	0.00	0.00	0.00	0.01
Urban	0.0	0.0	9.6	11.5	1.5	0.00	0.00	0.15	0.16	0.02
Media	0.0	0.2	0.0	0.1	0.1	0.00	0.01	0.00	0.00	0.00
military	0.2	0.5	0.1	0.4	1.9	0.00	0.01	0.00	0.01	0.03
nutrition	0.1	0.0	0.1	0.0	0.0	0.01	0.00	0.00	0.00	0.00

TABELLA 2.5 - I CONTRIBUTI ASSOLUTI E RELATIVI DELL'AC

Minori tipizzazioni si evidenziano sugli assi successivi. Il terzo e il quarto asse propongono la caratterizzazione degli argomenti Ecologici-Urbano-Ambientalisti col linguaggio della Statistica. Il quinto asse contrappone gli aspetti Socio-Psicologici e Geografico-Turistici al mondo della Finanza. Una maggior articolazione di questi raggruppamenti sarà sviluppata grazie ai risultati della Classificazione Automatica elaborata sui risultati dell'analisi fattoriale.

La distribuzione dei Settori sul primo piano fattoriale è evidenziata in Figura 2.1. Si nota infatti la caratterizzazione del settore *Health* sul primo piano fattoriale e dei settori *Food*, *Agriculture*, *Ecology*, sul secondo asse. I settori *tradizionali* si addensano in prossimità dell'origine del piano e rappresentano pertanto il linguaggio comune della Conjoint Analysis.

Si notano le prossimità di settori come *Marketing, Management, Statistics, Operational Research, Consumer, Economics, ecc.* che abbiamo visto rappresentare il riferimento teorico e applicativo tradizionale della Conjoint Analysis.

L’articolazione di tali linguaggi può essere analizzato nella Figura 2.2 dove vengono rappresentati i vocaboli specifici dei diversi settori. Anche in questo caso si notano le due appendici fortemente caratterizzate, mentre il corpo centrale appare più confuso. Come accennato il settore di riferimento resta quello del Marketing intorno al quale sembrano orbitare gli altri settori, sia tradizionali che innovativi.

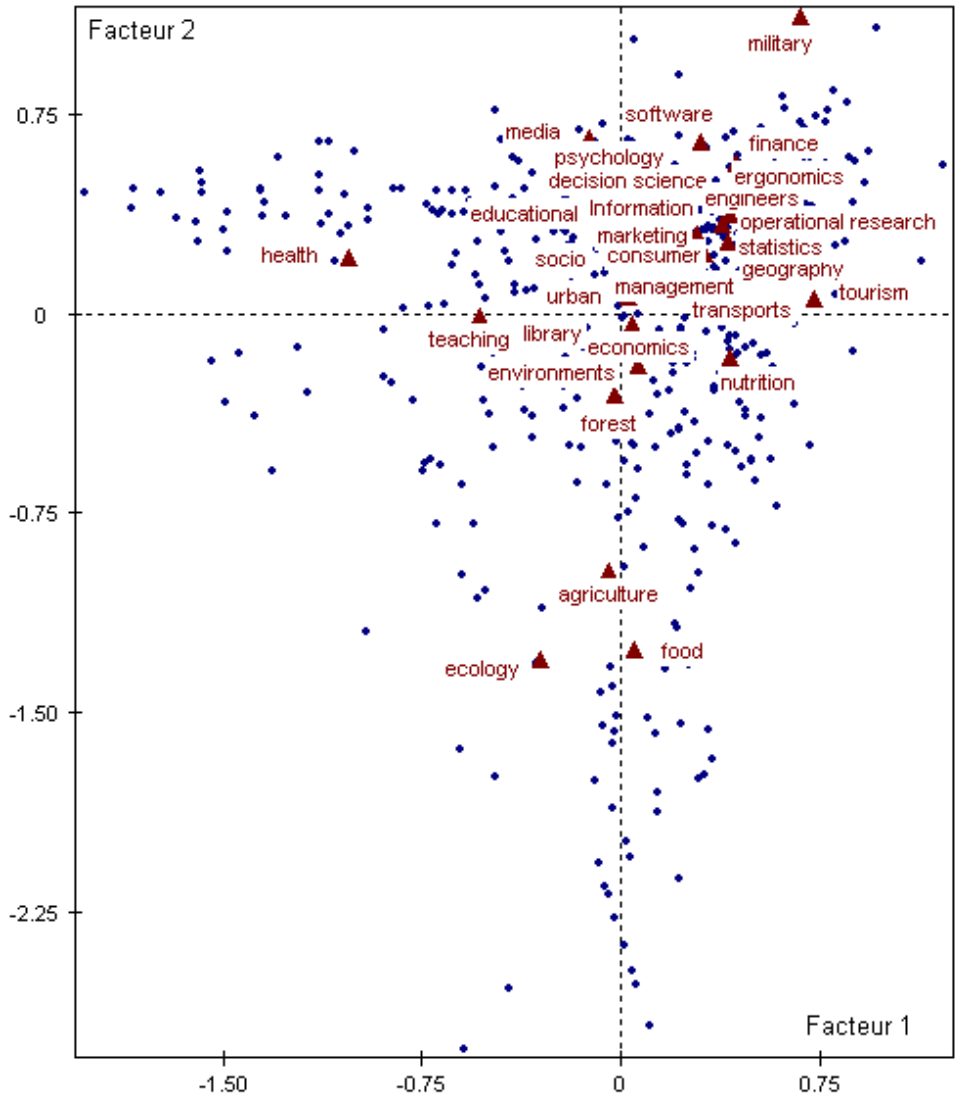

Figura 2.1 - La distribuzione dei settori nel primo asse fattoriale

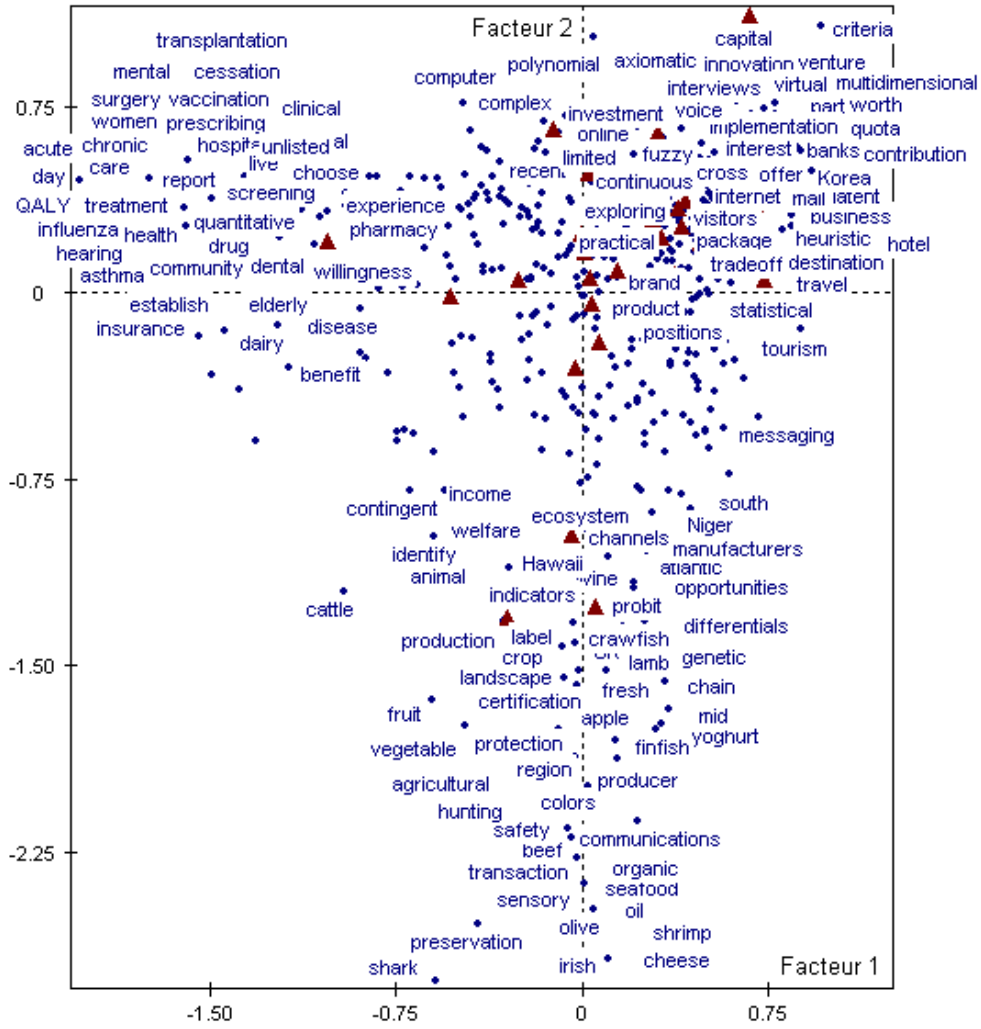

Figura 2.2 - La distribuzione del linguaggio sul primo piano fattoriale

La figura Figura 2.3 propone una lettura particolare del piano fattoriale, mettendo a fuoco proprio il ruolo centrale del Marketing ed evidenziando i settori che gravitano in misura minore o maggiore in

forza della relativa similitudine del vocabolario utilizzato. Occorre sottolineare che anche in questo caso la lettura va fatta guardando con attenzione i risultati riportati in Tabella 2.1 per quanto riguarda la qualità della rappresentazione dei punti sul piano. Il settore *Geography*, ad esempio, appare in proiezione prossimo al *Marketing*, ma esso è rappresentato in una dimensione del tutto differente (il quinto asse fattoriale).

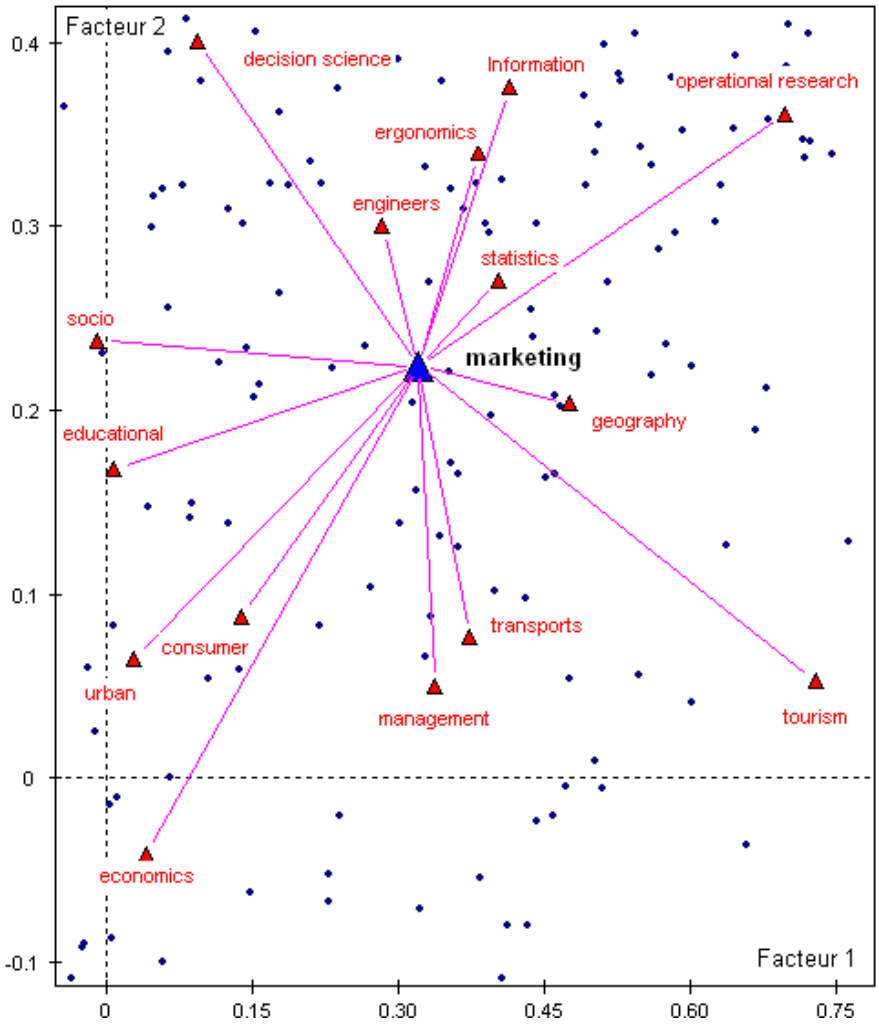

Figura 2.3 - Il Marketing in relazione agli altri settori disciplinari

CAPITOLO 3

3 La Text Classification

La text classification o text categorization (TC) ha l'obiettivo di associare documenti in linguaggio naturale ad un insieme di categorie predefinite.

