Text clustering based on centrality measures: an application on job advertisements

Domenica Fioredistella Iezzi[1], Mario Mastrangelo[2], Scipione Sarlo[3]

[1] « Tor Vergata » University, Rome – stella.iezzi@uniroma2.it
[2] «La Sapienza » University, Rome – m.mastrangelo@uniroma1.it
[3]«La Sapienza » University, Rome – scipione.sarlo@uniroma1.it

Abstract 1

Text clustering methods allow automatic classification of a large set of documents. Many algorithms can be applied using the proposed methods for structured data. However, the corpus, once transformed from unstructured information into structured data, presents a high dimensionality and an overlapping of the clusters that could jeopardize understandability of the cluster description.

In this paper, we introduce a new method of detecting centroids of clusters. Centroids represent prototypes of mutually exclusive partitions, and they can therefore facilitate interpretation of the results to describe groups.

In this approach, after the preprocessing step, we establish links between documents by using co-occurrence information, within some lexical units. We use centrality measures to weigh texts and classify documents. We analyze 1,650 job announcements, published from January 1st, 2010 to April 5th, 2011 by 496 companies on DB SOUL (System University Orientation and Job).

Abstract 2

Le tecniche di text clustering permettono di classificare automaticamente un insieme di documenti. Molti algoritmi possono essere applicati, usando i metodi proposti per dati strutturati. Tuttavia il corpus, dopo essere stato trasformato da informazione non strutturata a dato strutturato, presenta un'alta dimensione e, generalmente, una grande sovrapposizione dei gruppi, compromettendo l'interpretabilità dei risultati.

In questo lavoro, proponiamo un metodo per individuare i centroidi dei gruppi. I centrodi rappresentano i prototipi di partizioni mutuamente esclusive, e quindi possono facilitare l'interpretazione dei risultati. In questo approccio, dopo la fase di pre-processing, noi consideriamo la matrice dei legami tra documenti e co-occorrenze. Usiamo le misure di centralità per pesare i testi e classificare i documenti. Noi analizziamo 1,650 annunci di lavoro, pubblicati tra il 1 gennaio 2010 e il 5 aprile 2011 da 496 aziende sul DB SOUL (Sistema per l'Orientamento Università-Lavoro).

Keywords: centroid, text clustering, adjacent matrix, centrality measures.

1. Introduction

Text clustering is a set of techniques of classifying unlabeled texts (or words) into disjoint subsets of clusters; such texts within a cluster are very similar to each other and texts in different clusters are very different. Clustering problems arise in various areas of text mining and information retrieval. Typically, given a corpus, each document is reduced to representation by a vector of frequencies of terms selected in an appropriate way (Volkovich *et al.*, 2005). Each dimension corresponds to a separate term. The Vector Space Document model (VSD) is very widely used to represent documents. The common framework of this data model starts with a representation of any document as a feature vector of the words that appear in documents of the data set (Chim and Dem, 2007). Cosine distance is commonly used in document clustering algorithms, which could provide a reasonable division (Iezzi, 2010a). There are several main classes of methods in cluster analysis (Everitt *et al.*, 2011). According to the method adopted to define clusters, the algorithms can be largely classified into the following types: A) Partition clustering attempts to directly decompose the data set into a set of disjoint clusters; B) Hierarchical clustering proceeds successively by either merging smaller clusters into larger ones, or by splitting larger clusters; C) Density-based clustering groups neighboring objects into clusters based on density conditions; D) Grid-based clustering is mainly proposed for spatial data mining; E) Support Vector Clustering maps, by means of a gaussian kernel, a high dimensional feature space, where the algorithm seeks a minimal enclosing sphere (Iezzi, 2012, in press). In text mining, a widely-used clustering algorithm is the *k*-means (MacQueen, 1967), because it is relatively efficient in processing large numbers of high dimensionality cases (Iezzi, 2012a in press). Its weakness is sensitivity to outliers (Han *et al.*, 2001) and the need to specify *ab initio* the number (*k*) of desired clusters and optionally the location of *n* initial centroids (Brimicombe, 2007). This criterion function begins with an initial set of randomly selected centroids and iteratively refines this set so as to describe the sum of squared errors. Centroids are centers of the groups, and they do not coincide with real text. They are prototypes of the mutually exclusive partitions, and they can facilitate interpretation of the results to describe groups, especially when the corpus is composed of overlapped documents. A key limitation of *k*-means is that it is based on spherical clusters that are separable in a way that the mean value converges towards the cluster center. This condition occurs rarely in practice. Much more frequently, texts have many parts in common and *k*-means algorithm may lead to the partition not being well separated and internally cohesive. Compared to *k*-means algorithm, Partitioning Around Medoids (PAM) is more robust, because it operates on the dissimilarity matrix and minimizes Euclidean distance. Moreover PAM allows you to select the optimal number of clusters, using silhouette index (Rousseeuw, 1987). Cerioli (2005) underlines that the performance of clustering algorithms can be measured through their ability to recover clusters that are already known to exist. In text mining, if the corpus is composed of short and standardized documents, it will not be possible to detect the best clusters applying *k*-means or PAM algorithm without making changes, which take account of the fact that the data are not well separated.

