Multi-mode partitioning for text clustering to reduce dimensionality and noises

Livia Celardo[1], Domenica Fioredistella Iezzi[2], Maurizio Vichi[3]

[1]"La Sapienza" University – livia.celardo@uniroma1.it

[2]"Tor Vergata" University – stella.iezzi@uniroma2.it

[3]"La Sapienza" University – maurizio.vichi@uniroma1.it

Abstract 1

Co-clustering in text mining has been proposed to partition words and documents simultaneously. Although the main advantage of this approach may improve interpretation of clusters on the data, there are still few proposals on these methods; while one-way partition is even now widely utilized for information retrieval. In contrast to structured information, textual data suffer of high dimensionality and sparse matrices, so it is strictly necessary to pre-process texts for applying clustering techniques. In this paper, we propose a new procedure to reduce high dimensionality of corpora and to remove the noises from the unstructured data. We test two different processes to treat data applying two co-clustering algorithms; based on the results we present the procedure that provides the best interpretation of the data.

Abstract 2

Il co-clustering di dati testuali è stato proposto in letteratura per la classificazione simultanea di parole e documenti; tuttavia, anche se il principale vantaggio di questo approccio è l'interpretazione di cluster dai dati, ci sono ancora pochi studi su questo tema; mentre ad oggi la classificazione ad una via è largamente utilizzata per l'estrazione di informazione dai testi. Al contrario dei dati strutturati, i dati di tipo testuale soffrono di una molto ampia dimensionalità e della presenza di matrici largamente sparse, per cui diventa indispensabile attuare una fase di pre-processing sui testi prima di applicare qualsiasi procedura di clustering. In questo articolo, proponiamo una nuova proceduta per ridurre la dimensionalità dei corpora e per ridurre drasticamente il rumore presente all'interno dei dati testuali. Sono testate due differenti procedure per il trattamento dei dati applicando due diversi algoritmi di co-clustering; sulla base dei risultati presentiamo la procedura che fornisce la migliore interpretazione dei dati.

Keywords: co-clustering, disambiguation, *k*-means.

1. Introduction

Text Clustering is an unsupervised process that allows to classify large sets of documents in groups based on their attributes (Iezzi, 2012a) with the aim of reproducing the internal structure of the data (Iezzi, 2010); the main objective is to split the corpus in different subgroups on the basis of words/documents similarities (Iezzi and Mastrangelo, 2014). This technique corresponds to Cluster Analysis, and it doesn't require external information related to categories; in fact, clustering methodology is especially appropriate when no prior information is available about the data (Feldman and Sanger, 2007).

However, Feldman and Sanger (2007) identified in text clustering some distinctive characteristics that structured data partitioning have not: first of all, the major complexity and richness of internal structure in text documents implies a more problematic phase of dimension reduction; in fact, most of the words in a corpus are irrelevant to the categorization and represent only a noise in the data. Secondly, the problem of finding meaningful and concise descriptions of the clusters, that are not merely the centroids; lastly, the measurement of algorithm quality, that in text mining undoubtedly requires the subjective human judgment.

In text mining the most used partition algorithm for clustering is *k*-means (MacQueen, 1967), because of its efficiency in elaborating big data, even when it processes big sparse matrices (Iezzi, Mastrangelo and Sarlo, 2012; Iezzi, 2012b). Then, for general two-way data matrix (objects × variables) *k*-means algorithm is connected to one-way clustering, in which its objective is to classify objects; on the other hand, it is often necessary to identify of syntheses both in the direction of objects and variables, or, in text mining framework, texts and documents. In fact, very useful is the co-clustering approach (or two way clustering), that concerns simultaneous partitioning of rows and columns; the key idea is to identify sub-matrices of the observed data matrix, where each block specifies an object cluster and a variable cluster (Rocci and Vichi, 2008). In text mining contest, co-clustering is a very useful methodology (Balbi, 2012); the strength of this approach lies in finding clusters of documents characterized by groups of terms (Balbi, Miele and Scepi, 2010) with a high dimensionality reduction (Tjhi and Chen, 2006).

For its features, co-clustering is utilized in many studies in which it is involved in multiple attribute analysis; in text mining the study of co-clustering is proposed to deal with multi-partition of texts and words in digital library, because this approach is very useful in the observation of the co-occurrence of terms and documents in the same corpus (Xu, Zhang and Li, 2010).

In this paper, we propose a new procedure to reduce sparse Terms-Documents matrices and to remove most of the noise with the aim of words disambiguation; we apply on textual data two co-clustering partitioning methods to classify simultaneously terms and documents: the Double *K*-means (Vichi, 2001), and the Trimmed Double *K*-means (Vichi, 2013). The aim is to implement a co-clustering procedure to classify not only the terms, but also the documents on the basis of the distinctive contents within every text; this process is planned to obtain the best clustering of words/documents in terms of the higher level of results interpretability.

This article is structured as follows. In section no. 2, a brief literature review about co-clustering for text mining is exposed: in section no. 3, the methods are introduced; in section no. 4, some applications are discussed and conclusions and the future steps are drawn.

2. Co-clustering for text mining

Co-clustering is the problem of simultaneously clustering rows and columns of a data matrix, also known as bi-clustering, subspace clustering, bi-dimensional clustering, simultaneous clustering, block clustering. In the literature the ’70s, several algorithms have been proposed, and Hartigan (1972) explains that the principal advantage of this approach is the direct interpretation of the clusters on the data. Numerous algorithms and several applications are proposed in multiple domains, e.g in bioinformatics, in marketing, in medical science, in business and economics, and in many other fields. For a review on this topic, see Vichi (in press).

In text mining, the first step for applying clustering methods to corpus is to create a vector space model (Salton and Gill, 1983; Iezzi, 2012), and the corpus may be represented by a word by-document matrix **X** whose rows correspond to words and columns to documents. A non-zero entry in **X**, say x_{ij}, indicates the presence of word *i* in document *j*, while a zero entry indicates an absence. Typically, each document presents only a small number of words, then a corpus transformed into a term-document matrix **X** is very sparse with almost 99% of the matrix entries being zero. Co-clustering is more robust to sparseness, noise, and high-dimensional data, because the main aim is to exploit the "duality" between row and column clustering at all stages the row clusters incorporate column clustering, and vice versa. Co-clustering methods allow overcoming some typical issues of textual data transformed into matrices that are large, sparse and non-negative.

Agrawal *et al* (1999) underline that co-clustering is also related to the problem of sub-space clustering, in fact, the data is clustered by simultaneously associating it with a set of points and subspaces in multidimensional space. In this case, the data can be represented as sparse high dimensional matrices in which most of the entries are 0. Methods for subspace clustering can be adjusted to the co-clustering, e.g. Li et *al.* (2004) proposed an adaptive iterative subspace clustering for documents. Moreover, sub-space clustering can be considered a procedure of local feature selection, in which the words or repeated segments and/or documents selected are specific for each group. Principal Component Analysis (PCA) is a traditional way to select the features as linear combination of dimensions (Jolliffe, 1986). In text mining, traditional matrix approximations, such as Singular Value Decomposition/PCA do not preserve non-negativity or sparsity, because it has the disadvantage that the components extracted by this method have exclusively dense expressions, therefore interpretation can be very difficult.

In text clustering, two main approaches are well known: 1. Co-clustering with graph partitioning, in which **X**=[w_{ij}] is a term-document matrix of dimensions ($n \times p$), where *n* is the word types, *p* the documents, and w_{ij} is the weight of each word in a document, that corresponding to normalized term-frequency. **X** can be represented as a bipartite graph $G=(V_1 \cup V_2, E)$, where V_1 and V_2 are the vertex sets in the two bipartite positions of the G Graph, and E is the edge set (Dhillon, 2001; Iezzi, 2010). Each node in V_1 corresponds to one of the *p* terms, and each node in V_2 corresponds to one of the *n* documents. An undirected edge exists between $i \in V_1$ and node $j \in V_2$ if document *j* contains the term *i.* 2. Information-theoretic co-clustering, where the bag of words (BOW) table represents an empirical joint probability distribution of two discrete random variables. In this approach, the optimal co-clustering maximizes the mutual information between the clustered random variables subject to constraints on the number of row and column clusters. Dhillon *et al.* (2003) present a co-clustering algorithm that monotonically increases the preserved mutual information by intertwining both the row and column clustering at all stages. Balbi *et al.* (2010) proposed to use Goodman and Kruskal index on BOW as a criterion to prediction.