La categorizzazione automatica ha avuto inizio negli anni '60, e venti anni dopo gli approcci per la costruzione di classificatori automatici avvenivano ancora attraverso tecniche di *knowledge-engineering* consistente nella definizione manuale di regole da parte di esperti del settore.

Negli anni '90, con lo sviluppo tecnologico, la diffusione di internet e la conseguente disponibilità di documenti on-line, la

classificazione automatica di testi ha conosciuto un sempre più crescente e rinnovato interesse e l'approccio dominante nella ricerca è basato sulle tecniche di *machine-learning*: un processo induttivo che automaticamente costruisce regole di classificazione "imparando" da un insieme di documenti già etichettati.

La text classification è usata in molti contesti applicativi quali la ricerca automatica di documenti, la disambiguazione dei significati delle parole, l'organizzazione di risorse sul web, e in generale ogni applicazione che richiede una selettiva organizzazione di documenti (Sebastiani, 2002). D'altronde la TC ha grande utilità anche in settori commerciali legati alla Information and Communication Technology, ma è anche utile nell'estrazione di informazione rilevante che costituisce un punto di partenza per le applicazioni di Text Mining.

3.1 La text categorization

Obiettivo della text categorization è quindi la classificazione di documenti in un fissato numero di categorie predefinite.

Di solito agli algoritmi di TC non è fornita alcuna conoscenza relativa al significato o al contenuto delle categorie e non è data alcuna informazione esogena sui documenti: è necessario che la costruzione delle regole si costruisca esclusivamente sulla conoscenza endogena ricavata dal contenuto dei documenti.

Le fasi di un algoritmo di classificazione sono tre:

1. Codifica del testo: i documenti devono essere "codificati" per poter essere interpretati da un punto di vista statistico. Di solito è utilizzato il sistema "bag-of-word" (v. par. 1.3);
2. Riduzione dimensionale: talvolta è necessario procedere ad una riduzione dimensionale attraverso tecniche statistiche (come ad esempio il Latent Semantic Indexing) per risolvere problemi computazionali.
3. Costruzione dell'algoritmo di classificazione su un training set e validazione su un test set.

Formalmente, dato un insieme di documenti:

$$D = \{d_1, \dots, d_n\}$$

e l'insieme delle categorie:

$$C = \{c_1, \dots, c_r\}$$

La Text Categorization ha il compito di assegnare un valore booleano ad ogni coppia d_i,d_j tale che:

$$\begin{Bmatrix} \langle d_i, c_j \rangle = true \; se \; d \text{ è } classificata \; sotto \; c \\ \langle d_i, c_j \rangle = false \; se \; d \; non \text{ è } classificata \; sotto \; c \end{Bmatrix}$$

Lo scopo è pertanto quello di approssimare la funzione:

$$\bar{\phi}: DxC \rightarrow \{T, F\}$$

che classifica correttamente i documenti in una o più categorie, con una funzione:

$$\phi: DxC \rightarrow \{T, F\}$$

I gradi di coincidenza tra le due funzioni determinano l'efficienza dell'algoritmo di classificazione.

Per valutare la validità del classificatore ci si servirà della matrice di confusione, come quella riportata nella TABELLA 3.1 nel tipico esempio di due categorie.

	YES correct	NO correct
YES predicted	a	b
NO predicted	c	d

TABELLA 3.1 - LA MATRICE DI CONFUSIONE

Sulla diagonale principale della matrice si trovano gli elementi correttamente classificati.

Gli elementi fuori diagonale rappresentano errori di classificazione:

- di omissione quando un elemento appartenente alla classe considerata non vi è inserito;
- di inclusione quando un elemento è assegnato alla classe considerata pur non appartenendovi;

Pertanto la bontà (o accuratezza) del classificatore sarà valutata attraverso il tasso di corretta classificazione generale:

$$TCC = \frac{a + d}{a + b + c + d}$$

L'accuratezza totale dà una misura complessiva di quanto la classificazione è stata ben fatta, ma non distingue tra gli errori commessi nelle diverse classi, che sono trattate tutte allo stesso modo. In alcuni casi si è specificamente interessati ad una classe in particolare, perciò vengono definiti indici di accuratezza legati alle specifiche classi.

L'accuratezza per l' "utente" è definita come il rapporto tra il numero di elementi correttamente classificati nella classe considerata ed il numero di elementi assegnati in totale a quella classe. Per esempio per la classe "yes":

$$AU_{yes} = \frac{a}{a + b}$$

L'accuratezza per il "produttore" è definita invece come il rapporto tra il numero di elementi correttamente assegnati ad una determinata classe ed il numero totale di elementi pertinenti a quella classe nell'insieme di verifica. Per la stessa classe "yes" esso è dato da:

$$AP_{yes} = \frac{a}{a + c}$$

3.2 Le tecniche di classificazione

In questo paragrafo verranno inizialmente riproposte le tecniche di classificazione maggiormente diffuse in letteratura. Successivamente verrà proposta una nuova tecnica basata sulle regole di associazione.

3.2.1 Gli alberi di classificazione

Le procedure di classificazione ad albero (o di segmentazione Binaria o Multipla), sorte agli inizi degli anni '60 per il trattamento di dati numerici, si propongono di classificare un collettivo di individui in gruppi, omogenei al loro interno ed eterogenei tra loro, attraverso una

un serie successiva di suddivisioni dicotomiche pervenendo così alla costruzione di un albero di decisione binaria.

L'idea di base della segmentazione è di partizionare ricorsivamente un insieme di unità statistiche in gruppi sempre più fini (di numerosità inferiore) e sempre più omogenei internamente (rispetto alla distribuzione della variabile di risposta). Si determina in tal modo una partizione finale del gruppo iniziale presente al nodo radice in sottogruppi disgiunti ed esaustivi rappresentati dai nodi terminali dell'albero, a questi ultimi sarà assegnata una classe o un valore di risposta.

La radice dell'albero rappresenta quindi il campione iniziale, i nodi intermedi sono sottoinsiemi del campione iniziale e i nodi terminali le foglie dell'albero, cioè nodi non divisibili che identificano individui appartenenti ad una stessa classe.

La Figura 3.1 riporta un esempio di struttura ad albero binario nella quale è possibile distinguere la radice (T_1), i nodi intermedi (T_2, T_3 e T_7), i nodi terminali (T_4, T_5, T_6, T_8 e T_9), gli split (suddivisione dei nodi: s_1, s_2, s_3, s_4) e i labels (etichette dei nodi terminali: I_1, I_2, I_3, I_4 e I_5).

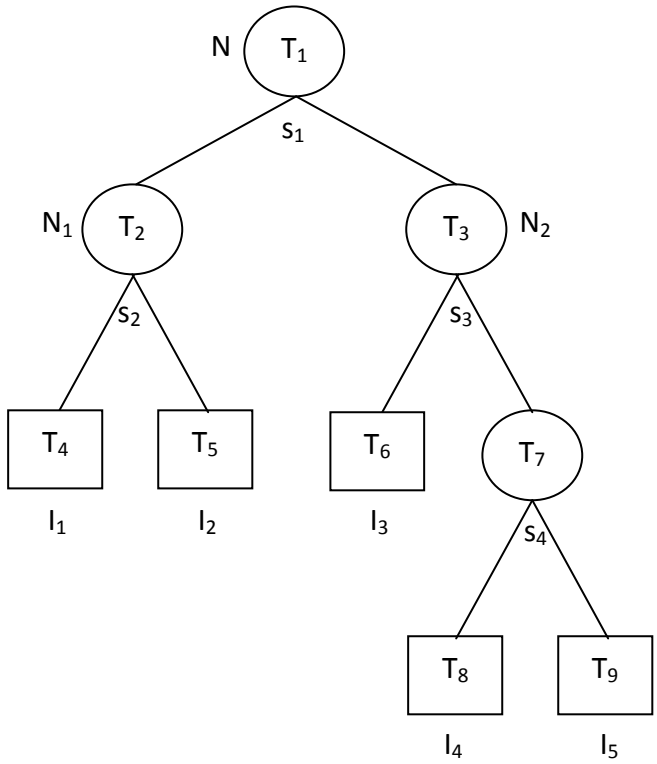

Figura 3.1 - La struttura ad albero

Le tecniche di segmentazione suddividono l'insieme dei documenti in base alle modalità di risposta (categoria predefinita) del predittore attraverso algoritmi iterativi. Ad ogni step verrà individuato il predittore che massimizza il "guadagno informativo", dato di solito dalla differenza tra l'impurità (calcolata attraverso idonee misure: tasso di errata classificazione, eterogeneità, entropia, ecc.) del nodo madre e la somma pesata dell'impurità dei nodi figli. L'algoritmo si

blocca quando il guadagno informativo è inferiore ad una soglia predefinita.

Nel tentativo di sfruttare tutta la informazione disponibile nei dati per discriminare tra i documenti, l'albero risultante tende ad essere molto grande ed inutilmente complesso finendo con lo spiegare l'insieme di apprendimento piuttosto che il fenomeno e risultando poco interpretabile. Pertanto si pone la necessità di ridurre l'albero mediante metodi di semplificazione con l'obiettivo di individuare le branche meno rilevanti o addirittura dannose ai fini della comprensione del fenomeno.

Gli algoritmi più usati per la costruzione degli alberi di classificazione sono il C4.5 (Quinlan, 1993), il CART (Holsen, et al., 1984) e le sue derivazioni FAST (Mola, et al., 1997) e TWO STAGE (Mola, et al., 1992).

3.2.2 La regressione

Il modello di regressione consente di derivare, sulla base di assunzioni verificabili, dai dati osservati una relazione statistica tra una variabile continua dipendente Y ed una o più variabili esplicative (X_1, X_2,…, X_p).

Nel campo della TC l'obiettivo è però ottenere un valore binario della Y che indichi l'appartenenza ad una categoria predefinita.

È Il metodo di regressione logistica binaria consente di individuare una relazione tra la variabile risposta Y dicotomica e le variabili esplicative; pertanto la variabile dipendente è una variabile casuale Bernoulliana.

La probabilità che la Y assuma valore 1 dipende da p repressori (X_1, X_2,…, X_p). Indicando con x il vettore dei regressori si ha:

$$P(Y = 1|X = x) = \pi(x)$$

E di conseguenza:

$$Y \sim Ber\big(\pi(x)\big)$$

In particolare si ha:

$$\pi(x) = \frac{exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right)}{1 + exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right)}$$

I parametri, stimati attraverso il metodo della massima verosimiglianza, consentono al modello di determinare l'appartenenza del documento ad una categoria.