We propose a new method of classifying short and standardized documents. We use an integrated approach: in the first step, we calculate centrality measures to identify different levels of connectivity among documents; and, in the second step, we use an adapted version of PAM algorithm to classify documents.

The paper is organized as follows: in paragraph 2, we present the method; in paragraph 3, we describe an application and the main results and, finally, in paragraph 4, we present the conclusions and future developments.

2. Data and methods

We consider each text represented by a vector of weighted terms of the form: $d_j = (w_{1j}, w_{2j}, \ldots, w_{ij}, \ldots, w_{pn})$, where w_{ij} represents the weight for term i, attached to document d_j. By joining these vectors, we get the **D** word-term-by-document-matrix (Iezzi, 2012b *in press*). We could read the **D** matrix as a two-mode network, which is usually represented by an affiliation matrix (Wasserman & Faust, 2008). An affiliation is a matrix **M** of dimension (n x k), where n is the number of the documents that belong to the corpus, and k is the number of keywords selected in the pre-processing. The generic element of $\mathbf{M}=[m_{ij}]$ represents how many times the word j is in the document i. From an affiliation matrix it is possible to extract two adjacent matrices one for the documents and one for the keywords. We calculate the adjacent matrix **A**, post-multiplying the affiliation matrix by its transpose: $\mathbf{A}=\mathbf{MM}^T$.

A is a matrix of dimensions (n x n), whose generic element a_{ij} is equal to the number of overlapping words of the documents i and j. The diagonal of **A** measures the number of words attended by text i. The ties between the documents are expressed by weights on a scale of intensity. We can visualize a corpus using a simple graph (Scott, 2000). It provides an efficient exploration tool for getting familiar with a document collection. The main benefit of those types of visualizations is their ability to organize the exploration of textual data (Feldman, 2007). In a corpus composed of documents that are similar to each other, we can speak of nested corpus (Figure 1). Generally, hierarchical clustering procedure produces a set of nested clusters organized as a hierarchical tree, and it is very complex to detect the number of groups. Partitional methods divide a set of objects into non-overlapping clusters. We propose a mixed approach that allows detecting hierarchical structure, but enables us to discover the best partition. A nested corpus is composed of very similar texts and presents few words different from each other. Figure 1 shows that the documents in the sub-corpus G1 are the most tightly linked. The boundary of subset G2 is drawn with a medium criterion of connection. All documents of the sub-corpus G1 together with additional texts are connected to the sub-corpus B in weak level. Gradually, up to sub-set G4 which is the very weakest criterion of connectedness and so it includes all connected documents.