3. Methods

Let $\mathbf{X}=[x_{ij}]$ be a $(n \times p)$ Terms-Documents matrix weighted by the Term Frequency - Inverse Document Frequency (TF-IDF) index, where n are the terms and p are the texts. The TF-IDF index allows the terms discrimination, or in other words it consents to individuate those terms that are able to distinguish certain individual documents from the remainder of the collection; so, the best terms should have high term frequencies but low overall collection frequencies (Salton and Buckley, 1988).

To simultaneously classify rows and columns we apply on data the Double K-means model (Vichi, 2001), that is a two-mode single-partition algorithm to split the data matrix into rectangular blocks of homogeneous values. The model is specified as follows:

$$\mathbf{X} = \mathbf{U}\overline{\mathbf{X}}\mathbf{V}' + \mathbf{E}$$

where unknown partitions for rows and columns, specified by membership matrices **U** and **V**, need to be identified in order to best reconstruct matrix **X**.

The least-square assessment of the model leads to the formulation of the following quadratic optimization problem with respect to variables u_{ik}, v_{jl} and $\bar{x}_{kl}$:

$$\min \left\| \mathbf{X} - \mathbf{U}\overline{\mathbf{X}}\mathbf{V}' \right\|^2$$

subject to

$$u_{ik} \in \{0, 1\} \qquad i=1,\ldots,I;\ k=1,\ldots,K;$$

$$\sum_{k=1}^{K} u_{ik} = 1 \qquad i=1,\ldots,I;$$

$$v_{jl} \in \{0, 1\} \qquad j=1,\ldots,J;\ l=1,\ldots,L;$$

$$\sum_{l=1}^{L} v_{jl} = 1 \qquad j=1,\ldots,J.$$

A robust version of the previous model is the Trimmed Double K-means (Vichi, 2013), that is specified as follows:

$$\mathbf{DXB} = \mathbf{DU}\overline{\mathbf{X}}\mathbf{V}'\mathbf{B} + \mathbf{E}$$

subject to

$$\mathbf{D} = diag(d_{11},\ldots,d_{KK}),\ d_{kk} \in \{0, 1\}\ \text{, for } k=1,\ldots,K;$$

$$\mathbf{B} = diag(b_{11},\ldots,b_{QQ}),\ b_{qq} \in \{0, 1\}\text{, for } q=1,\ldots,Q;$$

$$tr(\mathbf{D}) = n(1 - \alpha_O);$$

$$tr(\mathbf{B}) = J(1 - \alpha_V).$$

That model introduces two strengths on precedent algorithm:

1) the estimator of location of the clusters associated to the block matrices reported in the centroid matrix is the median;
2) it is given two trimming costants indicating the fraction of words (α_O) and texts (α_V) with maximal distance from medoids to be removed in the dataset by two matrices, **D** and **B**, indicating those terms and documents that are considered outliers.

The solution of the Trimmed Double K-means is al LS problem in the binary variables **U**, **V**, **D** and **B** and continues $\overline{\mathbf{X}}$:

$$\| \mathbf{DXB} - \mathbf{DU}\overline{\mathbf{X}}\mathbf{V'B} \|^2 \rightarrow \underset{\mathbf{U},\mathbf{V},\mathbf{D},\mathbf{B},\overline{\mathbf{X}}}{min}$$

4. Data, results and conclusions

We used for the analysis two different corpora. The first is composed by 662 reviews published on booking.com in Italian language about thirty-five different five stars hotels in Rome city (Table 1); we selected around twenty most recent reviews for each hotel in the period from 2013 to 2015.