La pesantezza computazionale del modello, che lo rende difficile da attuare per grandi dataset, tipici dell'analisi dei dati testuali, e la necessità di avere più individui che variabili per la stima

dei parametri rendono di fatto poco utilizzato questo modello nella TC.

3.2.3 Le reti neurali

Con rete neurale si intende di solito una rete di neuroni artificiali, che cerca di simulare il funzionamento dei neuroni del cervello umano.

I suddetti neuroni ricevono in ingresso degli stimoli e li elaborano. L'elaborazione può essere anche molto sofisticata, ma in un caso semplice si può pensare che i singoli ingressi vengano moltiplicati per un opportuno valore detto peso, il risultato delle moltiplicazioni viene sommato e se la somma supera una certa soglia il neurone si attiva attivando la sua uscita. Il peso indica l'efficacia sinaptica della linea di ingresso e serve a quantificarne l'importanza, un ingresso molto importante avrà un peso elevato, mentre un ingresso poco utile all'elaborazione avrà un peso inferiore. Si può pensare che se due neuroni comunicano fra loro utilizzando maggiormente alcune connessioni allora tali connessioni avranno un peso maggiore.

3.2.4 Il marcaggio simbolico

Il marcaggio simbolico (MS) (Gettler Summa, 1998) è una tecnica di segmentazione non binaria che ha l'obiettivo di trovare la struttura di associazione in un gruppo G_i ottenuta da una precedente analisi di classificazione.

Questa procedura, infatti, consente di interpretare una Cluster Analysis, condotta sui risultati di un'analisi fattoriale, attraverso la congiunzione logica delle diverse caratteristiche che descrivono i clusters. Il risultato può essere espresso in linguaggio naturale come regole logiche.

3.2.5 Altri metodi

I metodi presentati sono quelli maggiormente utilizzati nella TC. D'altronde come ben indicato da Sebastiani (Sebastiani, 2002) sono molte altre le tecniche adottate tra le quali si ricorda l'analisi discriminante, il Support Vector Machine, metodi basati sull'inferenza Bayesiana e algoritmi genetici.

In generale comunque qualsiasi algoritmo di classificazione supervisionata può, a diverso titolo, applicarsi.

3.3 Una strategia di text mining

Balbi e di Meglio (Balbi, et al., 2004) hanno proposto una strategia di text mining per affrontare l'analisi di corpora strutturati di grandi dimensioni. L'idea si basa sulla possibilità di operare sulla struttura sintattica del documento, concependolo come una struttura complessa costituita da livelli diversi e gerarchici di unità: le frasi e le parole. Di fatto quindi viene considerato un livello gerarchico intermedio tra la parola e il documento, la frase appunto, che può consentire una efficiente estrazione della conoscenza. La strategia proposta consta di più fasi:

1. I documenti sono suddivisi inizialmente in unità gerarchicamente intermedie (es.: frasi);
2. Le unità vengono suddivise in due campioni, un training set ed un test set;
3. L'intervento di un esperto esterno consente di classificare nel training set le frasi interessanti e le frasi non interessanti ai fini dell'indagine;
4. Un algoritmo di classificazione supervisionato, come quelli descritti nel par. 3.2, è considerato per l'individuazione di regole per la selezione di frammenti

(frasi) interessanti; in questa fase le unità considerate sono quelle provenienti da una codifica *bag-of-word;*

5. Le regole individuate al punto precedente sono applicate a tutto il test set;
6. I frammenti classificati non interessanti vengono eliminati dal corpus.

I vantaggi di questa procedura si possono sintetizzare nei seguenti:

- Vantaggi *"computazionali"*: la strategia non richiede enormi sforzi e consente la riduzione del corpus;
- Vantaggi *"applicativi"*: il contenuto del corpus ridotto dall'eliminazione dei frammenti "non interessanti" è più rilevante ai fini dell'indagine.

3.4 Le regole di associazione pesate nella TC

Per la corretta interpretazione del fenomeno linguistico è fondamentale considerare l'importanza di ogni unità coinvolta. La strategia qui proposta è quella di inserire un sistema di pesi specifico nell'ambito della TC. Nelle varie tecniche sopra esposte infatti le unità venivano pesate con schema booleano (presenza/assenza) o bag-of-

word (frequenza). Mancava però un qualsiasi riferimento alla capacità discriminatoria delle parole stesse, basato ad esempio sull'indice *idf*.

Si propone, quindi, l'uso di un sistema di pesi basato sull'indice TF-IDF che tiene conto della rarità relativa del termine rispetto all'intera collezione e quindi della sua capacità discriminatoria.

Da un altro punto di vista è utile individuare algoritmi di classificazione che consentano non solo di individuare relazioni tra la variabile di risposta e il contenuto dei documenti ma anche di ordinare le possibili regole permettendo poi al ricercatore di scegliere soglie di preferenza più adatte al''obiettivo da raggiungere.

Le regole di associazione (Agrawal, et al., 1993), strumento tipico della Market Basket Analysis, permettono di definire degli indicatori sulla validità delle regole, consentendone l'ordinamento.

Nel successivo paragrafo, dopo una breve introduzione alle regole di associazione, verrà presentata la strategia proposta.

3.4.1 Le regole di associazione

Le regole associative possono considerarsi come particolari tecniche di data mining con l'obiettivo di identificare set di attributi, denominati items, che frequentemente ricorrono insieme e formulare regole che caratterizzano le relazione tra loro esistenti.

Formalmente considerato un insieme I=(i_1, i_2, …, i_n) di items ed un database T di transazioni dove ogni transazione "t " è un insieme di attributi contenuto in I in modo tale che t $\subseteq$ I e vi è un unico identificatore associato ad ogni transazione.

Dato un attributo X $\subseteq$ I , una transazione t contiene X $\Rightarrow$X $\subseteq$ t. Una regola di associazione è un'implicazione del tipo X $\rightarrow$Y, dove X $\subseteq$ I e Y $\subseteq$ I e X$\cap$Y=$\emptyset$. Una regola X $\rightarrow$Y è confermata nel database di transazioni con confidenza c, se il c% delle transazioni nel database che contiene X contiene anche Y. Mentre ha supporto *s* se l's% delle transazioni nel database contiene X e Y .

In termini formali quindi data una regola X$\rightarrow$Y (dove X è detto antecedente della regola e Y conseguente) si definisce supporto della regola:

$$Sup(\mathrm{X} \rightarrow \mathrm{Y}) = P(X \cap Y)$$

e confidenza:

$$Conf(\mathrm{X} \rightarrow \mathrm{Y}) = P(Y|X)$$

In maniera sintetica si potrebbe dire che mentre la confidenza misura la forza della regola, il supporto ne misura la significatività statistica.

3.4.2 Le regole di associazione non simmetriche

In alcuni casi è possibile individuare delle gerarchie tra le informazioni di cui si dispone. In particolare, ogni qualvolta si hanno a disposizione due insiemi di variabili legate tra loro da un legame di dipendenza è possibile ricorrere alle regole di associazione non simmetriche(Balbi, et al., 2003) (Balbi, et al., 2005).

Talora è anche possibile considerare *m* variabili categoriche riferite alle singole transazioni. A titolo di esempio si potrebbero considerare le caratteristiche socio-demografiche degli acquirenti nella Market basket analysis. In questo scenario esiste una relazione di antecedenza logica tra i due spazi dal momento che sono le caratteristiche delle unità a determinare il comportamento delle transazioni.

Formalmente una regola di associazione non simmetrica è quindi definita dall'implicazione:

$$(\mathrm{X} \rightarrow \mathrm{Y}) \qquad con\ X \subseteq \mathrm{M}, Y \subseteq \mathrm{I}$$

3.4.3 Le regole di associazione pesate

Tao et al. (Tao, et al., 2003) hanno proposto l'introduzione di un sistema di pesi nelle regole di associazione con l'obiettivo di indirizzare l'estrazione delle regole verso quelle contenenti item più significativi, cioè con un peso maggiore.

Formalmente viene introdotto un sistema di pesi $W=w_1, w_2,\ldots,w_p$ reali e non negativi per cui ogni coppia (x,w) è un item pesato con $x \in I$ e $w \in W$. Una transazione è quindi un insieme di item pesati.

Il peso di un itemset è ovviamente funzione degli pesi degli item inclusi nell'itemset. La formulazione proposta è la media:

$$w_{is} = \frac{\sum_{k=1}^{is} w(i_k)}{|is|}$$

Allo stesso modo il peso di una transazione può essere calcolato come media dei pesi degli item presenti nella transazione:

$$weight(t_k) = \frac{\sum_{i=1}^{|WS_t(t_k)|} weight\big(item(i)\big)}{|WS_t(t_k)|}$$

Una regola è interessante se il suo supporto pesato è superiore ad una soglia prefissata. Il supporto pesato (wsp) di una regola X→Y, con X e Y sub-itemset non vuoti, è il rapporto tra i pesi

delle transazioni che contengono sia X che Y e la somma dei pesi di tutte le transazioni:

$$wsp(XY) = \frac{\sum_{k=1}^{|WS_T| \& (X \cup Y) \subseteq t_k} weight(t_k)}{\sum_{k=1}^{|WS_T|} weight(t_k)}$$

3.5 La strategia

Gli item delle regole sono le forme mentre le transazioni sono i documenti. Ogni documento si compone di più forme. Quindi, dopo aver effettuato tutte le procedure di pretrattamento, si ottiene una tabella lessicale **T** di dimensioni (n,p) il cui generico elemento t_{ij} rappresenta il peso assegnato alla *i-esima* forma nel *j-esimo* documento.

Il peso delle forme è pari ad una misura del TF-IDF (v. par. 1.5.3):

$$TF - IDF = \frac{tf_{ij} \log \frac{D}{df_i}}{\sum_i \left(tf_{ij} \log \frac{D}{df_i} \right)}$$

In tal modo il peso di ogni documento (transazione) sarà costante e uguale ad 1:

$$weight(t_i) = \frac{\sum_{k=1}^{|WS_t(t_i)|} weight\big(item(k)\big)}{|WS_t(t_i)|}$$

Alla matrice T verrà giustapposta una colonna Y con valori dicotomici presenti solo nel training set, così come è illustrato nella figura 3.2.

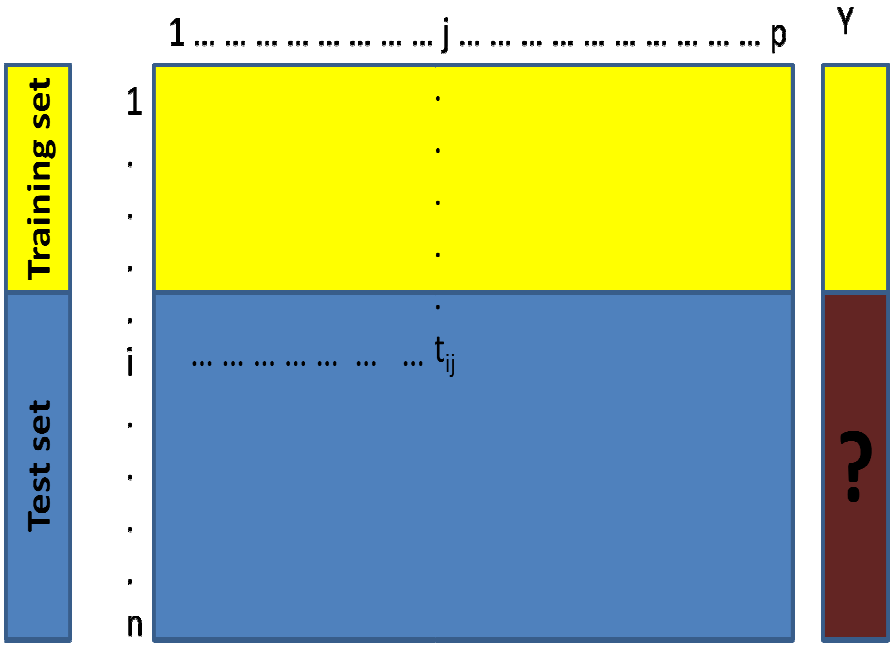

Figura 3.2 - Training set e test set

L'obiettivo è pertanto quello di assegnare la variabile di risposta Y nei documenti del test set.