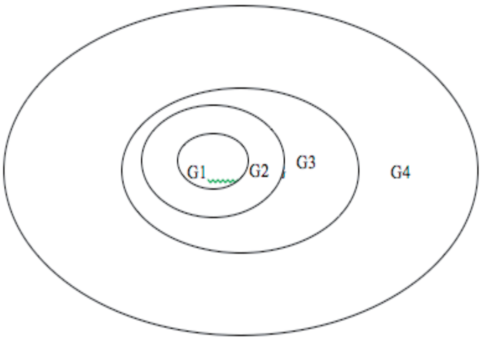


Figure 1 Nested documents

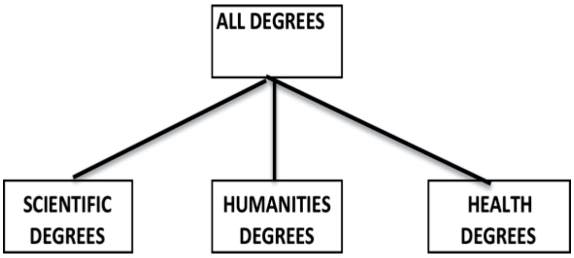
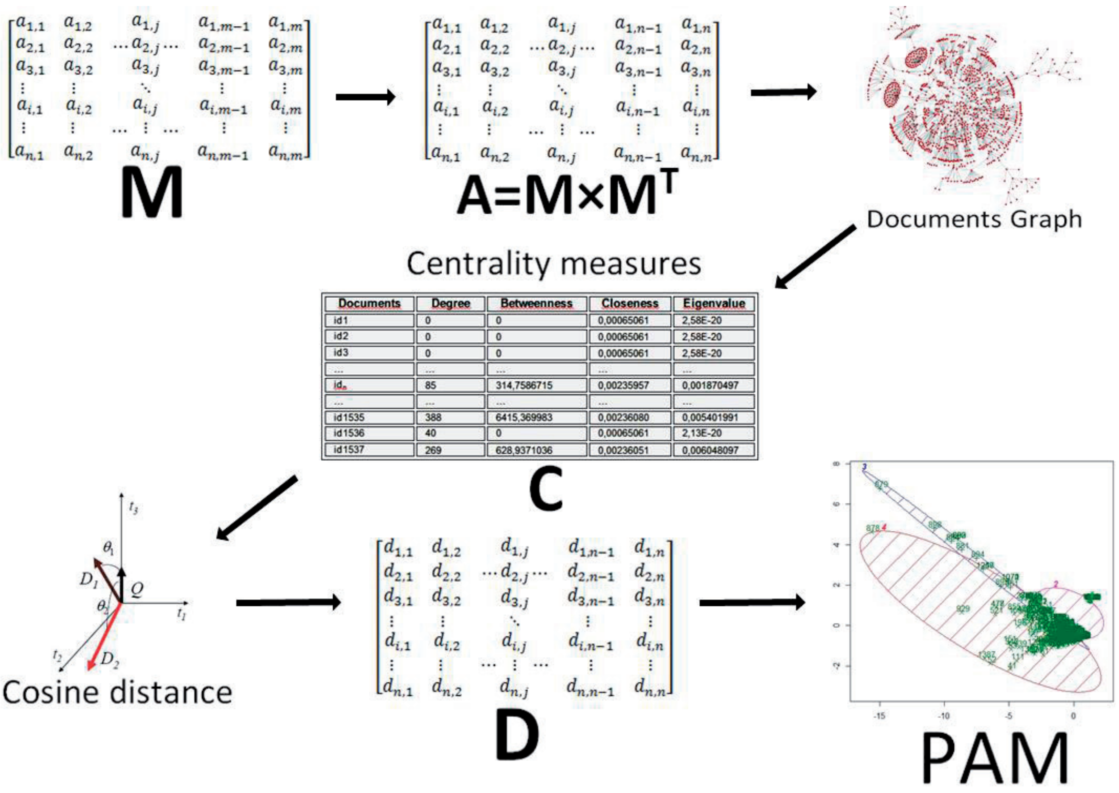


Figure 2 Nested Jobs respect to degrees required by companies

Figure 2 shows an example of nested documents. If we analyze job advertisements, e.g., we will find some jobs that require all degrees and others some specific degrees (Economics, Engineering, Humanities, Art, Law, Medicine and Surgery,…). In this case, the corpus generates a tree structure. If we apply a partitional approach to produce quick results, we will obtain a division of ads into badly classified groups.