Table 1 *Selected five stars hotels in Rome city*

1	Portrait	**13**	Hotel de Russie	**25**	Grand Hotel Palace
2	The First Luxury Art Hotel	**14**	Hassler	**26**	Grand Hotel de la Minerve
3	Aldrovandi Villa Borghese	**15**	Boscolo Exedra	**27**	Aleph Hotel
4	The Inn	**16**	Hotel Eden	**28**	Hotel Bernini Bristol
5	Gran Melia Rome	**17**	Hotel Intercontinental de la Ville	**29**	DOM Hotel
6	Rome Cavalieri	**18**	Jumeirah Grand Hotel	**30**	Palazzo Montemartini
7	Parco dei Principi Grand Hotel	**19**	Baglioni Hotel Regina	**31**	Ambasciatori Palace Hotel
8	St. Regis	**20**	Hotel Splendide Royal	**32**	La Griffe
9	Hotel Raphael	**21**	The Westin Excelsior	**33**	Grand Hotel Plaza
10	Sofitel	**22**	Hotel Indigo	**34**	Radisson Blu Hotel
11	Hotel Majestic	**23**	Hotel Lord Byron	**35**	J.K. Place
12	Palazzo Manfredi	**24**	Hotel d'Inghilterra		

The second corpus "parks" is composed of 6,905 reviews on 11 Italian National Parks (d'Abruzzo, Lazio e Molise, Arcipelago Toscano, dell'Asinara, del Circeo, delle Dolomiti Bellunesi, Foreste Casentinesi, del Gargano, del Gran Paradiso, Gran Sasso, della Maddalena, and Pollino), and there are 508,691 word tokens, 6,460 word types (Iezzi & Zarelli, 2015).

Concerning the first dataset, before the pre-processing phase the first corpus contained 62,221 tokens and 10,144 types; after tokenization, lemmatization and removal of some stop-words, e.g. AND, THAT, THE, A, AN, auxiliary verbs (TO BE, TO HAVE), and the vocabulary of unique words was truncated by only keeping words that occurred more than ten times, we obtain a Term-Document-Matrix **X** of size (733 × 35). About the second dataset, after the pre-processing phase we obtain a Term-Document-Matrix **X** of size (2,827 × 11). After that, we represented the datasets on a Terms-Documents matrix, weighted by TF-IDF index.

We use two methods to remove yet the noise into the data, and improve the results of classification: 1) in the pre-processing step, to select words with a high peculiarity in texts, we calculate the normalized variation coefficient (CV)[1] on rows of **X**, and delete the values less than 0.25. On this reduced matrix **R**, we apply Double *k*-means algorithm (DKM); 2) we apply the Trimmed Double *k*-Means (TDKM) on **X,** that selects words to eliminate, and classify terms and parks**.**

In the first method, the idea is to eliminate the most common words, that make noise, and preserving those words that are very specific. In this case, we delete for the first dataset 318

[1] The coefficient of variation (CV) is defined as the ratio of the standard deviation σ to the mean μ.

words (Table 2), getting a reduced matrix **R** that has a size (415 × 35), and for the second dataset 133 words (Table 3), obtaining a reduced matrix **R** that has a size (2,694 × 11).

Table 2 *The first 20 words deleted using the CV method for Hotels data*

Italian type	English type	Italian type	English type
Hotel	HOTEL	*Roma*	ROME
Essere	TO BE	*Servizio*	SERVICE
Camera	ROOM	*Bagno*	BATHROOM
Tutto	ALL	*Più*	MORE
Molto	A LOT	*Qualità*	QUALITY
Non	NOT	*Notte*	NIGHT
Colazione	BREAKFAST	*Fare*	TO DO
Avere	TO HAVE	*Disponibile*	AVAILABLE
Anche	ALSO	*Posizione*	POSITION
Bello	NICE	*Trovare*	TO FIND

Table 3 *The first 20 words deleted using the CV method for Parks data*

Italian type	English type	Italian type	English type
Natura	NATURE	*Naturale*	NATURAL
Fare	TO DO	*giornata*	DAY
Zona	AREA	*scoprire*	TO DISCOVER
Parco	PARK	*paesaggio*	LANDSCAPE
Conoscere	TO KNOW	*strada*	STREET
Verde	GREEN	*famiglia*	FAMILY
Naturale	NATURAL	*amare*	TO LOVE
Andare	TO GO	*vedere*	TO SEE
un_posto	A PLACE	*anno*	YEAR
Angolo	CORNER	*visitare*	TO VISIT

Table 2-3 show that using the CV method, we have deleted for both datasets theme words, e.g. NATURE, HOTEL, PARK; verbs of state e.g. TO LOVE, or of movement, e.g. TO GO, and generic words, to recall the story of the trip, e.g. DAY, FAMILY, STREET, ROOM.