Per poter applicare le regole associative pesate in un algoritmo di classificazione supervisionato (Larose, 2005) è necessario innanzitutto imporre una modalità della variabile di risposta quale conseguente della regola. In particolare è necessario individuare una gerarchia di interesse per le modalità in modo da concentrarsi su quella ritenuta maggiormente rilevante ai fini dell'indagine.

3.6 L'algoritmo

L'algoritmo (il cui codice in matlab è riportato in appendice) si compone di sette fasi:

1. Calcolo di tutte le possibili combinazioni tra i termini che costituiranno l'antecedente della regola. D'altronde, per ovviare ai problemi computazionali relativi all'elevato numero di combinazioni è necessario individuare un numero massimo di elementi, peraltro compatibili con obiettivo dell'analisi: regole troppo complesse non sarebbero interpretabili;
2. Calcolo della confidenza pesata per tutte le regole:

$$cp(A \rightarrow Y) = \frac{\sum_{i=1}^{|WS_T| \& (A \cup Y) \subseteq t_i} weight(t_i)}{\sum_{i=1}^{|WS_T| \& A \subseteq t_i} weight(t_i)}$$

3. Calcolo del supporto pesato per tutte le regole con cp>k con k>0:

$$sp(A \to Y) = \frac{\sum_{i=1}^{|WS_T| \& (A \cup Y) \subseteq t_i} weight(t_i)}{\sum_{i=1}^{|WS_T|} weight(t_i)}$$

4. Ordinamento decrescente delle regole in base al supporto pesato;
5. Calcolo dei tassi di corretta classificazione generale e per ogni modalità della variabile di risposta considerando ogni supporto calcolato quale supporto minimo;
6. Scelta della soglia minima per il tasso di corretta classificazione della modalità più rilevante della variabile di risposta;
7. Scelta del supporto minimo che massimizza il tasso di corretta classificazione generale;
8. Identificazione delle regole.

3.7 L'analisi delle declaratorie dei corsi di laurea

Uno degli obiettivi del Progetto di Ricerca ad Interesse Nazionale (PRIN 2005) coordinato dal Prof. Luigi Fabbris *"Il mercato delle competenze: metodi statistici per il confronto e l'analisi multidimensionale delle figure professionali offerte e domandate nel terzo settore"*, all'interno del quale è stata sviluppata la proposta, è quello di individuare le competenze offerte dalle università italiane nell'ambito del terzo settore. Per questo scopo assume importanza l'analisi delle declaratorie che possono essere viste come le "*competenze nominali"* offerte dalle università italiane.

Dall'elenco dei corsi di laurea esistenti in Italia sono stati selezionati quei 139 rivolti principalmente al terzo settore. Per terzo settore si è inteso *"quel complesso di organismi e associazioni che contribuisce a produrre beni e servizi di interesse collettivo. Si fonda sul principio della solidarietà e non prevede un profitto individuale o di gruppo; gli eventuali utili vanno reinvestiti a comune utilità"*.

La selezione ha comportato la scelta di 139 corsi di laurea afferenti a 5 classi diverse.

Le declaratorie sono state sottoposte a pre-trattamento con il software TaLTaC2; in particolare sono state individuate le forme

testuali quali unità di analisi e sono state eliminate tutte le forme vuote.

La scelta dei caratteri delimitatori di frase (**; . : ( ) -**) ha consentito la selezione di 2065 frasi da cui è stato selezionato casualmente il training-set con 512 frasi.

Il parere dell'esperto ha consentito di individuare le 104 frasi ritenute interessanti ai fini dell'indagine: si tratta cioè di tutte le frasi, contenute nel training-set, al cui interno erano chiaramente riportate le competenze del settore.

Poiché l'obiettivo dell'indagine era quello di individuare le competenze nominali offerte dal terzo settore si è scelta una soglia di tasso di corretta classificazione delle frasi *"con competenze"* pari a 0,80.

L'algoritmo di cui al par. 3.6 ha consentito di individuare le regole interessanti, in questo caso tutte con un solo termine all'antecedente, con un tasso di corretta classificazione del 77%, come illustrato nella TABELLA 3.2.

In pratica selezionando le frasi che presentano i termini capacità, possedere, sociale, gestione ed informazione permette di individuare ben il 77,218% di frasi interessanti.

X_1		X_2	CCR	CCR|Y=1
capacità	→	Y=1	0,83871	0,48077
possedere	→	Y=1	0,85887	0,66346
sociale	→	Y=1	0,85484	0,72115
gestione	→	Y=1	0,78629	0,77885
informazione	→	Y=1	0,77218	0,80769

TABELLA 3.2 - LE REGOLE E I TASSI DI CORRETTA CLASSIFICAZIONE

CAPITOLO 4

4 L'analisi dei dati testuali come informazione esterna

Molto spesso i dati testuali si trovano, ovvero vengono rilevati, in corrispondenza di dati numerici. In questi casi, oltre ai tipici problemi di scelta di unità e pesi, ci si pone il problema del loro utilizzo combinato.

Obiettivo di questo capitolo è evidenziare il ruolo che possono avere i dati testuali in ausilio a dati numerici e come la scelta delle unità e dei pesi sia rilevante a tale scopo.

Nella fattispecie viene proposto l'utilizzo di dati testuali come informazione esterna in un'analisi non simmetrica delle preferenze nel campo della segmentazione del mercato.

Un questionario all'uopo disegnato permette la rilevazione di dati numerici e testuali: in questo caso l'informazione testuale, proveniente da domande a risposta aperta, è codificata in presenza/assenza in considerazione della brevità dei documenti ed arricchisce le informazioni ottenute con la sola tecnica quantitativa. Un'applicazione al mercato degli orologi evidenzia le potenzialità della strategia adottata.

4.1 Introduzione

La definizione di prodotto ideale ottenuta con la conjoint analysis (Green, et al., 1990) (CA) dipende, tra l'altro dal metodo scelto per la rilevazione delle preferenze. La descrizione del prodotto ideale espressa in linguaggio naturale può fornire nuove prospettive alle strategie di marketing.

La strategia proposta è quella di introdurre una domanda a risposta aperta in un questionario tipicamente utilizzato per la CA con l'obiettivo di verificarne e migliorarne i risultati. In particolare la proposta è quella di considerare la descrizione testuale come

informazione esterna(Takane, et al., 1991) nell'analisi delle preferenze.

4.1.1 La conjoint analysis

La conjoint analysis è una tecnica statistica di analisi multivariata che permette di individuare e misurare il valore relativo attribuito dai clienti agli attributi di un prodotto/servizio al fine di ottimizzare la definizione del pacchetto di offerta (v. cap 2 per un'ampia rassegna bibliografica).

I fondamenti concettuali della metodologia risiedono nelle considerazioni sviluppate da Lancaster (Lancaster, 1966) e sostenute successivamente anche da Lambin (Lambin, 1996) in margine alla teoria del consumatore. Questa tesi sostiene che l'utilità d'uso di un bene è funzione dalle singole caratteristiche che lo compongono. È possibile quindi scomporre l'utilità che un consumatore ricava da un prodotto/servizio in varie utilità separate che traggono origine dalle diverse caratteristiche di quel bene.

Dal punto di vista matematico obiettivo della CA è quello di scomporre le valutazioni globali, espresse in corrispondenza di varie combinazioni (profili) di modalità (livelli) degli attributi (fattori) di un prodotto, nelle valutazioni delle singole modalità degli attributi stessi.

Poiché l'utilizzo di tutte le combinazioni possibili tra i vari livelli degli attributi (piano fattoriale completo) risulta particolarmente gravoso sia per gli intervistati, sia per il costo dell'operazione, i profili sottoposti a giudizio vengono individuati sulla base della teoria della programmazione degli esperimenti.

Formalmente **X** (*S,L*) è la matrice del disegno sperimentale (profili x livelli) partizionata in H matrici indicatrici $\mathbf{X}_h$ giustapposte (h=1,…,H), ognuna delle quali riferita ad un singolo attributo. **Y** (*S,G*) è la matrice delle preferenze espresse dai giudici che presenta in riga i P profili ed in colonna i g giudici.

Da un punto di vista statistico il modello classico della CA è un modello di regressione multipla multivariata:

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E}$$

Dove **E** (*S,G*) è la matrice degli errori e **B** (*L,G*) è la matrice dei *G* coefficienti individuali di utilità parziale associati agli L livelli degli attributi:

$$\widehat{\boldsymbol{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y}$$

Dove $()^{-}$ è l'inversa generalizzata di Moore e Penrose, poiché la matrice **X'X** non è di rango pieno.

4.1.2 L'analisi non simmetrica delle preferenze

Un approccio multidimensionale alla CA (Lauro, et al., 1998) è stato proposto per sintetizzare e visualizzare i coefficienti di utilità parziale in un sottospazio.

L'approccio, che da un punto di vista geometrico è basato sull'analisi in componenti principali in rapporto ad un sottospazio di riferimento, allo scopo di decomporre la varianza delle preferenze dei giudici spiegata dagli attributi, considera gli auto valori della matrice:

$$(\boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y})z_\alpha = \mu_\alpha z_\alpha \quad (\alpha = 1, \dots, m; m = rank(X))$$

Dove z_α e μ_α sono rispettivamente l'α-esimo auto valore e autovettore della matrice (**Y'X**(**X'X**)$^{-}$**X'Y**).

4.1.3 Un approccio fattoriale ai coefficienti di regressione

L'approccio fattoriale alla CA è stato esteso da Giordano e Scepi nel 1999 (Giordano, et al., 1999) considerando la matrice **Z** (*G,C*) delle caratteristiche socio-demografiche dei giudici quale informazione esterna alle righe della matrice **Y**. In altre parole propongono un

approccio fattoriale ai coefficienti di regressione della CA dove le classi di individui con preferenze ed abitudini simili sono proiettati in sottospazi di riferimento.

In pratica vengono definiti due insiemi di coefficienti di regressione: i coefficienti B definiti come utilità parziali spiegate dalle caratteristiche del prodotto e i coefficienti D definiti come utilità parziali spiegate dalle caratteristiche socio-demografiche. Pertanto viene introdotto un altro modello di regressione:

$$\boldsymbol{Y} = \boldsymbol{D}\boldsymbol{Z}' + \boldsymbol{F}$$

Dove **Z** (*G,C*) è la matrice che incrocia le cratteristiche socio-demografiche dei giudici.

La proposta è quella di decomporre in valori singolari la matrice di inter-relazione:

$$\boldsymbol{\Theta} = (\boldsymbol{Z}\boldsymbol{Z}')^{-}\boldsymbol{Z}\boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})'$$

Il cui generico valore rappresenta il parametro stimato della relazione tra le caratteristiche socio demografiche.

4.2 La conjoint analysis con informazione testuale esterna

La proposta è quella di inserire i dati testuali come informazione esterna per poter integrare i risultati ottenuti con la CA.