We applied four centrality measures (degree centrality, betweenness, closeness, and eigenvector centrality) to select centroids of the clusters. In fact, when a document in a corpus is strategically located on the "shortest communication" path connecting pairs of others, then the text is in a central position, this documents belongs to the most central group (G1). When a document presents a very low level of centrality, it is weakest criterion of likeness (G4).

Based on centrality measures, we calculate cosine distance to perform PAM algorithm (Figure 3).



Documents	Degree	Betweenness	Closeness	Eigenvalue
id1	0	0	0,00065061	2,58E-20
id2	0	0	0,00065061	2,58E-20
id3	0	0	0,00065061	2,58E-20
...	...	...	...	...
id_n	85	314,7586715	0,00235957	0,001870497
...	...	...	...	...
id1535	388	6415,369983	0,00236080	0,005401991
id1536	40	0	0,00065061	2,13E-20
id1537	269	628,9371036	0,00236051	0,006048097

Figure 3 The steps of the method

One of the practices of graph theory is the identification of "the most important" texts in a network. We apply four measures of centrality that are widely used in network analysis: degree centrality, betweenness, closeness, and eigenvector centrality on **A** matrix (Freeman, 1979; Opsahl *et al.*, 2010). The procedures for the calculation of centrality are based on considering the weights as a measure of proximity between two texts. The simplest definition of text centrality is that central documents must have most ties to other documents; the closeness focuses on how close a document is to all the other texts in the corpus. Betweenness centrality underlines that interactions between two nonadjacent documents might depend on the other texts in the corpus, especially the documents that lie on the paths between the two. Eigenvector centrality assigns relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. The four above-mentioned measures of centrality have different selection criteria of the key players, allowing us to identify documents that are characterized by a specific role.

Figure 3 shows the steps of the clustering process. Starting from the affiliation matrix **M**, we calculate an adjacent matrix **A** of dimension (*nxn*), where *n* are the texts; we calculate degree centrality, betweenness, closeness, and eigenvector centrality on the **A** matrix. In this way, we obtain a **C** centrality matrix of dimensions (*n* x *c*), where *n* are the documents and *c* are the number of centralities. In this case, *c*=4, because we apply 4 centrality measures. We compute cosine distance on matrix **C.** It measures the cosine of angle formed by two document vectors that describe two profiles, **A** and **B**. Formally, the cosine distance is:

$$\cos(\alpha) = \frac{<A,B>}{|A||B|}$$

The cosine value is 1 when two documents are identical, and zero if there is nothing in common between them.

To select the initial centers, we apply the *Partition Around Medoids (PAM)* algorithm on **C** matrix. Kaufman and Rousseeuw (1990) proposed PAM algorithm, which maps a distance matrix into a specified number of clusters. A particularly nice property is that PAM allows clustering with respect to any specified distance metric. We use cosine distance. In this note, we propose partitioning around medoids by maximizing the average silhouette criteria. PAM is more robust than *k*-means algorithm, because it minimizes a sum of dissimilarities instead of a sum of squared Euclidean distances. It provides a display, the silhouette plot, which allows the user to select the optimal number of clusters. PAM computes the first *k* representative documents, called medoids. A medoid can be defined as that object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal. In the classification literature, such representative objects are called centrotypes. After finding the set of medoids, each object of the data set, is assigned to the nearest medoid.

3. An Application

We analyzed 1,650 job announcements, published from January 1st, 2010 to April 5th, 2011 on DB SOUL (System University Orientation and Job) by 496 companies. SOUL is a network of eight Universities (Sapienza, Roma Tre, Tor Vergata, Foro Italico, Accademia di belle Arti, Tuscia, Cassino, LUMSA - Libera Università degli Studi Maria SS. Assunta) in the Lazio region. The main goal of SOUL is to create a link between the job market and the university, giving university students and graduates a chance to improve their employability. Currently DB SOUL collects 52,000 graduate CVs, of which about 27,000 come from "Sapienza", 7,500 from "Roma Tre", 2,500 from "Tor Vergata", and 15,000 from other universities not only from the Lazio region (LUMSA, LUISS, Tuscia and Cassino), but also from other regions (e.g., Napoli Federico II, Salerno, Bari, Bologna, Chieti-Pescara, Lecce).

