Quaderni di Statistica
Vol. 7, 2005

Pesi e metriche nell'analisi dei dati testuali

Simona Balbi, Michelangelo Misuraca
Dipartimento di Matematica e Statistica, Università di Napoli Federico II
E-mail: simona.balbi@unina.it; michelangelo.misuraca@unina.it

Summary: The paper goes through some tools considered nowadays "classical" in Text Mining procedures and software. We are speaking of Latent Semantic Indexing for dimensionality reduction, and the wide literature devoted to the problem of how to weight the word importance, and how to measure similarities between words and between words and queries. Visualisation is strongly affected by these choices. Here we compare some alternatives from a statistical viewpoint. A *corpus* consisting of six years of the Italian edition of *Le Monde Diplomatique* is analysed in order to show the effects of the different weighting systems together with the potentiality of Textual Data Analysis in summarising and representing newspaper information.

Keywords: Latent semantic indexing, Correspondence Analysis, tf/idf Index.

1. Introduzione

La crescente mole di fonti documentarie disponibili in formato elettronico ha reso possibile, e al contempo necessario, il ricorso a strategie sempre più complesse per l'estrazione, l'analisi e l'organizzazione della conoscenza, con lo scopo di soddisfare i diversi bisogni informativi. La ricerca della conoscenza in database di notevoli dimensioni (KDD, *Knowledge Discovery in Databases*) è messa in atto con un insieme di strumenti nati in ambito informatico, ma affrontata sempre più anche in ambito statistico, per la natura interdisciplinare del problema.

Negli ultimi anni la forbice tra l'analisi di dati strutturati e non strutturati (cioè non in formato numerico) si è ampliata a tal punto che *Data Mining* e *Text Mining* (TM) sono ormai considerati ambiti di ricerca

nettamente distinguibili, poiché soddisfano in modo differente bisogni informativi di natura diversa. Il TM, in particolare, ha come obiettivo l'estrazione di conoscenza a partire da grandi raccolte di fonti testuali (d'ora in avanti *documenti*).

In precedenti lavori è stato sottolineato come le tecniche proprie dell'Analisi Multidimensionale dei Dati rappresentino un utile strumento per l'estrazione di conoscenza da ampie collezioni di documenti, enfatizzando le peculiarità del punto di vista statistico (Balbi e Di Meglio, 2004). Analogamente, è stato mostrato come alcuni strumenti sviluppati nell'*Information Retrieval* (IR) possano fornire nuovi interessanti spunti di ricerca per gli analisti di dati testuali (Balbi e Misuraca, 2005).

Obiettivo del presente lavoro è approfondire le implicazioni connesse all'utilizzo di alcuni strumenti, considerati ormai "classici" nelle procedure (e nei software) di TM, ponendo in risalto le criticità delle scelte in termini di codifiche, distanze e sistemi di pesi non familiari ad utilizzatori con una formazione non prettamente statistica. In altre parole, ponendo una chiara distinzione tra le problematiche proprie dell'Analisi dei Dati Testuali (ADT) e di quelle proprie del TM, l'obiettivo è quello di approfondire i contributi che la prima può fornire al secondo, in particolare in tema di consapevolezza delle scelte effettuate e della validità dei risultati ottenuti.

L'attenzione sarà in particolare concentrata sugli strumenti utilizzati per ridurre la dimensionalità del fenomeno osservato e sui differenti sistemi di pesi per tener conto dell'importanza da attribuire alle singole parole contenute all'interno della base documentaria oggetto di analisi.

Il riferimento principale, nell'ambito del TM, sarà il ricorso al cosiddetto *Latent Semantic Indexing* e a misure di ponderazione quali il *tf/idf*. Un'ulteriore questione affrontata, strettamente connessa alle precedenti, riguarderà le implicazioni sulle rappresentazioni grafiche delle diverse misure di similarità fra parole (e fra parole e *query*).