Allo scopo è stato proposto un questionario diviso in tre sezioni:

- Nella prima sezione ogni giudice indica le proprie caratteristiche socio-demografiche;
- Nella seconda sezione ogni giudice descrive le proprie abitudini in relazione all'uso del prodotto servizio oggetto dell'indagine;
- Nella terza sezione i giudici rispondono ad una domanda aperta del tipo: "*qual è il tuo prodotto ideale?*"
- Infine, nella quarta sezione, i giudici emettono un giudizio quantitativo sui profili creati come in una CA classica.

4.2.1 La codifica dell'informazione testuale

La codifica dell'informazione testuale contenuta nella terza sezione del questionario proposto è di rilevante importanza. Anche in questo caso la scelta di unità e pesi non può prescindere dagli obiettivi dell'analisi e dal tipo di dati rilevati. Le risposte a domande aperte, se non sottoposte a testimoni privilegiati, si caratterizzano innanzitutto per la loro brevità. L'utilizzo di una codifica bag-of-word con un sistema di pesi ancorato quindi sulla frequenza della *i-esima* forma sul *j-esimo* documento con frequenze assolute (relativamente) basse tenderebbe a dare pesi eccessivi ad alcuni vettori in conseguenza di scarti di frequenza molto bassi. E molto spesso tali scarti dipendono più dallo stile di scrittura, ovvero dal livello culturale, del rispondente che a differenze sostanziali delle risposte date. In questa ottica è opportuno adottare un sistema di pesi booleano.

Da un altro punto di vista bisogna considerare che l'obiettivo della domanda aperta è quello di individuare nel dettaglio le preferenze dei giudici. In tale ottica la scelta delle unità non può non ricadere che sulle forme testuali (vedi par. 1.3.4)

4.2.2 La struttura dei dati

Sia **Y** (*S,G*) la matrice, centrata sia per riga che per colonna, delle preferenze rilevate nella quarta sezione del questionario i cui valori consistono nel giudizio espresso dai *G* giudici rispetto agli *S* profili. Sia **X** la matrice del disegno sperimentale in cui le *L* colonne sono i livelli degli attributi. Sia **Z** la matrice di presenza/assenza relativa alle informazioni sui giudici rilevate nelle prime due sezioni del

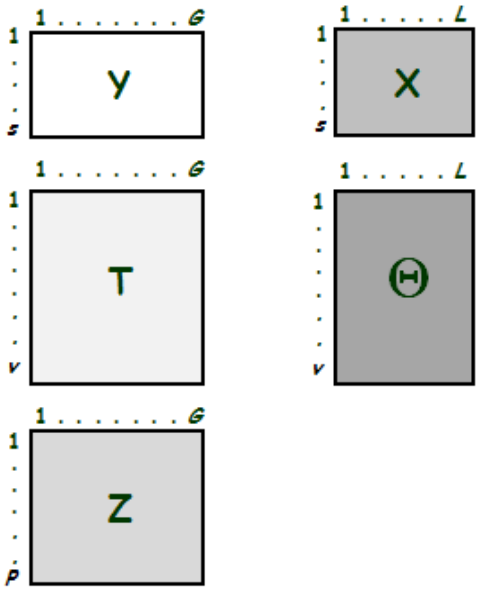

Figura 4.1 - La struttura dei dati

questionario.

Sia **T** la tabella lessicale che incrocia le *G* descrizioni rilevate nella terza parte del questionario e le *V* parole del vocabolario in cui

con il generico elemento t_{ij} ha valore 1 se la *i-esima* parola è presente nel *j-esimo* documento e 0 altrimenti.

La Figura 4.1 riepiloga la struttura dei dati.

4.2.3 La strategia adottata

L'informazione testuale può essere considerata come informazione esterna così come proposto da Takane e Shibayama (16). Introducendo **T** come vincolo lineare sulla matrice **B** può essere costruita la matrice **Q** di dimensioni (*V,L*):

$$\mathbf{Q} = (\boldsymbol{TT}')^{-}\boldsymbol{TB}' = \boldsymbol{W}\Xi\boldsymbol{V}'$$

Dove q_{il} è la stima del parametro che collega il livello dell'attributo alla descrizione testuale del giudice sul prodotto ideale.

Applicando una decomposizione in valori singolari alla matrice **Q** con i vincoli di normalità:

$$\mathbf{W}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{W} = \mathbf{I}_{\mathbf{L}}$$

$$\mathbf{V}(\boldsymbol{TT}')\boldsymbol{V} = \mathbf{I}_{\mathbf{V}}$$

In questa ottica è possibile visualizzare le parole della descrizione insieme ai livelli della CA su uno stesso piano fattoriale. Le informazioni in **Z** possono inoltre essere proiettate in supplementare (17) arricchendo così l'interpretazione dei risultati della CA.

4.3 Un'applicazione al mercato degli orologi

Per verificare la validità della proposta è stata svolta un'indagine sul mercato degli orologi da polso. In particolare un questionario come quello descritto nel par. 4.2 è stata sottoposta ad un campione di 150 individui, stratificati rispetto all'età ed al genere, residenti nella provincia di Napoli.

Pertanto il questionario prevedeva nella prima sezione la richiesta delle caratteristiche socio-demografiche, nella seconda sezione domande inerenti l'uso del prodotto (numero di orologi posseduti, interesse verso le marche, le ragioni d'acquisto, ecc.) e nella terza sezione, oltre all'ordinamento di profili di prodotti individuati attraverso un piano frazionato ridotto composto da attributi e livelli (Figura 4.2) individuati attraverso conoscenza esperta, è stata inserita la domanda a risposta aperta: *"descrivi il tuo orologio da polso ideale".*

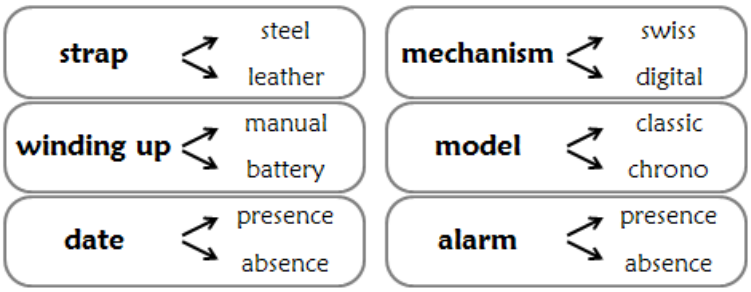

Figura 4.2 - I fattori e i livelli della CA

L'analisi delle risposte aperte, avvenuta come indicato nel par. 4.2.1, ha permesso l'individuazione di 215 forme testuali.

La figura Figura 4.3 riporta il primo piano fattoriale che spiega il 44% della variabilità totale. Il primo asse oppone orologi meccanici (*mechanic*) con carica manuale *(manual_WUP)* ad orologi digitali *(digital)* con batteria *(battery_WUP)*. È chiaro quindi che questo asse contrappone orologi classici con orologi moderni.

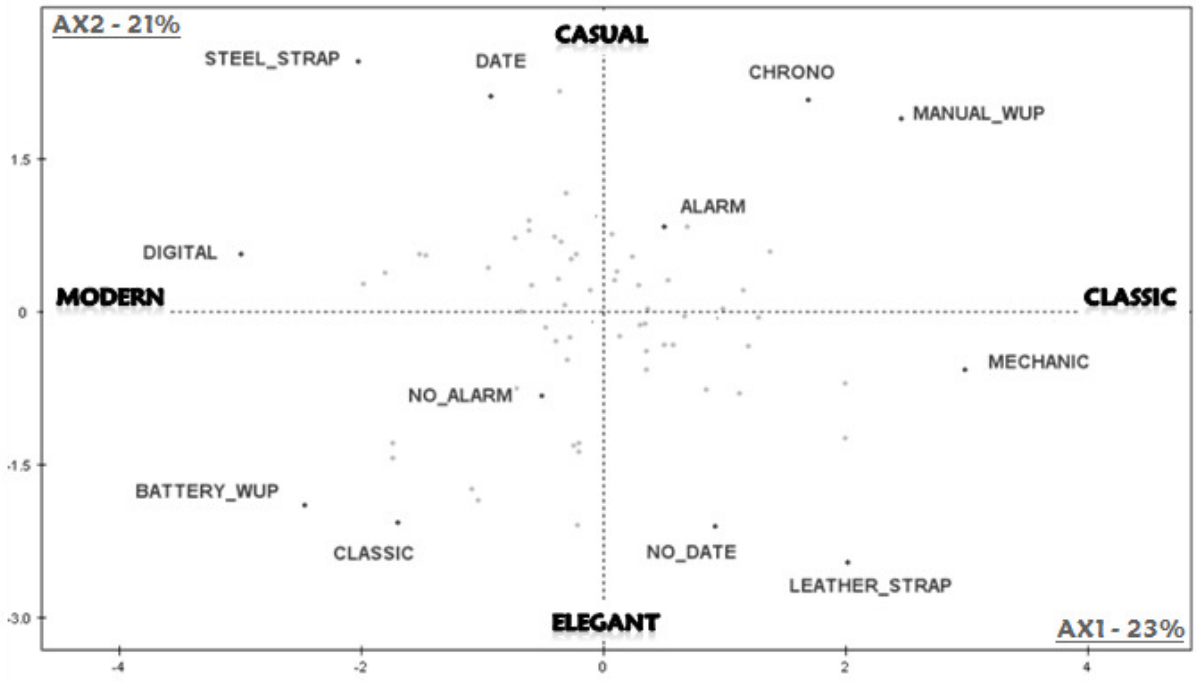

Figura 4.3 - Il primo piano fattoriale

Nel secondo asse fattoriale possiamo si nota invece la contrapposizione tra orologi *casual* e orologi eleganti in funzione della presenza (sopra) e dell’assenza (sotto) di allarme e data e al tipo di cinturino *(steel and leather strap)*.

La segmentazione del mercato può essere rielaborata considerando l’informazione testuale rilevata nell’indagine (Figura 4.4).

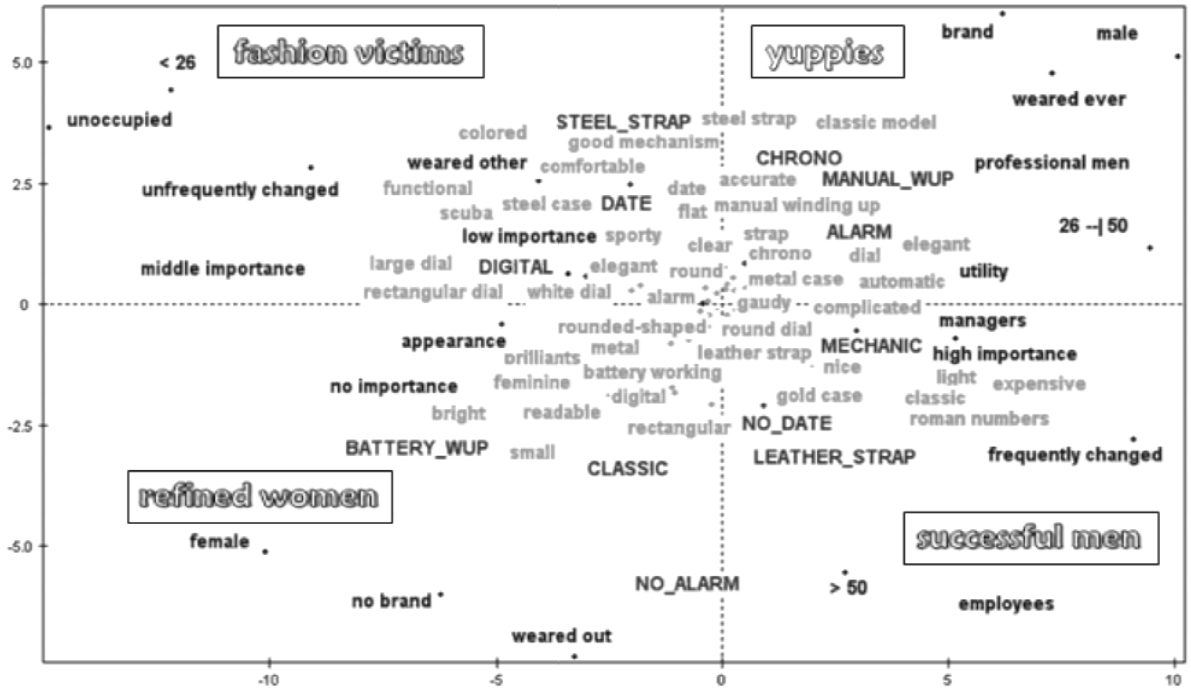

Figura 4.4 - La segmentazione del mercato

Nella Figura 4.4 si ripropone il primo piano fattoriale proiettando oltre ai livelli dei fattori della CA (inlettere maiuscole) sia le descrizioni verbali (in chiaro) sia, in supplementare, le caratteristiche socio economiche demografiche (in nero minuscolo).