The main problem concerning our data is their nature: short texts sharing a similar language. Due to this, we decided to operate on a particular term-document matrix (**M**) generated after a restricted selection of terms.

In a previous work carried out on the same database we not only put together "soft" normalization, but also based this on lists (Bolasco, 2005). Normalization based on lists acts by recognizing multiple words, grammatical phrases and nominal groups to preserve their specificity within the corpus. To this aim, we used several lists, some of these were provided as resources by software Taltac2 (Bolasco, 2010; Giuliano, La Rocca, 2008), other lists were specifically built during the pre-processing (Iezzi, Mastrangelo, Sarlo, 2011). Among the latest kind of lists, we used one as a manual thesaurus to select words which were deemed to be about the same topic.

In our case, we focused on a particular topic: candidates' education, which means educational level and characteristics required to apply for a job position. Basically, any job announcement contains at least two different kinds of information: one referred to job position, the second one usually concerns the ideal candidate's profile. The affiliation matrix **M** is a dimension (1,537 x 401), where 1,537 are the selected job announcements and 401 the education of the ideal candidate. We built on a semantically based selection of terms that allow you to enhance the capability of discriminating our documents, at least on the basis of one of the two components of information contained in a typical job announcement. In other words, the aim of this procedure is to cluster together the documents sharing the same terms of educational topic, while maintaining in different clusters the documents which represent different aspects of educational topic.

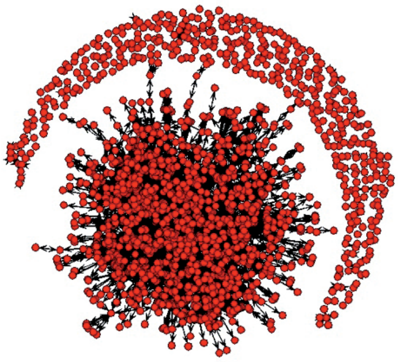


Figure 4 One-mode graph of adjacency network of texts

With our data, two types of networks may be generated by the affiliation matrix: one is a network with nodes representing documents and lines that indicate the extent to which they share the same terms; the other, a network with nodes representing terms and lines that indicate the extent to which they are used in the documents. These two kinds of networks can be derived from adjacency matrices related to the original affiliation matrix. In particular, we constructed the adjacency matrix **A**=**MM**^{T} (1,537 x 1,537). This matrix is square and represents links between documents: the relation between documents, as above-mentioned, is based on terms in common. Our assumption is that co-occurrence of terms in documents is an indicator of an underlying semantic connection among documents.

Figure 4 depicts one-mode graph of adjacency network of texts. It shows nested clusters. The relation we consider here is level of education between pairs of job announcements. In this graph there are *g*=1,537 documents and L=1,890 lines between the pairs of nodes.

We select the optimal number of clusters using silhouette measure. It measures the degree of confidence in the clustering assignment of a particular observation, with well-clustered observations having values near 1 and poorly clustered observations having values near -1. Table 1 shows the results of silhouette measures in respect to two procedures: 1) Lexical correspondence analysis applied on matrix **A**, and we classify documents based on 25 latent dimensions that explain 50% of variability (Method 1); 2) the proposed method in this paper (Method 2). Method 1 presents a classic way to classify textual data (Lebart *et al*, 1998)

For Method 1 the ideal number is 9 and for Method 2 it is 5, but also other solutions are good. The silhouette index is a dimensionless measure of the extent of clustering structuring that has been discovered by the PAM algorithm. It can be interpreted as follows: from 0.71 to 1.0 means that a strong structure has been found; from 0.51 to 0.70 a reasonable structure has been found; from 0.26 to 0.50 the structure is weak and could be artificial. Try additional methods of data analysis; less than 0.25 no substantial structure has been found. The silhouette index performs very badly for Method 1, in fact this measure is how closely it is matched to data within its cluster and how loosely it is matched to data of the neighbouring cluster, i.e. the cluster whose average distance from the datum is lowest.