La rilevanza delle problematiche affrontate, anche in relazione alla fase di pre-trattamento del testo e all'analisi dei risultati ottenuti, sarà illustrata sulla base di un *corpus* di documenti costituito da sei annate dell'edizione italiana della rivista *Le Monde Diplomatique*.

2. La struttura dei dati e la loro codifica: il Bag-of-Words

In un processo di analisi dei dati il primo passo è necessariamente quello di definire la struttura dei dati rispetto al fenomeno oggetto di studio. Il TM può essere definito come "il processo di estrazione di informazione e conoscenza, interessante e non banale, da un testo non strutturato" (Hearst, 1999).

È evidente, quindi, che un approccio quantitativo all'analisi del testo necessita una trasformazione di quest'ultimo, una sua "strutturazione", attraverso un sistema di codici idonei. Questa fase preliminare, detta di *numerizzazione*, è articolata in una serie di sottofasi legate alle leggi del linguaggio naturale (quali l'identificazione dell'unità minima di senso, la lemmatizzazione, la disambiguazione, il *tagging* grammaticale, ecc.), e rappresenta una prima immissione di valutazioni soggettive in un processo automatico. Nella letteratura della statistica testuale esiste, ad esempio, un vivace dibattito sulla scelta dell'unità d'analisi (Bolasco, 1999), completamente ignorata (anche a causa dell'enorme peso computazionale) nei software di TM, che operano "naturalmente" sulle cosiddette *forme grafiche* (termini, parole).

La successiva fase di codifica pone all'analista delle scelte circa il peso da attribuire alle singole parole all'interno delle parti del testo. Per poter trattare in modo automatico un testo si ricorre generalmente al cosiddetto *Bag-of-Words*, nel quale si opera una trasformazione dei documenti in vettori in uno spazio multidimensionale.

Un generico vettore/documento $\mathbf{d}_j$ è rappresentato come

$$\mathbf{d}_j = [w_{1j}, w_{2j}, \ldots, w_{ij}, \ldots, w_{pj}] \qquad (1)$$

dove w_{ij} è espressione dell'importanza della i-esima parola nel j-esimo documento in termini di contenuto informativo ($i = 1,\ldots,p$ e $j = 1,\ldots,q$), nel caso in cui la base documentaria sia costituita da p parole differenti presenti in q documenti.

Si tratta, quindi, di introdurre un adeguato sistema di pesi che tenga conto della diversa importanza delle parole. Gli schemi di ponderazione più diffusi sono:

[a] il *booleano*, in cui w_{ij} assume valore 1 se la parola i è presente nel documento j e 0 altrimenti;
[b] il *frequentista*, in cui w_{ij} è uguale a n_{ij}, frequenza della parola i nel documento j;
[c] il *frequentista normalizzato*, in cui $w_{ij} = n_{ij} / max\ n_j$, con $max\ n_j$ frequenza della parola più presente all'interno del documento j;
[d] *tf/idf* (*term frequency* / *inverse document frequency*), proposto per problemi di IR da Salton e Buckley (1988), in cui:

$$w_{ij} = \frac{n_{ij}}{max\ n_j} \cdot \log \frac{q}{q_i} \tag{2}$$

dove q è il numero totale dei documenti e q_i è il numero di documenti in cui è presente la parola i. La *ratio* per l'introduzione del secondo rapporto risiede nel potere discriminante di una parola all'interno della base documentaria.

Quest'ultimo sistema di pesi, semplice ed efficace, ha avuto un grande successo e ne sono state proposte numerose varianti. Bisogna in ogni caso sottolineare come il *tf/idf* sia un indice basato su valutazioni pragmatiche più che statistiche.

3. Riduzione e rappresentazione dell'informazione testuale

La *decomposizione in valori singolari* (DVS) è una operazione algebrica utilizzata sia nell'ADT sia in alcune tecniche di IR per decomporre la matrice di dati originaria e ricostruirla come matrice di rango ridotto (Eckart e Young, 1936). Questo denominatore comune consente di far riferimento ai processi di IR e a quello di *Information Mining* (IM) o di ADT come ad un contesto analogo in cui inscrivere e quindi discutere le scelte effettuate in termini di distanze e pesi.