To detect the number of clusters both for rows and columns into **X**, we apply a graphical approach using Ward's method. Regarding the first dataset, we individuated three clusters for rows and three cluster for columns; the Double *k*-means algorithm got nine centroids on the basis of the words and documents groups (Table 4).

Table 4 *Double K-means centroids of Hotels data*

		Hotel		
		CCL1 (14 units)	*CCL2 (12 units)*	*CCL3 (9 units)*
Terms	*RCL1 (336 units)*	0.309	0.245	0.277
	RCL2 (57 units)	0.418	0.779	0.349
	RCL3 (22 units)	0.510	0.306	1.437

Table 5 *Hotels clusters of Hotel data*

CCL1	**CCL2**	**CCL3**
The Inn Hotel Raphael Palazzo Manfredi Hassler Boscolo Exedra Hotel Intercontinental de la Ville Jumeirah Grand Hotel Hotel Indigo Hotel d'Inghilterra Grand Hotel Palace Aleph Hotel Hotel Bernini Bristol La Griffe Radisson Blu Hotel	The First Luxury Art Hotel St. Regis Hotel Majestic Baglioni Hotel Regina The Westin Excelsior Hotel Lord Byron Grand Hotel de la Minerve DOM Hotel Palazzo Montemartini Ambasciatori Palace Hotel Grand Hotel Plaza J.K. Place	Portrait Sofitel Aldrovandi Villa Borghese Hotel de Russie Gran Melia Rome Rome Cavalieri Parco dei Principi Grand Hotel Hotel Eden Hotel Splendide Royal

Table 6 *Terms clusters of Hotel data*

RCL1	**RCL2**	**RCL3**
difficult - to avoid - noise noisy - to soundproof - defect to waste - expensive - to exceed beauty - super - nice spectacular - comfortable unforgettable - wine - evening extraordinary - panoramic pleasant - hydro massage suggestive - gym - magical fabulous - to lose	elegance - furniture - silent marble - old - chef - special charm - refined - ancient palace - design - enchanting tranquillity - art	dream - luxurious - spa appetizer - landscape - family villa - pool - breathtaking garden - park - quiet - holiday

The first cluster of words (Table 6) contains most of the terms of the corpus; the average values of TF-IDF in the three blocks of RCL1 are almost the same, then these words are lowly utilized approximately in the same way by hotel guests. Only the first group of hotel (Table 5) has an average value a little higher than others, so this group of terms are used a lot more for this cluster. These words are both positive and negative, and from those we can deduce that five stars hotels in Rome have not any equipments against city noises, that are felt by customers as a problem and that the hotels are considered too expensive related to the level of perceived services. On the other hand, words like EXTRAORDINARY or MAGICAL are used for each hotel in the same way, so the location of the city plays a fundamental role in the final perception of the guests.

The second group of terms is related to architectural qualities of the hotels; in the partition, the higher level of TF-IDF in connection with this cluster of terms is associated to the second group of hotels, that are the most distinctive and characteristics in terms of buildings.

The third cluster of words concerns mainly to the outdoors of the hotels and to the family holidays; an high average value of TF-IDF for these terms is connected to the third group of hotels, that are most evaluated for those features.

Concerning the second dataset, we individuated 6 clusters for the words and 3 for the parks. Applying Double *k*-means algorithm (DKM), we obtained 6 classes well interpretable for the contents, and three groups for the parks (Table 7).