È possibile così individuare quattro segmenti di mercato:

- *"Giovani informali"*: giovani (fino a 26 anni) che vogliono un orologio comodo e sportive con data e cinturino d'acciaio.
- *"Donne raffinate"*: donne che vogliono un orologio classico con batteria; possibilmente piccolo e femminile, quasi un gioiello.

- *"Uomini di successo":* sono adulti che cambiano sesso orologio. Vogliono un orologio elegante con cinturino in pelle, numeri romani e senza data. Per questo segmento non è importante se l'orologio è costoso.
- *"Yuppies"*: professionisti che vogliono un chrono con batteria manual e allarme.

4.4 Conclusioni

La strategia proposta consente di verificare e migliorare il processo di segmentazione del mercato attraverso l'analisi congiunta di dati strutturati (numerici) e non strutturati (testuali). L'interpretazione

L'interpretazione dei dati ottenuti attraverso un approccio classico alla CA è stata infatti arricchita dalle forme testuali elaborate dalla descrizione verbale del prodotto ideale facendo emergere anche l'interpretazione soggettiva dei giudici sulle diverse caratteristiche individuate.

CAPITOLO 5

5 Il confronto tra corpora

Un problema complesso nell'analisi dei dati testuali è lo studio delle relazioni tra corpora differenti. Può accadere infatti che si sia interessati a verificare delle differenze esistenti, ad esempio, a risposte a domande aperte sottoposte a due campioni differenti o quando si voglia confrontare due imprese in funzione delle loro *mission* espresse in linguaggio naturale. Nel caso di relazione di antecedenza/conseguenza tra i due linguaggi, si può ricorrere all'Analisi Non Simmetrica(Balbi, 1995) delle Corrispondenze per visualizzare le forme caratteristiche dei due vocabolari e allo stesso tempo per valutare il livello di predicibilità di un linguaggio rispetto all'altro (Grassia, et al., 2004).

Il problema si può porre però anche in termini di differenze di linguaggio utilizzato. In questo capitolo si propone una metodologia basata sui dati simbolici per la risoluzione di questi tipi di problemi.

5.1 Gli oggetti simbolici

Una definizione di unità statistiche in termini di dati simbolici è stata proposta da Diday nel 1987 (Diday, 1987). Tale definizione trova il suo campo di applicazione quando le unità statistiche non sono dei semplici individui, caratterizzati per una sola modalità di ogni descrittore, ma delle unità complesse che difficilmente possono essere riprodotte in una matrice unità x variabili per essere analizzate con le classiche tecniche dell'analisi dei dati.

Formalmente un oggetto simbolico (Bock, et al., 2000) è definito da una terna s = (a, R, d), dove d = (d_1, . . . , d_j , . . . , d_p) è la descrizione dell'oggetto (l'intensione), costituita dai valori assunti da un insieme di p descrittori (Y_1, . . . ,Y_j , . . . ,Y_p), a è la funzione di riconoscimento, e R = (R_1, . . . ,R_j , . . . ,R_p) è la relazione su cui si basa il confronto tra la descrizione fornita a livello concettuale, in intensione, da d e le singole osservazioni.

I descrittori di un oggetto simbolico possono essere di tipo nominale, continuo o discreto, e presentare più modalità o valori per ciascun oggetto. La funzione booleana a assume valori {vero, falso} e

consente d'individuare gli elementi che appartengono all'insieme di descrizione d e che definiscono ext(**s**), l'estensione dell'oggetto.

5.2 Gli oggetti testuali

Le prime applicazioni degli oggetti simbolici ai dati testuali risalgono agli inizi di questo decennio. In particolare i contributi di Balbi, Bolasco e Verde (Balbi, et al., 2002) (Bolasco, et al., 2002) propongono un'analisi su dati derivati dalla definizione formale di concetti e dalla loro composizione in termini di unità elementari. Tali concetti si fondano sulle strutture lessicali complesse basate su tutte le flessioni possibili di un lemma.

Di fatto è possibile costruire degli oggetti testuali attraverso l'elaborazione di informazioni elementari rilevate nella fase di raccolta dei dati, ovvero ricorrendo al giudizio di un esperto.

L'utilizzo degli oggetti nel campo dell'analisi dei dati testuali permette la soluzione di problemi relativi alla presenza di matrici sparse nell'analisi. Se, infatti, l'aggregazione dei documenti, quando necessaria, avviene generalmente attraverso una variabile di interesse conosciuta a priori, l'aggregazione del vocabolario si ottiene attraverso le varie fasi di pre-trattamento del testo e con l'eliminazione delle forme a bassa frequenza.

Gli oggetti testuali operano una riduzione delle forme organizzandole in variabili modali dove ogni singola forma diventa una modalità della variabile corrispondente.

Pertanto dalla matrice lessicale **T**, già precedentemente definita, si ottiene una matrice **S**, detta matrice simbolica, con in riga i documenti (ovvero una loro classificazione) ed in colonna le variabili modali, ognuna con s_{m_i} modalità corrispondenti alle forme testuali in essa inserite. Il *j*-esimo oggetto simbolico della matrice S è definito come:

$$o_j = \bigwedge_{k=1}^{Y} \left[S_k = \{ s_{k,m}(f_{k,m}) \}_{m=1,2,\ldots,m_i} \right]$$

dove $f_{k,m}$ è la frequenza relativa di $s_{k,m}$, *m*-esima modalità della variabile S_k.

Questo modo di operare, se da un lato consente la riduzione del Vocabolario della matrice **T**, dall'altro può consentire di confrontare corpora provenienti da diverse fonti. In particolare si può operare in due modi:

- si può ricondurre il vocabolario agli oggetti testuali di interesse e su questi calcolare misure di dissimilarità;

- si possono essere costruiti oggetti simbolici sui documenti delle tabelle lessicali prendendo in considerazione il solo vocabolario comune.

È possibile così calcolare misure di dissimilarità tra i documenti dei due corpora.

5.3 I confronti tra corpora

Da un punto di vista metodologico è impossibile confrontare tabelle lessicali provenienti da due corpora in quanto le corrispondenti matrici lessicali non hanno in comune né le righe né le colonne.

Una possibile strategia di analisi è la seguente:

1) Considerare i due corpora come generati da un vocabolario comune;
2) Individuare k temi, mediante conoscenza esperta, ed aggregare le forme del vocabolario "comune" interessanti ai fini dell'obiettivo dell'indagine;
3) Costruire oggetti simbolici relativi ai documenti di entrambe le matrici ed esprimerli in forma modale rispetto alle forme;
4) Calcolare le dissimilarità fra gli oggetti simbolici delle due matrici.

SI ritiene che nell'attuazione di questa strategia, rivolta alla ricerca di unità di ordine superiore quali gli oggetti simbolici, ma con riferimento alle forme comuni nei due corpora, sarà necessario considerare quali unità le forme testuali così come descritte nel paragrafo 1.3.4.

Da un punto di vista formale si rilevano due tabelle lessicali $\mathbf{T_1}$ e $\mathbf{T_2}$ a partire dai due corpora tra i quali si vuole calcolare la dissimilarità. Il sistema di pesi adottato è quello booleano per cui in ciascuna tabella lessicale l'elemento generico è la presenza/assenza della *i*-esima forma testuale nel documento *j*-esimo.

Al passo successivo si individuano le H unità comuni alle matrici lessicali $\mathbf{T_1}$ e $\mathbf{T_2}$ ed interessanti ai fini dell'indagine. Si costruiscono così le matrici di presenza/assenza del vocabolario comune $\widetilde{\boldsymbol{T_1}}$ e $\widetilde{\boldsymbol{T_2}}$.

Nella fase seguente vengono individuati gli oggetti simbolici relativi ai documenti. In particolare vengono definite le matrici $\widetilde{\boldsymbol{O_1}}$ e $\widetilde{\boldsymbol{O_2}}$ le cui righe sono gli *S* oggetti simbolici modali e le colonne sono le *H* unità lessicali comuni.

Il generico elemento delle matrici $\widetilde{\boldsymbol{O_1}}$ e $\widetilde{\boldsymbol{O_2}}$ è dato dalla media dei valori delle forme appartenenti all'oggetto considerato:

$$o_{jm} = \frac{\sum_{i=1}^{n_{classe_j}} z_i}{n_{classe_j}}$$

È così possibile calcolare una misura dissimilarità (Bock, et al., 2000) (Bruzzese, et al., 2002), basata sul confronto tra le rispettive distribuzioni marginali, tra gli oggetti della matrice $\widetilde{\boldsymbol{O_1}}$ e gli oggetti della matrice $\widetilde{\boldsymbol{O_2}}$:

$$d(\widetilde{\boldsymbol{O_1}}, \widetilde{\boldsymbol{O_2}}) = \frac{1}{m_k} \sum_{k=1}^{m_k} \frac{\left| f_{s_k}(\widetilde{\boldsymbol{O_1}}) - f_{s_k}(\widetilde{\boldsymbol{O_2}}) \right|}{max\left(f_{s_k}(\widetilde{\boldsymbol{O_1}}); f_{s_k}(\widetilde{\boldsymbol{O_2}}) \right)}$$

L'indice di dissimilarità assume valore 0 se i due oggetti presentano per ogni variabile la stessa distribuzione di frequenza e valore 1 se differiscono completamente.

5.4 Il mercato delle competenze

Obiettivo successivo all'individuazione delle competenze offerte dalle università italiane del PRIN 2005 (v. par. 3.7) coordinato dal Prof. Luigi Fabbris è valutare la rispondenza fra la domanda di competenze e di professionalità e il tipo di formazione che il sistema universitario offre.

Sono state pertanto realizzate 19 interviste semi-strutturate a testimoni privilegiati con l'obiettivo di far emergere la domanda di

competenze e nuove professionalità richieste dal terzo settore. Si è così individuato. Dalle risposte a domanda aperta, dopo il necessario pre-trattamento per l'identificazione delle forme testuali, si è potuta così costruire la tabella lessicale delle competenze domandate dal mercato. Si è provveduto poi ad aggregare i 19 individui in funzione della dimensione (misurata attraverso il numero di addetti) della struttura rappresentata; sono stati così costruiti i tre oggetti "testimoni di piccole dimensioni", "testimoni di medie dimensioni" e "testimoni di grandi dimensioni".

Dal lato dell'offerta i 135 corsi di laurea già indicati nel par. 3.7 sono stati aggregati in 16 oggetti in funzione sia della classe di laurea di appartenenza, sia della denominazione prescelta.

Al fine di poter confrontare i due corpora sono state selezionate le 95 forme testuali comuni, con l'esclusione delle forme vuote, comuni ai due vocabolari.

Le forme sono state poi, con l'ausilio di un esperto, divise in 6 classi logiche:

- Attitudini personali (saper essere): *coerenza, disponibilità*…;
- Capacità (saper fare): *interagire, dirigere, analizzare*…
- Conoscenze (sapere): *psicologia, diritto, pedagogia*…
- Termini generali: *competenze, conoscenze, formazione professionale*…
- Professionalità: *mediatore, animatore*…

- Servizi: *immigrati, disagiati, minori…*

Si è potuto così procedere alla giustapposizione delle due tabelle, come evidenziato nella Figura 5.1.

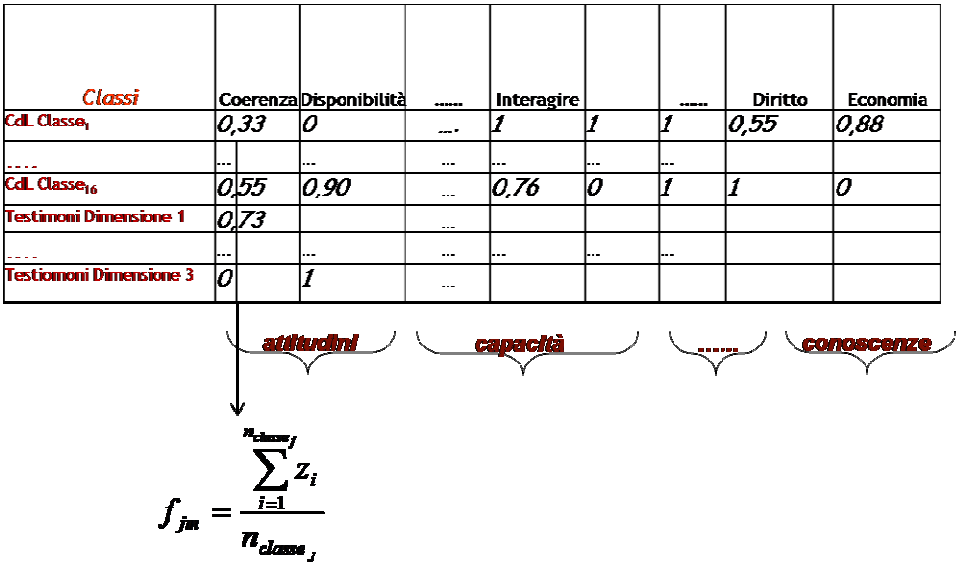

Figura 5.1 - Gli oggetti simbolici

La Figura 5.2 mostra la matrice delle dissimilarità tra la domanda (in riga) e l'offerta (in colonna) di competenze in relazione alle classi individuate.

CLASSI	servizio sociale	servizio sociale internazionale	scienze sociali e cooperazione e sviluppo	scienze sociali e cooperazione e sviluppo	cooperazione e sviluppo	economia noprofit	discipline e economiche e aziendali	scienze per la pace	scienze dell'educazione	infanzia	educatore sociale, culturale e territoriale	professioni educative	educatore interculturale	formatore promozione e sviluppo risorse umane	scienze dell'educazione e della formazione	scienze dei processi formativi
Numerosità interna	44	2	12	2	4	4	1	2	16	9	8	12	2	5	9	3
attitudini_1	0,43188	0,14583	0,19896	0,23958	0,27188	0,22396	0,27083	0,26458	0,33475	0,09766	0,36458	0,34422	0,33333	0,14583	0,41059	0,39583
capacità_1	0,41053	0,23526	0,43333	0,34263	0,3774	0,38574	0,23397	0,36346	0,47989	0,28405	0,36409	0,42998	0,34744	0,36406	0,48275	0,51058
conoscenza_1	0,45334	0,32339	0,42986	0,31102	0,38324	0,37254	0,35484	0,47715	0,49634	0,30956	0,37788	0,43434	0,39866	0,3833	0,50862	0,4578
generale_1	0,5919	0,35119	0,56245	0,56548	0,56746	0,50099	0,69048	0,57738	0,58624	0,51171	0,49235	0,5	0,39762	0,45655	0,56002	0,44762
professione_1	0,54962	0,50476	0,5131	0,40476	0,48651	0,46429	0,4881	0,54167	0,41916	0,51736	0,4951	0,54589	0,48095	0,48782	0,54498	0,43571
servizi_1	0,50094	0,3451	0,41085	0,41373	0,39951	0,34926	0,43137	0,44559	0,40396	0,38268	0,49272	0,47304	0,38186	0,32497	0,4254	0,45245
attitudini_2	0,52656	0,27083	0,44792	0,36458	0,52083	0,34896	0,39583	0,48958	0,49579	0,24479	0,38988	0,52312	0,45833	0,27083	0,4974	0,39583
capacità_2	0,31968	0,35577	0,36394	0,42115	0,41341	0,44396	0,38141	0,41955	0,43151	0,3237	0,3553	0,37999	0,41442	0,42185	0,41613	0,54359
conoscenza_2	0,5207	0,32473	0,42472	0,20806	0,36398	0,34149	0,24731	0,39624	0,57402	0,30009	0,40296	0,50842	0,38548	0,40022	0,57849	0,4707
generale_2	0,58708	0,46429	0,59828	0,49405	0,55992	0,44643	0,83333	0,74405	0,59351	0,50802	0,56719	0,70476	0,57024	0,54476	0,68542	0,60833
professione_2	0,58845	0,63095	0,41483	0,44643	0,48909	0,48909	0,5	0,52976	0,39156	0,35863	0,36074	0,50795	0,3619	0,29214	0,57184	0,30476
servizi_2	0,55852	0,26912	0,29724	0,25	0,29412	0,19771	0,22059	0,40686	0,55609	0,45343	0,60011	0,55508	0,22794	0,44549	0,56878	0,48088
attitudini_3	0,40093	0,24107	0,32708	0,24033	0,39658	0,20424	0,28571	0,37946	0,46801	0,31052	0,25122	0,42634	0,33408	0,24107	0,44647	0,23532
capacità_3	0,35771	0,39267	0,40023	0,45746	0,4476	0,46696	0,40659	0,46346	0,42938	0,30547	0,33103	0,38878	0,45888	0,42668	0,42033	0,55458
conoscenza_3	0,46191	0,312	0,36014	0,24979	0,28961	0,30679	0,3341	0,40486	0,52194	0,33094	0,34763	0,48756	0,32295	0,32248	0,54351	0,49409
generale_3	0,58518	0,49082	0,52833	0,55616	0,37466	0,48963	0,70408	0,55485	0,45528	0,52612	0,44625	0,62132	0,50298	0,30969	0,63278	0,54082
professione_3	0,5831	0,60306	0,56347	0,54974	0,625	0,6014	0,53061	0,53444	0,47076	0,56035	0,50055	0,4991	0,50485	0,51017	0,53496	0,48135
servizi_3	0,4192	0,43309	0,46512	0,47234	0,50683	0,4475	0,48739	0,49247	0,4284	0,48319	0,45574	0,46075	0,47847	0,34819	0,48567	0,42248

Figura 5.2 - la matrice delle dissimilarità

Per ottenere una sintesi di tale matrice si è provveduto a calcolarne le medie (pesate in funzione della numerosità delle classi), come illustrato nella Figura 5.3.

	Piccola Dimensione	Media Dimensione	Grande Dimensione
attitudini	0,33	0,46	0,37
capacità	0,41	0,37	0,39
conoscenza	0,43	0,47	0,43
generale	0,55	0,59	0,54
professione	0,51	0,48	0,55
servizi	0,44	0,49	0,44

Figura 5.3 - la dissimilarità tra domanda e offerta

La tabella mostra una forte differenza tra la domanda e l'offerta di competenze. I valori più elevati , ma comunque appena superiori a 0,5, sono presenti nella categoria generale, quella più propriamente riguardante i linguaggi utilizzati. Per le categorie maggiormente rilevanti quali *attitudini*, *capacità* e *conoscenze* esistono, evidentemente, delle differenze enormi che mostrano come l'università abbia difficoltà ad adeguarsi alle sempre mutevoli esigenze del mercato.

5.5 Conclusioni

L'analisi adottata ha dimostrato le potenzialità degli oggetti per confrontare corpora differenti. Tuttavia l'analisi, trattando il solo vocabolario comune, esprime probabilmente le sole differenze di linguaggio utilizzate nei due diversi corpora. L'uso degli oggetti testuali

(Balbi, et al., 2002) potrebbe consentire l'utilizzo di tutte le forme consentendo nuove e forse meno "linguistiche" interpretazioni ai dati.

Da un punto di vista metodologico si potrebbe inoltre introdurre un sistema di pesi nel calcolo dell'indice di dissimilarità, quale ad esempio il TF-IDF (v. par. 1.5.3), che tenga conto del potere discriminante dei termini.

CAPITOLO 6

6 Conclusioni

La presenza dello statistico nel processo di text mining, come illustrato nei capitoli precedenti, assume un ruolo centrale soprattutto nella fase di codifica dei dati. La scelta di unità e pesi è una fase assolutamente rilevante nel processo; da queste scelte sono fortemente condizionati i risultati e quindi queste scelte devono dipendere dagli obiettivi della ricerca.

la scelta del tipo di "forme" (grafiche, lemmatizzate o testuali) da utilizzare nell'analisi dipende essenzialmente dagli obiettivi che ci si pone considerando i vincoli economici e di tempo. Bisognerà valutare, di caso in caso, se il vantaggio derivante da un'approfondita pulizia del testo, in termini di raggiungimento degli obiettivi preposti, sia

maggiore dei tempi e dei costi sostenuti. Si tratta cioè di valutare l'efficacia e l'efficienza del processo di pre-trattamento. In quest'ottica assumono un ruolo rilevante gli studi di sociolinguistica e il progresso tecnologico (soprattutto nel campo dell'informatica) che stanno consentendo lo sviluppo di software in grado di automatizzare e semplificare il processo di pre-trattamento e, quindi, di "individuazione" delle forme testuali.

La tesi vuole però aprire a nuovi spunti di riflessione: in letteratura sono infatti emerse proposte di gerarchizzazione delle unità. In sostanza si concepisce il documento come una struttura complessa, costituita da unità di ordine inferiore. Si può pertanto passare in grado discendente dal documento, alla frase (par. 3.3) o all'oggetto testuale (par. 5.2), alle forme testuali (par. 1.3.4) e, infine, alle forme grafiche (par. 1.3.1). La scelta del "*livello*" dipende essenzialmente dagli obiettivi dell'indagine e dal tipo di analisi da applicare. E ovviamente non puo' prescindere dal sistema di pesi da adottare.

È pertanto la giusta "combinazione" unità-pesi che consente una efficace ed efficiente estrazione della conoscenza da dati testuali.

Appendice: il codice MATLAB per le regole di associazione

```
% Nel workspace di matlab dovrà essere presente:
% 1. la matrice "data" (n x (p+1)) - campione di
% apprendimento - contenente n transazioni, p TF-IDF
% per ogni transazione ed 1 variabile risposta binaria
% Y nell'ultima colonna;
% 2. il vettore "textdata" contenente la denominazione
% dei p itemes e la Y nell'ultima cella;
% 3. la cardinalità massima delle regole "lung".
% Generazione regole contenenti y come antecedente e
% relativi supporti pesati
[r,c] = size(data)
itemsets = textdata(:,1:c-1);
vetWSP = [];
if or((lung < 2),(lung > c))
   itemsets = {};
   return
end
itemsets = combnk(itemsets,lung-1);
vetY(1:size(itemsets,1),1) = textdata(:,end);
itemsets = [vetY itemsets];
vetWSP = zeros(size(itemsets,1),1);
vetWt = sum(data, 2)./sum(data > 0, 2);
wT = sum(vetWt);
matrBin = (data>0);
for i = 1 : size(itemsets,1)
   vetBin = ismember(textdata,itemsets(i,:));
   numWSP = 0;
   for j = 1 : r
      if isequal(and(matrBin(j,:),vetBin),vetBin)
         numWSP = numWSP + vetWt(j);
      end
   end
vetWSP(i,1) = (numWSP/wT)*100;
i
end
% Ordinamento decrescente del vettore dei supporti
% pesati peresi una sola volta
vetWspOrd = flipud(unique(vetWsp));
```

```
vetTcc = [];
vetTcc1 = [];
vetTcc0 = [];
matrBin = (data(:,1:end-1)>0);
vetY = (data(:,end));
for i = 1 : 100
% Ricavo delle regole significative ed dei
% relativi supporti per ogni supporto pesato
   freqItemsets = {};
   vetFreqWsp = [];
   pos = find(vetWsp >= vetWspOrd(i,1));
   rpos = size(pos,1);
   for j = 1 : rpos
      freqItemsets(j,:) = itemsets... (pos(j,1),2:end);
      vetFreqWsp(j,1) = vetWsp(pos(j,1),1);
   end
% Classificazione delle regole significative
% impostando ogni supporto pesato come minimo
   l = size(freqItemsets,1);
   r = size(data,1);
   vetClassY = zeros(r,1);
   for h = 1 : l
      vetBin = ismember(textdata(:,1:end-1)...
               ,freqItemsets(h,:));
      for j = 1 : r
         if isequal(and(matrBin(j,:),vetBin)... ,vetBin)
            vetClassY(j,1) = 1;
         end
      end
   end
% Calcolo del Tasso di Corretta Classificazione
% per ogni supporto pesato
   tcc = 0;
   tcc1 = 0;
   tcc0 = 0;
   matrConf = zeros(2,2);
   y1 = sum(vetY == 1);
   y0 = r - y1;
   for j = 1 : r
      if and((vetY(j,1)==0),(vetClassY(j,1) == 0))
         matrConf(1,1) = matrConf(1,1) + 1;
      elseif and((vetY(j,1)==1),... (vetClassY(j,1)==1))
         matrConf(2,2) = matrConf(2,2) + 1;
      end
   end
   vetTcc(i,1) = (matrConf(1,1) + matrConf(2,2))/r;
   vetTcc1(i,1) = matrConf(2,2)/y1;
   vetTcc0(i,1) = matrConf(1,1)/y0;
end
```

```
% Ricavo del massimo Tasso di Corretta Classificazione
% degli 1 e della sua posizione
posMaxTcc1 = find(vetTcc1 >= 0.8);
troncTcc1 = vetTcc1(posMaxTcc1,1);
tcc1 = min(troncTcc1);
% Estrazione del TCC e TCC0 in corrispondenza del TCC1
tcc = max(vetTcc(posMaxTcc1,1));
tcc0 = max(vetTcc0(posMaxTcc1,1));
% Estrazione del Minimo Supporto Pesato
minWsp = max(vetWspOrd(posMaxTcc1,1))
% Estrazione delle regole significative per la
% classificazione
regole = {};
vetWspReg = [];
pos = find(vetWsp >= minWsp);
rpos = size(pos,1);
for j = 1 : rpos
   regole(j,:) = itemsets(pos(j,1),:);
   vetWspReg(j,1) = vetWsp(pos(j,1),1);
end
```

Bibliografia

Agrawal R., Imielinski T. e Swami A. Mining association rules between sets of items in large dataset. - New York : ACM Press, 1993. - Vol. in: Proceedings of the 1993 ACM SIGMOD International : p. 207–216.

Balbi S. e Di Meglio E. Una strategia di Text Mining basata su regole di asociazione - Roma : Università la Sapienza, 2004. - Vol. Applicazioni di analisi statistica dei dati testuali.

Balbi S. Non symmetrical correspondence analysis of textual data and confidence regions for graphical forms In JADT. - 1995. - Vol. 2. - p. 5-12.

Balbi S., Bolasco S. e Verde R. Text mining on elementary forms in complex lexical Structures in A. Morin, P. Sèbillot: Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles. - Paris : Irisa-Inria, 2002. - p. 89-100.

Balbi S., Bruzzese D. e Grassia M.G. Chi Cerca Cosa. Esplorare le competenze richieste dalle imprese mediante tecniche di mining in Atti di convegno: Efficacia esterna della formazione universitaria: il progetto OUTCOMES; a cura di L. Fabbris, 2005. - p. 189-201.

Balbi S., Bruzzese D. e Scepi G. Analyze the E-commerce impact on business to consumer transactions through Non Simmetrical Association rules. Statistica Applicata, 2003. - p. 131-148.

Balbi S., Infante G. e Misuraca M. Conjoint analysis with textual external information. In stampa. - 2008.

Benzécri J.P. Histoire et prehistoire de l'analyse des données. - Paris : Dunod, 1982.

Benzécri J.P. L'analyse des données - Paris : Dunod, 1973.

Bock H. e Diday E. Symbolic data analysis. Springer Verlag, 2000.

Bolasco S. Criteri di lemmatizzazione per l'individuazione di coordinate semantiche in Ricerca qualitativa e computer: teoria, metodi ed applicazioni. Franco Angeli, 1995.

Bolasco S. Statistica testuale e text mining: alcuni paradigmi applicativi. Quaderni di Statistica, Liguori, 2005. - Vol. 7 : p. 17-53.

Bolasco S., Verde R. e Balbi S. Outils de Text Mining pour l'analyse de structures lexicales à elements variables in A. Morin, P. Sèbillot: (eds.), Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles. - Paris : Irisa-Inria, 2002. - p. 197-208.

Bruzzese D. e Davino C. Post Analysis of Association Rules in a symbolic framework in Atti della XL Riunione scientifica della Società italiana di Statistica, Sessioni Plenarie e Specializzate. - Firenze : [s.n.], 2002. - p. 91-102.

Di Meglio E. Text Mining: a statistical perspective. Tesi di dottorato in Statistica computazionale, Dipartimento di Matematica e Statistica, Università di Napoli Federico II, 2003.

Diday E. Introduction à l'approche symbolique en analyse de données. Premier journées Symbolique-Numerique. - CERAMADE, Université Paris IX Dauphine, 1987. - p. 21-56.

Dulli S., Polpettini P. e Trotta M. Text mining: teoria e applicazioni. Franco Angeli, 2004.

Fisher R.A. The precision of discriminant functions. Ann. Eugene, 1940.

Gettler Summa M. Marking and Generalization by Symbolic Objects in the Symbolic Official Data Analysis Software. Paris : Université Dauphine LISE CEREMADE, 1998.

Giordano G. e Scepi G. Different informative structure for quality design. Journal of italian statistical society, 1999. - Vol. 8 (2-3) : p. 139-149.

Giordano G. L'analisi multidimensionale dei dati di preferenza: una strategia esplorativa per la Conjoint Analysis. Tesi di dottorato in statistica computazionale ed Applicazioni, Dipartimento di Matematica e Statistica, Università di Napoli Federico II. , 1998.

Giuliano L. L'analisi automatica dei dati testuali. Software e istruzioni per l'uso. Edizioni Universitarie di Lettere Economia Diritto, 2004.

Grassia M.G., Misuraca M. e Scepi G. Relazioni non simmetriche tra corpora. Le poid des mots. Actes des 7es Journ´ees internationales d'Analyse statistique des Données Textuelles / a cura di in G. Purnelle C. Fairon, A. Dister (eds.), - UCL Presses Universitaires de Louvain, 2004. - p. 524-532.

Green P.E. e Srnivason V. Conjoint analysis in marketing: new devlopments with implications for research and practise - Journal of Marketing, 1990. - Vol. 54 : p. 3-19.

Greenacre M.J. Theory and application of correspondence analysis - London : Academic Press, 1984.

Guiraud P. Les caractères statistiques du vocabulaire- Paris : Puf, 1954.

Holsen R.A. [et al.] Classification and regression trees - New York : Wadsworth Statistical Press, 1984.

Hotelling H. Analysis of a complex statistical variables into principal components. Journal of educational Psychology. - 1933. - Vol. 24.

Infante G. e Giordano G. Una rassegna dei contributi sulla Conjoint Analysis mediante l'analisi dei dati testuali. Giornata di Studio "La Conjoint Analysis. Orientamenti metodologici e recenti contributi all'analisi dei dati di preferenza nelle scienze socio-economiche". - 2004.

Infante G. e Misuraca M. Text mining strategies for analysing semi-structured corpora. Meeting of the Classification and Data Analysis of the italian Statistical Society. - 2007. - p. 267-270.

Lambin J.J. Marketing strategico. una prospettiva europea. McGraw-Hill, 1996.

Lancaster K. A new approach to consumer theory. Journal of political economics, 1966.

Larose D.T. Discovering knoledge in data. Wiley-intersciense, 2005. - p. 196-197.

Lauro N.C., Giordano G. e Verde R. A multidimensional approach to conjoint analysis. Applied Stochastic Models and Data Analysis, 1998. - p. 265-274.

Lebart A. e Salem S. Analyse statistique des données textuelles. Paris : Dunod, 1988.

Lebart L. e Salem A. Statistique textuelle. - Paris : Dunod, 1994.

Lebart L., Morineau A. e Warwick K.M. Multivariate Descriptive Statistical Analysis. Wiley & Sons, 1984.

Misuraca M. La visualizzazione dell'informazione testuale. Contributi metodologici ed applicativi. Tesi di dottorato in Statistica, Dipartimento di Matematica e Statistica, Università di Napoli Federico II, 2004.

Mola F. e Siciliano R. A Fast Splitting Procedure for Classification and Regression Trees. Statistics and Computing. - Chapman Hall, 1997. - Vol. 7. - p. 208-216.

Mola F. e Siciliano R. A two-stage predictive splitting algorithm in binary segmentation. Computational Statistics: COMPSTAT 92. Physica Verlag, 1992. - p. 179-184.

Morrone A. Alcuni criteri di valutazione della significatività dei segmenti ripetuti. S.J. Anastex, 1993. - p. 445-453.

Muller C. Principes et methode de la statistique lexical. Paris : Hachette, 1977.

Nishisato S. Analysis of categorical data: dual scaling and its applications. Toronto : University of Toronto Press, 1980.

Oren N. Reexamining tf.idf based information retrieval with Genetic Programming. Proceedings of the 2000 Annual Research Conference of the South African Institute of Computer scientists and Information Technologists on Enblement through technology / aut. libro P. Kotzé. - 224-234 : SAICSIT, 2002.

Pearson K. On lines and planes of closet fit to systems of points in space. Phil Mag., 1901. - Vol. 11. - p. 559-572.

Quinlan J.R. C4.5: programs for machine learning, Morgan Kaufmann Publishers, 1993.

Reinert M. Un logiciel d'analyse des donnès textuelles: Alceste. Data analysis and informatics / aut. libro E. in Diday. NH, 1988. - Vol. 5.

Salton G. e Buckley C. Term weighting approaches in automatic text retrieval. Information Processing & management, 1988. - Vol. 24. - p. 513-523.

Sebastiani F. Machine Learning in Automated Text Categorisation. ACM Computing surveys, 2002. - p. 2.

Spearman C. General intelligence, objectively determined and measured. Journal of Psychology. Amer, 1904. - Vol. 15. - p. 201-293.

Takane Y. e Shibayama T. Principal component nalysis with esternal information on both subjects and variables. Psychometrika, 1991. - Vol. 56 : p. 97-120.

Tao F., Murtagh F. e Farid M. Weighted association rule mining using weighted support and significance framework. New York : ACM Press, 2003. - p. 661-666.