Data una matrice rettangolare $\mathbf{A}$ (p,q) con $p > q$, per semplicità di rango q, si ha che:

$$\begin{aligned} \mathbf{A} &= \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}} \\ \mathbf{U}^{\mathrm{T}}\mathbf{U} &= \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I} \end{aligned} \tag{3}$$

dove $\mathbf{\Lambda}$ (q,q) è una matrice diagonale di numeri positivi λ_α (con α = 1,2,…,q) posti in ordine decrescente detti *valori singolari*, mentre **U** (p,q) e **V** (q,q) sono le matrici dei *vettori singolari* di sinistra e destra.

In Balbi e Di Meglio (2004) e in Misuraca (2005) si è ampiamente sottolineato come una rilettura della decomposizione in termini di DVS *generalizzata* (DVSG, Greenacre, 1984) consenta di riportare il problema della scelta del sistema di ponderazione per parole e documenti ad un problema di definizione di metriche Euclidee ponderate in spazi multidimensionali.

Nella DVSG la ricerca di un sottospazio che meglio approssimi la struttura dei dati è espressa in modo equivalente da:

$$\begin{matrix} \mathbf{\Omega}^{1/2}\mathbf{A}\,\mathbf{\Phi}^{1/2} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}} \\ \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I} \end{matrix} \quad \Leftrightarrow \quad \begin{matrix} \mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}} \\ \mathbf{U}^{\mathrm{T}}\mathbf{\Omega}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{\Phi}\mathbf{V} = \mathbf{I} \end{matrix} \tag{4}$$

dove $\mathbf{\Omega}$ (p,p) e $\mathbf{\Phi}$ (q,q) sono due matrici simmetriche definite positive rappresentanti i sistemi di pesi e le metriche Euclidee ponderate scelte nei due spazi di rappresentazione degli elementi posti sulle righe e sulle colonne della matrice **A**.

4. Il Latent Semantic Indexing

Lo scopo principale degli strumenti di IR è consentire all'utente di "navigare" in basi documentarie di notevoli dimensioni per estrarre informazione "utile" su uno specifico argomento. Affinché tale ricerca sia proficua e risponda alle esigenze dell'utente è necessario definire le parole chiave che caratterizzano i diversi documenti ed indicizzarle con un sistema di pesi che assicuri una buona rispondenza tra la richiesta ed il risultato ottenuto.

Se nell'ambito dell'ADT il problema della ricerca di un sottospazio è fortemente sentito, con lo scopo di identificare una struttura di associazione tra le variabili osservate e visualizzarla graficamente, esso è centrale in egual misura nelle strategie di IR.

Una delle soluzioni comunemente adottate, particolarmente vantaggiosa in termini di identificazione dei documenti di interesse, è quella basata sul *Latent Semantic Indexing* (LSI, Deerwester et al., 1990). Tale famiglia di metodi supera i limiti dell'effettiva presenza, nei documenti, di tutte le parole chiave utilizzate nella ricerca (*query*) e, sotto determinate condizioni, anche della lingua in cui quest'ultima è formulata (Littman et al., 1998).

L'obiettivo finale è quello di soddisfare una bisogno informativo specifico. Lo schema d'analisi utilizzato prevede il ricorso alla DVS, così come espressa nella (3), per la ricerca di *fattori* ortogonali che approssimino la matrice dei dati originaria {*parole* × *documenti*}, e nella successiva proiezione su un piano fattoriale dei documenti e delle *query*, in modo da individuare i documenti d'interesse in termini di prossimità.

Dallo schema proposto si evince l'utilizzo di un sistema di pesi unitario, che privilegia quindi la frequenza delle parole come misura della loro importanza, e di una metrica Euclidea per misurare la distanza tra documenti e tra documenti e *query*, privilegiando implicitamente i documenti più lunghi.