K	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Method 1	0,489	0,438	0,452	0,477	0,484	**0,493**	0,444	0,454	0,414
Method 2	0,792	**0,804**	0,764	0,76	0,764	0,758	0,758	0,762	0,766

Table 1 Silhouette measures from 4 to 12 groups

Method 2 selected five prototypes, starting from 401 possible education levels. Table 2 shows educational profiles obtained with the Method 1 and Table 3 with the Method 2.

Cluster	**Size of the group**	**Educational Profiles**
g1	1217	Mixed group
g2	21	Engineering
g3	78	Engineering, economics and technical-scientific degrees
g4	97	Engineering and informatics
g5	45	Engineering, economics and technical-scientific degrees
g6	52	Graduates (indicating general area :juridical, economic, humanistic,..)
g7	18	Recent graduates (indicating general area :juridical, economic, humanistic,..)
g8	3	Pharmacology and nutrition science
g9	6	Business economics

Table 2 Educational profiles of the centroids of the Method 1

Method 1 identifies one large group, that is difficult to interpret and small groups of very few texts.

Cluster	**Size of the group**	**Educational Profiles**
g1	386	No qualifications
g2	416	Graduates in marketing, social sciences and humanities
g3	186	Engineering management
g4	403	Informatics, computer engineering, mathematics and statistics)
g5	146	Economics

Table 3 *Educational profiles of the centroids of the Method 2*

The first group (g1) contains 386 ads that do not require qualifications; the second group (g2), composed of 416 ads, encloses a range of degrees; graduation in marketing, publishing and journalism, social sciences and humanities. This cluster requires less technical-scientific degrees than groups 4 and 5. The third group (g3), made of 186 ads, is a rather heterogeneous cluster, because it classifies job ads for engineers and experts in business administration. This profile is very close to Engineering management or other new professional profiles which were designed by the Italian university reform (Aureli, Iezzi, 2006; Iezzi, 2011). The fourth group (g4) collects 403 ads which require technical and scientific training. In particular, the most requested degree is computer engineering, but also mathematics, physics, and chemical biology. The fifth group (g5) assembles 146 ads, for graduates in Economics

4. Conclusions

We proposed studying the structure of the corpus using centrality measures. In this way, we could visualize links between documents and the structure of the clusters. To select the centroids, we calculated different levels of centrality. The identification of “the most important” texts in a network will help us to detect centroids of the groups that represent prototypes of clusters. This method uses as input, a collection of real-valued similarity between texts in a corpus. The similarity indicates how well the documents are linked in respect to the centrality of texts.

This method presents the advantage that we can also detect clusters that are nested. If we apply a partition algorithm on term document matrix, we will identify one large group and small groups of very few texts.

The case study focuses on educational profiles of job announcements on DB SOUL. We detect “natural clusters” (Gordon, 1999) as the cluster of the new degrees with skills in technical scientific disciplines. Applying the PAM algorithm on Term Document Matrix we obtained one group with 1,250 job ads and 4 groups with few documents.

References

Aureli E., Iezzi D.F. (2006). Recruitment via web and information technology: a model for ranking the competences in job market, in *JADT 2006*, Vol.1 : 79-88.

Bolasco S. (2005). Statistica Testuale e Text mining: alcuni paradigm applicative. *Quaderni di Statistica*, vol.7.

Bolasco S. (2010). *TALTAC 2.10. Sviluppi, esperienze ed elementi essenziali di analisi automatica di testi*, ROMA: LED.

Bolasco S., Pavone P. (2008). Multi-class categorization based on cluster analysis and TFIDF, in S. Heiden & B. Pincemin (eds.) } *JADT2008*, Presses Universitaires de Lyon, vol. 1, pp. 209-218.

Brimicombe A. J. (2007). A dual approach to cluster discovery in point event data sets. *Computers, Environment and Urban Systems*, 31 (2007) 4–18.