Table 7 *Description of the clusters using DKM for Parks data*

cluster no.	Short description	The most important words	size
1	**SUSTAINABLE TOURISM**	CONSERVATION, PROTECTION, WELFARE, EXEMPLARY, DAMAGE	976
2	**DESCRIPTION OF THE MOUNTAIN PARKS**	SCALE, TERRACES, RUIN, ARCHAEOLOGICAL, CRAG, PEAK, SLOPE, ICED, TO SNOW	771
3	**WHAT TO SEE AND DO IN THE MOUNTAIN PARKS**	MOUNTAIN-BIKE, BEAR, FOX, SUEDE, GOAT,	569
4	**SENTIMENT AND ACTION OF TOURIST OF THE MARINE PARKS**	RELAX, PEACE, WATER-HIKING, TO SWIM, WONDERFUL	226
5	**WHAT TO SEE AND DO IN THE OF THE MARINE PARKS**	BEACH, FERRY, BOAT, PINEWOOD, BAY, RAFTING	126
6	**NATURE AND ACTIVITIES IN THE PARKS BOTH MARINE AND MOUNTAIN**	WATER-HIKING, POLLINO, PINUS_NIGRA, BASILICATA	26

Regarding to the procedure using TDKM, we experimented 4 thresholds of truncated mean 0.10, 0.15, 0.20 and 0.25.

As for the first dataset, Table 8 displays the first 14 words deleted using the TDKM *(thresholds=0.10).* The algorithm removed 73 terms, that are both generic words, e.g., EACH, NIGHT, than specific words, e.g. POOL, SPA, EXCELLENT.

Table 8 *The first 14 words deleted using the TDKM on Hotels data(thresholds=0.10)*

Italian type	**English type**	**Italian type**	**English type**
Avere	TO HAVE	*Personale*	STAFF
Scoprire	TO DISCOVER	*Perfetto*	PERFECT
Ogni	EACH	*Piscina*	POOL
Notte	NIGHT	*Via*	STREET
Curare	TO TREAT	*Spa*	SPA
Primo	FIRST	*Ottimo*	EXCELLENT
Cambiare	TO CHANGE	*Unico*	UNIQUE

For the second dataset, Table 10 shows the most conservative form of the corpus, with trimmed mean equals to 0.10, in which are deleted 284 words. Table 9 displays the first 20 words deleted using the TDKM *(thresholds=0.10).* The algorithm removed both theme words, e.g. PARK, NATURE, TO TRAVEL and specific words, e.g. BEAR, PORT.

Table 9 *The first 14 words deleted using the TDKM on Parks data (thresholds=0.10)*

Italian type	**English type**	**Italian type**	**English type**
abbazia	ABBEY	*Motonave*	MOTOR-BOAT
parco	PARK	*natura*	NATURE
cascata	WATERFALL	*Orso*	BEAR
traghetto	FERRY	*Pineta*	PINE_FOREST
isola	ISLAND	*Porto*	PORT
mare	SEA	*Traghetto*	FERRY
mediterraneo	MEDITERRANEAN	*Viaggiare*	TO TRAVEL

Table 10 *Description of the clusters using TDKM for Parks data*

cluster no.	Short description	The most important words	size
1	**DESCRIPTION OF ENVIRONMENT AND ACTIVITIES OF MOUNTAIN**	HOLY, STREAM, PINK, CLIMBING, SKIING, SQUIRREL DREAMING	768
2	**FOOD**	DRINK, BEER, FRY, SHRIMP, ICE CREAM, FOOD, INN, COOKHOUSE	635
3	**FAUNA OF THE MOUNTAIN PARKS**	HORSE, WILD BOAR, BIRD, AQUILA, DAINO, SEAGULL	410
4	**ACTIVITIES, AND SPECIALITY FOODS OF THE MOUNTAIN PARKS**	SHEEP'S GRILLED SKEWERS, TREKKING WATER, ABRUZZO, BASILICA, RAFTING, SPIRITUAL	357
5	**DESCRIPTION OF THE PARKS**	FAUNA, FLORA, FOREST, DAY, LOVELY, WONDERFUL, EYE, PRISTINE	212
6	**SENTIMENT OF THE VISITORS**	HOT, COLD, RELAX, SMELLING, REMEMBER, RESERVE	161

In this method, the marine parks have disappeared, in fact, the terms SEA, BEACH, MARINE, MEDITERRANEAN have been deleted. Moreover it emerges clearly a group that discusses about food. In the method no.1, the topic about food is inserted in the discussions in which people describe the activities to be performed in the parks.

In both cases, not only for the first dataset but also for the second, the method n.1 is more performing to eliminate the noise, because it cuts only the words that are common in all reviews, then, they are not very relevant for the classification.

On the other hand, the TDKM method by cutting the right and left tails of the matrix **X,** eliminates in each dataset in addition to the common terms also the words that are very important for the classification.

References

Agrawal R., Gehrke J., Raghavan P.. Gunopulos D. (1999). Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications, *ACM SIGMOD Conference*.

Balbi, S., Miele, R., and Scepi, G. (2010). Clustering of documents from a two-way viewpoint. In *10th Int. Conf. on Statistical Analysis of Textual Data*.

Balbi, S. (2012). Beyond the curse of multidimensionality: high dimensional clustering in text mining. *Bolasco S. and Iezzi DF (by) Advances in Textual Data Analysis and Text Mining-Special Issue Statistica Applicata-Italian Journal of Applied Statistics*.

Dhillon I. (2001). Co-clustering Documents and Words using bipartite spectral graph partitioning, *ACM KDD Conference*.

Dhillon, I., Mallela, S., and Modha, D. S. (2003). Information-theoretic co-clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 89-98). ACM.

Feldman, R., and Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press.

Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the american statistical association*, *67*(337), 123-129.

Iezzi, D. F. (2010). Topic connections and clustering in text mining: an analysis of the JADT network. *Statistical Analysis of Textual Data, Rome, Italy*, *2*(29), 719-730.

Iezzi, D. F. (2012a). Centrality measures for text clustering. *Communications in Statistics-Theory and Methods*, *41*(16-17), 3179-3197.

Iezzi, D. F. (2012b). A new method for adapting the *k*-means algorithm to text mining. *Ital. J. Appl. Stat*, *236*(22), 1.

Iezzi, D. F., Mastrangelo, M., and Sarlo, S. (2012). Text clustering based on centrality measures: an application on job advertisements. *11es Journées Internationales d'analyse statistique des données textuelles*, 515-524.

Iezzi, D. F., and Mastrangelo, M. (2014). The IEMA Fuzzy c-Means Algorithm for Text Clustering. *12es Journées Internationales d'analyse statistique des données textuelles,* 239-248.

Iezzi D.F.& Zarelli F. (2015). *What tourists say about the Italian national parks: a web mining analysis.* RIVISTA ITALIANA DI ECONOMIA, DEMOGRAFIA E STATISTICA, vol. LXIX, p. 73-82.

Jolliffe I.T. (1986) *Principal Component Analysis*. New York: Springer.

Li T., Ma S., Ogihara M. (2004). Document Clustering via Adaptive Subspace Iteration, *ACM SIGIR Conference*.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1(14), 281-297.

Rocci, R., and Vichi, M. (2008). Two-mode multi-partitioning. *Computational Statistics & Data Analysis*, *52*(4), 1984-2003.

Salton G., and McGill M. J. (1983). *Introduction to Modern Retrieval*. McGraw-Hill Book Company.

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

Tjhi, W. C., and Chen, L. (2006). A partitioning based algorithm to fuzzy co-cluster documents and words. *Pattern Recognition Letters*, *27*(3), 151-159.

Vichi, M. (2001). Double *k*-means clustering for simultaneous classification of objects and variables. In *Advances in classification and data analysis* (pp. 43-52). Springer Berlin Heidelberg.

Vichi, M. (2013). Robust Two-mode clustering. *Proceedings of the 59th World Statistics Congress of the International Statistical Institute.*

Vichi, M. (in press). Two-mode Clustering.

Xu, G., Zhang, Y., and Li, L. (2010). *Web mining and social networking: techniques and applications* (Vol. 6). Springer Science & Business Media.