Sempre più frequentemente nei software commerciali per il TM, la matrice dei dati è sottoposta ad una trasformazione preliminare, in modo da considerare non la frequenza assoluta di ogni parola nei documenti ma il *tf/idf*, calcolato generalmente secondo la (2). Il vantaggio derivante dall'introduzione di questo doppio sistema di pesi, *locale* (rappresentato dal *tf*) e *globale* (rappresentato dall'*idf*), risiede nel fatto che si ottengono migliori risultati qualora nei documenti non siano riportate esattamente le parole chiave utilizzate nelle *query*.

5. L'Analisi delle Corrispondenze Lessicali

La matrice di dati tipica dell'ADT ha nelle *p* righe le parole presenti nel *corpus* oggetto d'analisi e nelle *q* colonne i documenti (Lebart e Salem, 1994). Si tratta della giustapposizione di *q* vettori/documento così come definiti nella (1), considerando un sistema di ponderazione [b].

Su questa matrice sono state proposte numerose analisi, la più diffusa delle quali è la cosiddetta *Analisi delle Corrispondenze Lessicali* (ACL),

che ha come obiettivo la descrizione da un punto di vista geometrico e algebrico delle relazioni tra le parole, tra i documenti e, indirettamente, tra parole e documenti.

L'analisi è svolta calcolando una serie di fattori a partire dalle variabili originarie, ognuno dei quali rappresenta una dimenzione latente del tipo di associazione presente nei dati, espressa in termini di χ^2. La successiva rappresentazione in forma grafica consente una interpretazione semplice della struttura, evidenziando gli aspetti non rilevabili dalla lettura diretta dei dati.

Come è noto (Lebart et al., 1995), in questa tecnica le distanze fra documenti e le distanze fra parole sono calcolate facendo ricorso ad una metrica Euclidea ponderata, nota come metrica del χ^2:

$$d(i,i') = \sum_{j=1}^{q} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 \qquad d(j,j') = \sum_{i=1}^{p} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2 \tag{5}$$

con $f_{ij} = n_{ij} / n_{..}$, $f_{i.} = n_{i.} / n_{..}$ e $f_{.j} = n_{.j} / n_{..}$. In questo modo parole poco frequenti e documenti brevi contribuiscono all'analisi in egual misura, rispetto a parole frequenti e documenti lunghi. In alcuni specifici contesti (Balbi, 1995) si è suggerito di dare un peso alle parole che tenesse conto della loro frequenza, normalizzando il contributo dei documenti:

$$d(i,i') = \sum_{j=1}^{q} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 \qquad d(j,j') = \sum_{i=1}^{p} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2 \tag{6}$$

La ricerca dei fattori è eseguita attraverso una DVSG della matrice delle frequenze relative **F**, secondo lo schema riportato nella (4):

$$\begin{matrix} \mathbf{D}_p^{-1/2} \mathbf{F} \, \mathbf{D}_q^{-1/2} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{V}^{\mathrm{T}} \\ \mathbf{U}^{\mathrm{T}} \mathbf{U} = \mathbf{V}^{\mathrm{T}} \mathbf{V} = \mathbf{I} \end{matrix} \quad \Leftrightarrow \quad \begin{matrix} \mathbf{F} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{V}^{\mathrm{T}} \\ \mathbf{U}^{\mathrm{T}} \mathbf{D}_p^{-1} \mathbf{U} = \mathbf{V}^{\mathrm{T}} \mathbf{D}_q^{-1} \mathbf{V} = \mathbf{I} \end{matrix} \tag{7}$$

Le matrici diagonali $\mathbf{D}_p^{-1}$ e $\mathbf{D}_q^{-1}$ sono costruite a partire dalle distribuzioni marginali di riga e colonna $f_{i.}$ e $f_{.j}$ (rispettivamente frequenza relati-

va dell'*i*-esima parola e lunghezza del *j*-esimo documento), e rappresentano il sistema di pesi e la metrica nello spazio in cui vengono proiettate le parole; viceversa, le stesse, rappresentano la metrica ed il sistema di pesi nello spazio in cui vengono proiettati i documenti.

6. Analisi fattoriali basate sul tf/idf

Quando l'analisi è focalizzata sui documenti, o più specificatamente sulle categorie di documenti, è possibile utilizzare una peculiare metrica Euclidea ponderata in termini di *tf* per misurare le similarità, assumendo un sistema di pesi unitario (Balbi e Misuraca, 2005).

La base teorica di tale tecnica risiede nell'idea che le differenti categorie hanno la medesima importanza, ma è necessario considerare il diverso peso delle parole nel valutare la distanza tra categorie differenti.

Data la matrice **A** (*p*,*q*) ed una matrice in codifica disgiuntiva completa **Q** (*q*,*h*) che considera per tutti i documenti di un *corpus* una certa caratteristica, si costruisce la matrice **K** = **AQ**, che ha sulle colonne le *h* categorie di documenti. Da **K** (*p*,*h*) si ricava la matrice **F**, il cui elemento generico f_{ic} rappresenta il numero di volte in cui la parola *i* si presenta nella *c*-esima categoria, con *c* = (1,…,*h*), rispetto al numero totale di parole prese in considerazione.

L'indice *tf* per la parola *i* è calcolato utilizzando il sistema di pesi [c], dove $max\, f_c$ è la frequenza della parola più presente nella categoria *c*.

Detto q_c il numero di documenti presente in ogni categoria, è possibile ottenere per ogni forma il *tf* medio come media ponderata dei *tf* relativi alla *i*-esima parola:

$$atf_i = \frac{1}{h}\sum_c \left(\frac{f_{ic}}{max\, f_c} \right) q_c \qquad (8)$$

Dalla (8) si costruisce la matrice diagonale dei *tf* medi $\mathbf{D}_{atf}$, utilizzata come sistema di pesi nello spazio in cui sono proiettate le parole e conseguentemente come metrica nello spazio dei documenti.

La rappresentazione che si ottiene consente di valutare su un grafico piano la *ricchezza lessicale* delle categorie di documenti.

Da un punto di vista matematico la ricerca degli assi principali con cui costruire la mappa fattoriale è effettuata con la DVSG della matrice:

$$\begin{aligned} \mathbf{F} &= \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}} \\ \mathbf{U}^{\mathrm{T}}\mathbf{D}_{atf}^{-1}\mathbf{U} &= \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I} \end{aligned} \tag{9}$$

con $\mathbf{D}_{atf}^{-1/2}\mathbf{U}$ e $\mathbf{V}$ matrici dei vettori singolari sinistri e destri.

Ricorrendo all'indice *tf/idf* è possibile anche costruire, come già visto, una matrice {*parole* × *documenti*} che ha come elemento generico il *tf* della *i*-esima parola nella *c*-esima categoria di documenti, ponderato per l'indice *idf* calcolato sulla intera collezione di documenti.

Da un punto di vista algebrico la matrice *tf/idf* può essere ottenuta a partire dalla matrice **F** con il prodotto $\mathbf{D}_{idf}^{-1}\mathbf{F}\mathbf{D}_{tf}^{-1}$, dove la matrice diagonale $\mathbf{D}_{tf}^{-1}$ ha come elemento generico $1/max\ f_c$, mentre la matrice $\mathbf{D}_{idf}^{-1}$ ha come elemento generico $\log(h/h_i)$, con h numero di categorie presenti nel *corpus* e h_i il numero di categorie in cui è presente la parola *i*.

Tale schema può essere ricondotto all'utilizzo nell'analisi dei dati di informazione esterna, facendo riferimento in questo caso ad una informazione *intra-testuale*, di tipo quantitativo, che tiene conto delle diverse relazioni tra le forme e i documenti ed è ottenuta a partire dai dati contenuti nel corpus (Misuraca, 2005).

La matrice **F** è decomposta secondo l'espressione:

$$\begin{aligned} \mathbf{F} &= \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}} \\ \mathbf{U}^{\mathrm{T}}\mathbf{D}_{idf}^{-1}\mathbf{U} &= \mathbf{V}^{\mathrm{T}}\mathbf{D}_{tf}^{-1}\mathbf{V} = \mathbf{I} \end{aligned} \tag{10}$$

dove $\mathbf{D}_{idf}^{-1/2}\mathbf{U}$ e $\mathbf{D}_{tf}^{-1/2}\mathbf{V}$ rappresentano le matrici dei vettori singolari sinistri e destri.

Se utilizziamo la formulazione alternativa della DVSG espressa nella (4), è evidente come una LSI sulla matrice *tf/idf* in realtà dia un peso doppio tanto alle parole quanto ai documenti rispetto alla (10):

$$\begin{matrix} \mathbf{D}_{idf}^{-1}\mathbf{F}\,\mathbf{D}_{tf}^{-1} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}} \\ \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I} \end{matrix} \quad \Leftrightarrow \quad \begin{matrix} \mathbf{F} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}} \\ \mathbf{U}^{\mathrm{T}}\mathbf{D}_{idf}^{-2}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{D}_{tf}^{-2}\mathbf{V} = \mathbf{I} \end{matrix} \tag{11}$$

A questo punto diventano determinanti gli obiettivi dell'analista. Se infatti lo scopo dell'analisi è quello di ricercare una conoscenza specifica, tipicamente nelle strategie di IR, allora la LSI consente di enfatizzare le parole più presenti e i documenti più lunghi. Se lo scopo dell'analsi è quello di una conoscenza più ampia, in un'ottica esplorativa propria dell'ADT, allora l'analisi fattoriale con pesi e metrica *tf/idf* fornisce dei risultati più interessanti.

7. Un caso studio: l'edizione italiana de Le Monde Diplomatique

Il linguaggio utilizzato dalla stampa è stato nel corso degli ultimi anni, per la sempre maggiore disponibilità di fonti e di strumenti avanzati, più volte investigato. In Italia l'esempio più significativo è certamente dato dal *corpus* Rep-90, costituito da dieci annate del quotidiano *La Repubblica* (Bolasco e Canzonetti, 2003).

Nell'applicazione presentata viene utilizzato un *corpus* di dimensione minore, ma interessante per i contenuti linguistici espressi. La base di dati analizzata comprende sei annate (1998-2003) dell'edizione italiana del mensile francese *Le Monde Diplomatique* (nel seguito LMD).

Si tratta della traduzione dell'edizione originale, con l'aggiunta di alcune recensioni letterarie e cinematografiche prodotte dalla redazione italiana. Il linguaggio utilizzato è sufficientemente omogeneo poiché le traduzioni sono affidate ad un numero limitato di traduttori, anche se comunque risente di alcune scelte soggettive che, ad esempio, portano a rendere alcune parole "uniche" in lingua originale in modo diverso.

Le sei annate complete sono state scaricate con un software dedicato dal sito web `http://www.ilmanifesto.it/MondeDiplo/` e convertite in un unico file in formato testo utilizzando un programma in *Java*.

Per ogni articolo è stata considerato solo il corpo, escludendo titolo e occhiello. Sul sito gli articoli sono parzialmente categorizzati, per facilitare la ricerca, in base ad un indice tematico e ad un indice geografico; poiché il sistema di categorie utilizzato è risultato incompleto, con

l'ausilio di conoscenza esperta sono state create 32 categorie sulla base dell'argomento principale degli articoli. Dai circa 2000 articoli iniziali, ne sono stati selezionati manualmente 1914.

I codici delle pagine web in *Html* sono stati eliminati dal *corpus* per mezzo delle cosiddette *regular expression*. È stata effettuata una normalizzazione degli articoli per ridurre i casi di possibile duplicazione dei dati, ad esempio conformando la traslitterazione delle parole provenienti da alfabeti diversi.

Dopo aver messo in atto una lessicalizzazione approfondita per individuare le unità minimali di senso ed eliminare i casi più comuni di ambiguità, è stato costruito un vocabolario di più di 100.000 *forme testuali*, marcate con un sistema di etichette del tipo *Part of Speech*.

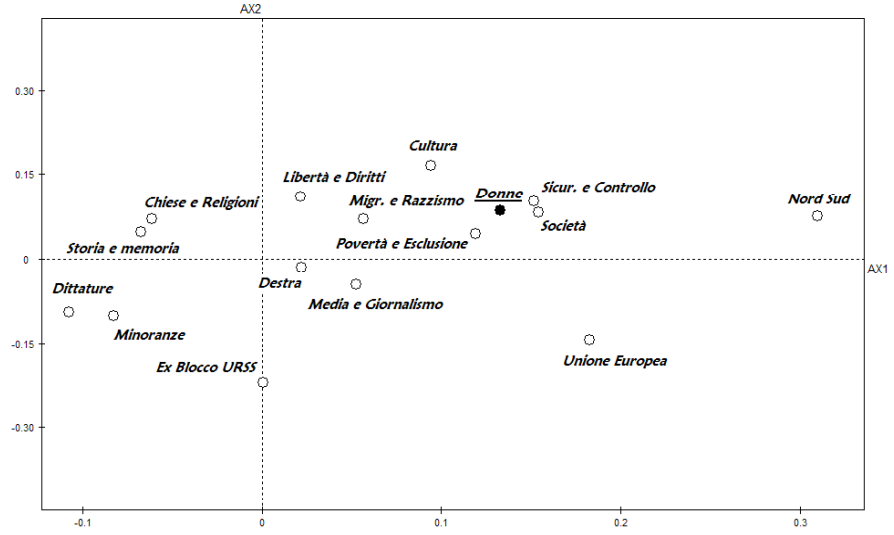

Figura 1. Analisi delle Corrispondenze Lessicali: I e II asse

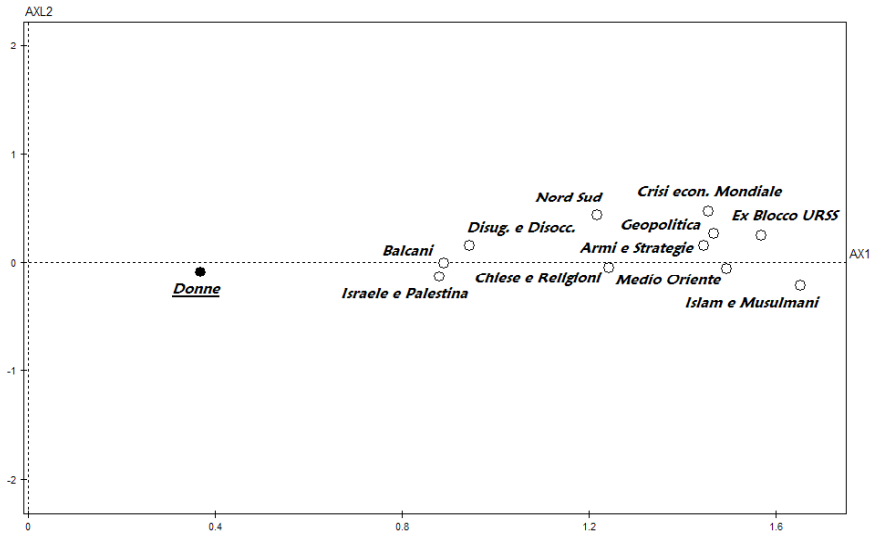

Figura 2. Analisi fattoriale con pesi e metriche tf/idf: I e II asse

Nel confronto tra le rappresentazioni grafiche della ACL (Figura 1) e dell'analisi fattoriale con pesi e metriche basati sul *tf/idf* (Figura 2) si nota, ad esempio, come la categoria "Donne" cambi la propria posizione sul piano e soprattutto il suo sistema di relazioni con le altre categorie.

Da una analisi approfondita degli articoli di tale categoria emerge la presenza di parole fortemente caratterizzanti, legate al tema della *mutilazione sessuale* tipica di alcune culture del mondo arabo e musulmano.

Nella Figura 1 le coordinate fattoriali sul primo e secondo asse non sono molto elevate, come conseguenza della metrica χ^2, mentre il ricorso al *tf/idf* nella Figura 2 evidenzia la forte caratterizzazione linguistica della categoria.

8. Conclusioni e prospettive

In questo lavoro si è mostrato come cruciale a qualsiasi trattamento quantitativo di basi documentarie sia la scelta di un sistema di pesi che esprima l'*importanza* delle parole, in relazione al loro potere discrimi-

nante ed alla loro portata informativa, nonché la scelta di un sistema di pesi da attribuire ai documenti in cui si articola la base documentaria ed in cui appaiono le differenti parole.

La conclusione cui si è pervenuti è che non è possibile ricorrere a sistemi di pesi "ottimi" in tutte le circostanze, ma che questi devono essere strettamente connessi agli specifici obiettivi d'analisi.

L'aver ricondotto le differenti strategie comunemente utilizzate ad una unica operazione algebrica (la DVSG) consente, a nostro avviso, di lasciare aperta la questione, potendosi preconizzare un utilizzo a maggior vocazione informativa dei pesi, non necessariamente legato alle nozioni di "frequenza" per le parole e "lunghezza" per i documenti, ma che altresì consenta d'immettere nell'analisi ulteriori informazioni "esterne", sui documenti e/o sulle parole.

Si pensi per queste ultime, ad esempio, alla possibilità di differenziare i ruoli a seconda della natura grammaticale, ovvero tenendo conto di informazioni relative alla maggiore o minore attinenza di un termine all'oggetto di studio. O, ancora, per i documenti, ad assegnare una diversa importanza in relazione al soggetto che lo ha prodotto.

In tale prospettiva si riafferma la necessità di un approccio interdisciplinare al testo, lasciando allo statistico il compito di alimentare il quadro metodologico di riferimento, inserendo i contributi, sostanziali o formali, dei ricercatori degli altri ambiti disciplinari coinvolti.

Riferimenti bibliografici

Balbi S. (1995), Non symmetrical correspondence analysis of textual data and confidence regions for graphical forms, in Bolasco S. et al. (eds.), *Actes des 3es Journées internationales d'Analyse statistique des Données Textuelles*, CISU, Roma, II, 5-12.

Balbi S., Di Meglio E. (2004), Contributions of Textual Data Analysis to Text Retrieval, in Banks D. et al. (eds.), *Classification, Clustering and Data Mining Applications*, Springer-Verlag, Berlin, 511-520.

Balbi S., Misuraca M. (2005), Visualization Techniques in Non Symmetrical Relationships, in Sirmakessis S. (ed.), *Knowledge Mining (Studies in Fuzziness and Soft Computing)*, Springer-Verlag, Heidelberg, [in corso di stampa]

Bolasco S. (1999), *Analisi multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione*, Carocci, Roma.

Bolasco S., Canzonetti A. (2003), Some insight on the evolution of 1990s' standard Italian, by Text Mining techniques and automatic categorization using the lexicon of daily "La Repubblica", *CLADAG03.*

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshmanet R. (1990), Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, 41(6), 391–407.

Eckart C., Young G. (1936), The approximation of one matrix by another of lower rank, *Psychometrika*, 1, 211-218.

Greenacre M.J. (1984), *Theory and Application of Correspondence Analysis*, Academic Press, London.

Hearst M. (1999), Untangling Text Data Mining, in *Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics*, Kaufmann Publishers, San Francisco, 3-10.

Lebart L., Morineau L., Piron M. (1995), *Statistique exploratoire multidimensionnelle*, Dunod, Paris.

Lebart L., Salem A. (1994), *Statistique textuelle*, Dunod, Paris.

Littman M.L., Dumais S.T., Landauer T.K. (1998), Automatic cross-language information retrieval using latent semantic indexing, in Grefenstette G. (ed.), *Cross Language Information Retrieval*, Kluwer.

Misuraca M. (2005), *La visualizzazione dell'informazione testuale. Contributi metodologici ed applicativi*, Tesi di Dottorato, Napoli

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