Cerioli A. (2005). K-means cluster analysis and Mahalanobis metrics: a problematic match or an overlooked opportunity?, *Statistica Applicata – Italian Journal of Applied Statistics*, 17: 61-73.

Chim H, Deng. X. (2007) A new suffix tree similarity measure for document clustering. In *WWW*: 121-130.

Everitt, B.S., Landau, S., Leese, M. (2001), *Cluster Analysis*, Fourth edition, Arnold.

Feldman R, Sanger J. (2007), *The Text Mining Handbook*, Cambridge: Cambridge University Press.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks*: **1**(3), 215-239.

Giuliano L., La Rocca G. (2008). L'Analisi automatica e semi-automatica dei dati testuali. Software e istruzioni per l'uso. LED, Milano.

Gordon A. (1999). *Classification*, Chapman & Hall, London.

Han, J., Kamber, M., & Tung, A. (2001). Spatial clustering methods in data mining. In H. J. Miller & J. Han (Eds.), Geographic data mining and knowledge discovery (pp. 188–217). London: Taylor & Francis.

Houle, M. E., Kriegel H. P.; Kröger P. Schubert, E., Zimek, A. (2010). Can Shared-Neighbor Distances Defeat the Curse of Dimensionality? (PDF). Scientific and Statistical Database Management. Lecture Notes in Computer Science. 6187. pp. 482. doi:10.1007/978-3-642-13818-8_34. ISBN 978-3-642-13817-1. Edit

Iezzi D.F. (2010a). Intimate femicide in Italy: a model to classify how killings happened. In: PALUMBO F., LAURO C.N., GREENACRE M.J.. *Data Analysis and Classification*. p. 85-92, BERLIN: Springer-Verlag, ISBN/ISSN: 978-3-642-03738-2, doi: 10.1007/978-3-642-03739-9.

Iezzi D.F. (2010b). Topic connections and clustering in text mining: an analysis of the JADT network. In: *Statistical Analysis of Textual Data*. Rome, Italy, 9-11 June, vol. 2, 2(29): 719-730.

Iezzi D.F. (a cura di) (2011) *Indicatori e metodologie per la valutazione dell'efficacia del sistema universitario*. Vol. **1**. CLEUP, Padova ISBN 978 88 6129 621 3

Iezzi D.F. (2012a, *in press*). A new method for adapting *k*-means algorithm to text mining, *Statistica Applicata - Italian Journal of Applied Statistics.*

Iezzi D.F. (2012b, *in press*). Centrality measures for text clustering. *COMMUNICATIONS IN STATISTICS. THEORY AND METHODS*, ISSN: 0361-092.

Iezzi D.F., Mastrangelo M., Sarlo S. (2011). A text Classification Method to Measure Distance between Graduate Profiles and Labour Market Offers, in: Cerchiello P., Tarantola C. (eds.) *CLADAG 2001 Book of Abstracts 8th Scientific Meeting of the CLAssification and Data Analysis Group of the Italian Statistical Society*, University of Pavia September 7-9, 2011. ISBN 978-88-96764-22-0.

Kaufman L, Rousseeuw P. J. (1990) *Finding groups in data*. New York, Wiley.

Lebart L., Salem A., Berry L. (1998) *Exploring Textual Data*, Dordrecht, Kluwer Academic Pulisher

MacQueen J.B. (1967): "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297

Opsahl T., Agneessensb F., Skvoretzc J., (2010). Node centrality in weighted networks: Generalizing degree and shortest paths . *Social Networks*: **3**: 245-251 .

Rousseeuw P.J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20: 53-65.

Scott J. (2000) *Social Network Analysis.* A Handbook, London, Sage.

Volkovich Z, Kirzhner V.,,A. Bolshoy A., Nevo E., Korol A. (2005) The method of N-grams in large-scale clustering of DNA texts. *Pattern Recognition* 38: 1902-1912.

Wasserman S., Faust K. (2008). *Social Network Analysis*. Method and Applications, Cambridge (MA), Cambridge University Press.Wörterbuchnetz. (2011). Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier.