***Network text analysis: a two-way classification approach***

**Livia Celardo1, Martin G. Everett2**

*1 Department of Social Sciences and Economics, Sapienza University of Rome, Italy*

*2 School of Social Sciences, University of Manchester, United Kingdom*

***Abstract***

*Text clustering is a well-known method for information retrieval and numerous methods for classifying words, documents or both together have been proposed. Frequently, textual data are encoded using vector models so the corpus is transformed in to a matrix of terms by documents; using this representation text clustering generates groups of similar objects on the basis of the presence/absence of the words in the documents. An alternative way to work on texts is to represent them as a network where nodes are entities connected by the presence and distribution of the words in the documents. In this work, after summarising the state of the art of text clustering we will present a new network approach to textual data. We undertake text co-clustering using methods developed for social network analysis. Several experimental results will be presented to demonstrate the validity of the approach and the advantages of this technique compared to existing methods.*

***1. Introduction***

Automatic text analysis refers to the set of techniques that are used to describe and analyse textual data. In the data revolution era and with the constant growth of data availability (Gandomi & Haider, 2015; Rekik et al., 2018), text analysis has become an essential tool for analysing web contents (social media, newspapers, blogs, websites, forums, etc.) and other textual information for research or business purposes (Iezzi & Mastrangelo, 2014). Several methods and procedures for text analysis have been already proposed; however, the extraction of meaningful information from text remains a challenging task, and often the existing techniques are not satisfactory in terms of results. For this reason, in this paper we propose a new procedure for content classification where text mining and social network analysis approaches are merged.

As already discussed, there are a variety of procedures and methods that can be used to analyze text, each with its own particular objective and goals. In the Social Sciences, Content Analysis is a standard approach for analysing the content of texts (Krippendorff, 2004). It consists of making inference about some characteristics of a message from its content (Weber, 1983; Nasir, 2005; Aswani et al., 2018), by describing syntactic and semantic features of communication (Van Cuilenburg et al., 1988). The basic idea is that those words occurring together in relatively close proximity are related to a common theme or concept in the data analysed (Brier & Hopp, 2010). Content analysis techniques can be classified according to different criteria (Scharkow, 2013), such as being thematic/semantic/networking (Roberts, 2000) or supervised/unsupervised (Hillard et al., 2007). In the first classification, text analysis techniques differ according to the type of variable employed in each (Roberts, 1997): in thematic text analysis the focus is on occurrences of themes or concepts; in semantic text analysis the examination is of the interrelation of themes within sentences; and network text analysis extracts a network of interrelated themes. These three analyses are not mutually exclusive as they can be combined in the same analysis. For example, Semantic Network Analysis (Malrieu, 1994) or Relational Content Analysis (Popping, 2000) merges semantic and network text analysis, so instead of being simply role-based, this analysis also uses a relational model; this combination results in increased flexibility while still retaining the syntactic simplicity of semantic grammar. In the second categorization, techniques differ from being supervised or unsupervised; in the first case, content analysis is implemented using prior knowledge through category schemes (Van Cuilenburg et al., 1988); in the second case, themes are inferred from clusters of words (Hogenraad et al., 2003), which are identified using cluster analysis methods for textual data – i.e. text clustering algorithms.

Text clustering is an unsupervised data mining problem for document organization and browsing, corpus summarization and document classification (Aggarwal & Zhai, 2012). Many classes of algorithms, such the *k*-means or hierarchical methods, are extended to textual data, but the results are not always very satisfactory. For example, a key limitation of *k*-means algorithm is that it is based on spherical clusters that are separable in a way that the mean value converges towards the cluster centre. This condition occurs rarely in practice; in fact, much more frequently texts have many parts in common and the *k*-means algorithm may lead to the partition not being well separated and internally cohesive (Iezzi et al. 2012). Moreover, one-mode clustering synthetizes only words or documents, but often it is desirable to identify at the same time groups of words and texts. The simultaneous partitioning of rows and columns of a matrix is known as “co-clustering”, where the re-ordering creates rectangular blocks of non-zero entries. Text co-clustering is more robust with respect to sparseness, noise and high-dimensional data, because the main aim is to exploit the duality between rows and columns, so at all stages the row clusters incorporate column clusters, and vice versa (Celardo, 2017).

In text mining, some algorithms of co-clustering have been proposed. Three well-known methods for text co-clustering use graph theory (1), information theory (2) and latent block models (3). The co-clustering approach with graph partitioning (Dhillon, 2001; Beckett, 2016; Du et al., 2008; Liu & Murata, 2010) represents the term-document matrix as a bipartite graph, where two vertex sets – corresponding to documents and terms – are projected. The second model (Dhillon et al., 2003) defines optimal clustering as the one which maximizes the mutual information between the clustered random variables. The third algorithm (Govaert & Nadif, 2003) allows embedding simultaneous clustering of objects and variables in a mixed approach, based on the CEM algorithm. These methods, whilst they lead to a fast and scalable partitioning procedure, also have some drawbacks: in fact, for all of these methods it is not specified how to identify the optimal number of clusters nor the way to present the results (in terms of contents found in the data). Moreover, textual data differ in their own characteristics on the basis of the data source – e.g. tweets are not equal to news.

On the basis of the data type and according to the objectives of the research, several techniques can be used to analyse and classify the content of the text. Starting from this, our objective is to implement a text co-clustering procedure for automatic content classification where network theory is used. The idea is that, using a relation model based on word co-occurrences where the two dimensions are simultaneously classified – terms and documents, we can obtain a better understanding of the subjects inside the text. The novel contribution of this procedure is the merging of different approaches – from text mining and social network analysis fields – that could improve the analysis of textual data.

The paper is structured as follow: in the section 2, the theoretic framework is defined; in section 3, the methodology is presented; in section 4, the data used for the analysis is given; in section 5, the analyses and the results are presented; and in section 6, conclusions are drawn.

***2. Text vs. Network Analysis: the general framework***

There are several ways to encode a corpus – that is, a collection of documents; the most common is by using the Space Vector Model (Salton et al., 1975), where texts are represented through a term-document matrix **X**, in which rows correspond to words and columns to texts. The entries in **X** are weights, directly proportional to the importance of words in each text. The choice of the scheme depends on the objective of the analysis; one of the most used weighting schemes in text mining is the Term Frequency (TF) scheme, where a non-zero entry in **X**, say *xij*, indicates the number of occurrences of word *i* in document *j*, while a zero entry indicates an absence. Another way to encode a text is by using graph theory, interpreting the corpus as a network of entities. Let **X**=[*wij*]be aterm-document matrix of dimensions (*k* × *n*),where *k* is the word types,*n* the documents, and *wij* is the weight of each word in a document, that corresponding to normalized term-frequency. **X** can be represented as a bipartite graph, where *V1* and *V2* are the vertex sets in the two bipartite positions of the *G* graph, and *E* is the edge set (Iezzi, 2010). Each node in *V1* corresponds to one of the *k* terms, and each node in *V2* corresponds to one of the *n* documents. An undirected edge exists between and nodeif document *j* contains the term *i.*

Linking textual data and Social Network Analysis has many examples, which can be classified in three main categories (Diesner, 2012):

* Semantic Network Analysis;
* Network metrics definition for assessing relational data from texts;
* Methods definition for extracting relational data from texts.

When we extract one-mode networks from texts in order to identify theoretical representations of the information that people conceive in their minds – i.e. "concepts", we obtain Semantic Network Analysis (Sowa, 1984). A semantic network is a structure for representing knowledge as a pattern of interconnected nodes and arcs (Sowa, 2014); in fact, semantic network analysis represents the content of the text as a network of objects (van Atteveldt, 2008). In semantic network analysis human coding could be used to identify relations or not; in the last case the method is limited to the extrapolation and the analysis of the co-occurrence networks of words (Bullinaria & Levy, 2007) and it is also called network text analysis, that is a method where links between words in a text are encoded and networks of the linked words are built (Diesner & Carley, 2004). The use of co-occurrence statistics for semantic representation exploits the idea that words with similar meanings will tend to occur in similar contexts (Bullinaria & Levy, 2007). In fact, in a natural language words interact among themselves in many ways, and their relations cannot be inferred from word frequency distributions (Choudhury et al., 2010). Co-occurrence relations are based on three concepts (Dagan et al., 1995): co-occurrence within a consecutive sequence of words (*n*-grams), within syntactic relation (for example, verb-object) and within a limited context. In general, generating semantic vectors is achieved by simply going through a written corpus and counting the number of times each context word occurs within a window of a certain size around each target word (Bullinaria & Levy, 2012). In general terms, a word co-occurrence network can be represented by an undirected graph *G = (V, E)*, where *V* is the set of vertices representing all the different words and *E* is the set of edges representing the different adjacency relations of the word forms in sentence formation; so, two vertices are linked by an edge if the two corresponding words are adjacent within at least one sentence (Liu & Cong, 2013).

Network approaches to textual data can also be used also with a two-mode matrix in which words stand for objects and texts for variables, this matrix as often referred to as an affiliation matrix (Iezzi, 2012; Iezzi et al., 2012). An affiliation matrix is a two-mode matrix where rows and columns refer to different sets of nodes (Borgatti et al., 2013); so we can represent affiliations as graphs where nodes correspond to entities (such as words and texts) and edges correspond to ties of affiliation among the entities. Given the affiliation matrix X, we could construct two non-directional adjacency matrices, respectively **W** = XXT and **T** = XTX, where are shown the links among words or texts. Therefore, the adjacency matrices **W** and **T** are a (*n* × *n*) and a (*p* × *p*) symmetric matrices, where *n* and *p* are, respectively, the number of texts and words in a corpus. The entry values in **W** and **T** are weights, corresponding to the strength of the connection between pairs of words or documents.

Starting from an adjacent matrix, we could be interested in classifying the nodes, in order to identify the contents within the text. Using Social Network Analysis, the classification objective is to find groups of connected entities and is called Community Detection. Documents with more words in common and words more frequently occurring together in the same documents will have a higher probability of belonging to the same group or community. In general terms, a community is defined as a sub-group of nodes in a network that are more connected to each other than to nodes outside the group. In the field of text analysis a community of words refers to a group of terms that frequently co-occur in the same documents, while a community of documents indicates a group of texts that have more words in common.

In Social Network Analysis there are two way to dealing with two-mode data (Everett & Borgatti, Forthcoming): the first is to convert data in two adjacent matrices - also called projections - while the other consists of a direct approach on the affiliation matrix, such as blockmodelling of multi-mode data (Doreian et al., 2004). In the first case the original data matrix is transformed in two adjacent matrices that are partitioned separately, so we can obtain a partition for rows and a partition for columns. Moreover, using the Dual Projection approach (Everett & Borgatti, 2013; Melamed, 2014) the partitioned projections are finally combined using the original data. Then, the final result is the original data matrix divided into blocks with a density value associated to each one. The strengths of this method compared to the others are that a different number of clusters for rows and columns could be set and no information is lost in the process.

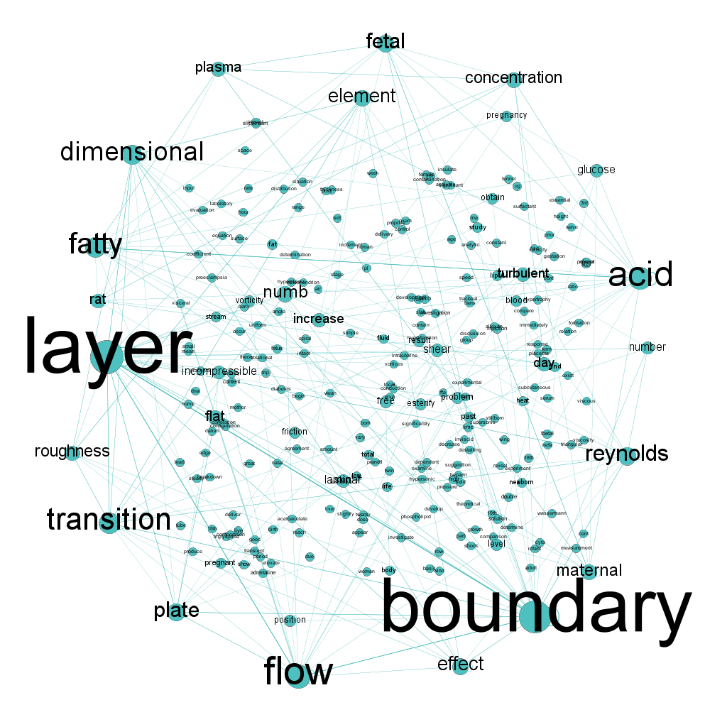
Once the affiliation matrix is transformed into two adjacent matrices, there are many partitioning methods for one-mode data that can be applied on these. For single mode networks the most popular techniques are those that optimization of modularity is achieved. Modularity is a numerical score that compares the edges in the network that connect vertices in the same community minus the expected value of the same quantity in a network with the same community divisions but random connections between the vertices (Newman & Girvan, 2004).When modularity is higher than zero, the number of within-community edges is better than random while a value equal to one means that we have a network with strong community structure. Newman’s community detection algorithm (Newman, 2006) is the most used among those where modularity is optimized, mostly because of its accuracy. So, for relatively small networks this method is suitable, but for bigger dataset - such as textual ones - the computational complexity means this approach is not feasible. A similar community detection method created especially for large networks in which modularity is optimized is the Louvain algorithm (Batagelj et al., 2014; Blondel et al., 2008); its computational efficiency allows to find communities within large datasets in a short time, but with a small loss in the accuracy of the result

***3. Method***

The procedure for content classification we show in this paper (Figure 2) is based on the idea of applying network theory to two-mode matrices in order to propose a text co-clustering method. This choice is motivated by the following:

1. The text could be interpreted as a network of words instead of a set of terms. In this way, words are read not only in terms of frequencies but also in terms of relations (Figure 1);
2. Clustering methods designed for structured data – like *k*-means or Ward’s algorithms – are not able to perform a good classification on textual data;
3. In text co-clustering, in spite of a better interpretation of the texts, there are still few methods, while one-way partition is widely utilized for information retrieval.

*Figure 1 – Example of word co-occurrence network, extracted from the dataset CRAN+MED\_10*



*Source: our elaboration.*

Differently from semantic network analysis, where one-mode networks are constructed on the basis of the word sequences, the idea here is to start from the term-document matrix, using the word co-occurrences. A major advantage of using word co-occurrence lies in their unambiguity, because they are defined and extracted from the language data in a theory-neutral manner (Liu & Cong, 2013). So, the purpose is not to analyse the semantics of the corpus but to identify the contents within the texts starting from the relationship between words and documents.

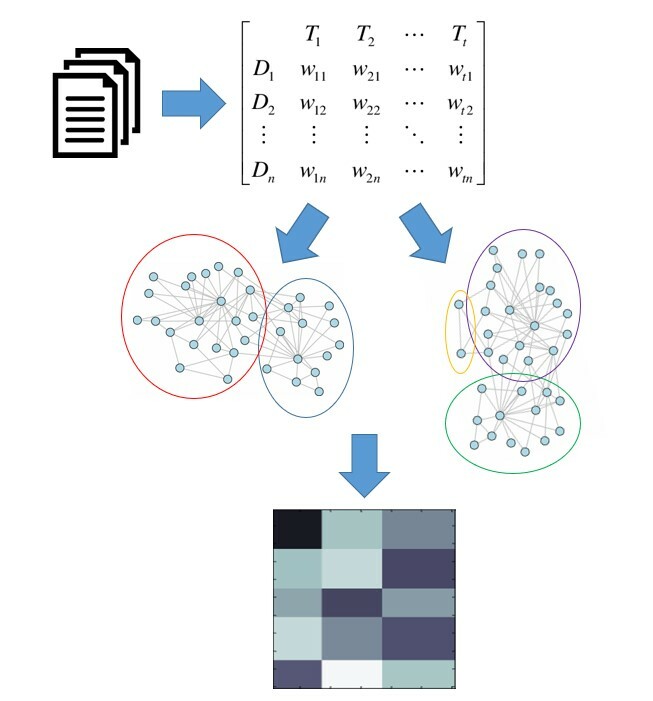
The procedure proposed is articulated in several steps (Table 1); in order to implement it, two software packages have been used. The first three steps are carried out with IRAMUTEQ, a free software package using the R interface for multidimensional analysis of texts. The remaining phases are realised with R software, using the *igraph* (Csardi & Nepusz, 2006) and *blockcluster* (Bhatia et al., 2014) packages.

*Table 1 – Procedure for text co-clustering with network approach, description of the steps*

|  |
| --- |
| **Procedure steps** |
| 1. Collection of documents 2. Pre-processing phase 3. Construction of the term-document matrix 4. Generation of the two adjacent matrices 5. Use of the Louvain algorithm on the two networks for the identification of the groups 6. Dual-projection approach 7. Interpretation of the results |

To do that, we start collecting a set of documents (1); on these texts, we apply a pre-processing phase (2) to identify the text units – i.e. the tokens – and select the keywords, in order to reduce the dimensionality and remove the noise. Encoding the texts, we construct a term-document matrix (3) weighted by the term frequency, and then we transform it in two adjacent matrices, one for the words and another for the documents (4); to produce the adjacent matrices we use the cross-product formulas. On these two networks, we use the Louvain algorithm (5) in order to identify the “communities” of terms and texts. The Louvain algorithm finds high modularity partitions of large networks in a short time and unfolds a complete hierarchical community structure for the network. The use of this method solves a crucial issue in classification problems, that is the definition of the number of clusters. To compare the results in terms of correct classification rate of the Louvain algorithm on our data we repeated the analysis using the blockcluster method, an algorithm based on latent block model (see section 1) which takes into account the block clustering problem on both the words and documents sets. After generating the groups, we use the dual projection approach (6) for returning to the two-mode matrix, identifying the significant intersections of words-documents through the density matrix (7). In this way, we are able to clearly detect the contents, and also the groups of documents where those contents are characterized.

*Figure 2 – Procedure for text co-clustering with network approach, from the collection of documents to the interpretation of the results.*



*Source: our elaboration.*

***4. Data***

In order to validate the procedure proposed here, we collected several corpora (Table 2) with different characteristics, both in terms of *size* – i.e. the number of occurrences – rather than in terms of *average length of the documents* – i.e. the number of words. In this approach, the average length of the documents is very important because it determines the numbers of co-occurrences, and so the number of edges in the networks we construct. So, we decided to test the procedure on three famous collections already used for Information Retrieval and Latent Semantic Indexing:

* 1. Cranfield (CRAN): 1,398 abstracts in aeronautics and related areas originally used for tests at the Cranfield Institute of Technology in Bedford, England;
  2. Medline (MED): 1,033 abstracts in biomedicine received from the National Library of Medicine;
  3. Cisi (CISI): 1,454 abstracts in library science and related areas published between 1969 and 1977 and extracted from Social Science Citation Index by the Institute for Scientific Information;
  4. on a mixture of 10 papers related to two different subjects, that are DIABETES DISEASE and SOLAR ENERGY themes. The papers have been selected from Google Scholar using as keywords *diabetes disease* and *solar energy* expressions;
  5. on a mixture of two collections of tweets, extracted from Twitter, regarding BREXIT and WEDDING subjects; the tweets have been scraped from Twitter with the R software using as keywords *Brexit* and *Wedding* expressions. In order to validate the results obtained analysing the tweets, we created also a dataset where multiple tweets are merged together. In this way, we can be sure that the outcome of the analysis implemented on Twitter data are due to the short length of the texts and not to the general word use in tweets.

We created several mixtures for the tests in order to verify the goodness of the method through the correct classification rate – i.e. the percentage of documents correctly assigned to their group.

*Table 2 – Mixture of corpora used for testing the procedure; for each dataset, details are exposed.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Datasets** | **Details** | **Size** | **Document average length** |
| CRAN+MED+CISI | Mix of 1,398 aeronautical systems, 1,033 medical and 1,454 information retrieval abstracts | 562,471 occurrences | 144 tokens |
| CRAN+MED | Mix of 1,398 aeronautical systems and 1,033 medical abstracts | 386,771 occurrences | 159 tokens |
| CRAN+CISI | Mix of 1,398 aeronautical systems and 1,454 information retrieval abstracts | 402,462 occurrences | 141 tokens |
| MED+CISI | Mix of 1,033 medical and 1,454 information retrieval abstracts | 335,933 occurrences | 135 tokens |
| CRAN+MED\_10 | Mix of 10 aeronautical abstracts and 10 medical abstracts extracted from the complete datasets CRAN and MED | 2,627 occurrences | 131 tokens |
| PAPERS | Mix of 5 papers about diabetes and 5 papers about solar energy | 33,717 occurrences | 3,371 tokens |
| TWITTER\_1 | Mix of 5,000 tweets about Brexit and 5,000 tweets about wedding | 178,109 occurrences | 18 tokens |
| TWITTER\_2 | Mix of 1,000 tweets about Brexit and 1,000 tweets about wedding | 36,383 occurrences | 18 tokens |
| TWITTER\_3 | Mix of 1,000 tweets about Brexit and 1,000 tweets about wedding; the tweets have been merged together, creating 200 texts (each one composed of 10 tweets) | 36,383 occurrences | 180 tokens |

1. ***Analyses***

As mentioned in the previous section, for the data analysis several corpora were used. For all the corpora, in the pre-processing phase we lemmatized the texts, we eliminated the hapax forms and we removed stop-words – articles, auxiliaries, conjunctions, pronouns and prepositions. For the corpus obtained by the mixture of MED, CRAN and CISI datasets, we will show the complete procedure and all the results. For the other corpora, we describe for each one a synthesis of the results obtained.

For the first corpus (CRAN+MED+CISI), after the text was pre-processed and then encoded, we got a term-document matrix composed of 5,456 words and 3,885 documents. Starting from this affiliation matrix (*terms* × *documents*), we created two adjacent matrices – (*terms* × *terms*) and (*documents* × *documents*) – using the cross-product formula. Using the *igraph* package, the adjacent matrices were then transformed in two undirected graph objects. We applied on these networks the Louvain algorithm for community detection. The algorithm clearly identified the number of clusters, dividing both documents and words in three groups, composed as exposed in Tables 3-4.

*Table 3 – Confusion matrix, clusters of documents obtained by using the Louvain algorithm (modularity = 0.24).*

|  |  |  |  |
| --- | --- | --- | --- |
|  | *C1* (n=1,373) | *C2* (n=1,008) | *C3* (n=1,504) |
| CRAN | **1,345 out of 1,398 (96%)** | 17 | 36 |
| MED | 24 | **979 out of 1,033 (95%)** | 30 |
| CISI | 4 | 12 | **1,438 out of 1,454 (99%)** |

The confusion matrix (Table 3) shows that the algorithm assigned to the first group of documents the 96% of the CRAN abstract, to the second cluster the 95% of the MED abstract and to the third group the 99% of the CISI abstract. Then, the misclassification rate in this test is very low – 3% on average.

*Table 4 – Cluster of words, first 25 words, listed by degree centrality score (modularity = 0.23).*

|  |  |  |
| --- | --- | --- |
| **R1** (n=2251) | **R2** (n=2042) | **R3** (n=1163) |
| Study | Time | Result |
| Case | Form | Present |
| Increase | Well | Show |
| Type | Develop | Find |
| Suggest | Discuss | Effect |
| Produce | System | High |
| Great | Datum | Method |
| Occur | Problem | Large |
| Normal | Include | Numb |
| Control | Development | Obtain |
| Group | Report | Test |
| Factor | Mean | Condition |
| Level | Information | Small |
| Significant | Base | Determine |
| Observe | Subject | Value |
| Change | Term | Analysis |
| Total | Paper | Consider |
| Early | Year | Function |
| Patient | Process | Low |
| Specific | General | Lead |
| Treatment | Provide | Compare |
| Period | Work | Theory |
| Long | Relate | Point |
| Cause | State | Order |
| Activity | Technique | Flow |

In order to reconstruct the connection between documents and words, the dual projection was applied on the classification results. To do that, the permutation of rows and columns on the basis of the membership matrices was performed. On the permutated matrix, for each block (intersection of clusters) we calculated the density score (Table 5); the highest score on the row profile specifies a connection between the two groups.

*Table 5 – Density score matrix.*

|  |  |  |  |
| --- | --- | --- | --- |
|  | *C1* | *C2* | *C3* |
| *R1* | 0.001 | **0.004** | 0.001 |
| *R2* | 0.002 | 0.001 | **0.006** |
| *R3* | **0.14** | 0.003 | 0.002 |

The density score matrix indicates the connections between row and column clusters. As shown in the table, the first group of documents, mostly composed of CRAN abstract, is linked to the third group of words. The second group of documents, represented by MED abstracts is associate with the first group of terms; finally, the third cluster of documents, where are most of the CISI abstracts, is related to the second group of terms. From the dual projection approach, we confirmed that the procedure has correctly co-classified the two dimensions of the matrix, both for rows and columns.

Then, we repeated the same procedure implemented on the first corpus on the other corpora presented before in order to confirm the validity of the method. Furthermore, we compared the results we obtained with our procedure with another method of co-clustering (Table 6) belonging to the category of latent block models (see section 1), used to classify words and documents starting from the term-document matrix. This method, called *block mixture model* (Govaert & Nadif, 2003), allows to embed simultaneous clustering of objects and variables in a mixture approach. It is implemented within an R package, named *Blockcluster*.

*Table 6 – Comparison of the results obtained on the same corpora, Louvain vs. Blockcluster methods for text co-clustering.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Datasets** | **Documents classification results** | | | |
| ***Louvain method*** | | ***Blockcluster method\*\*\**** | |
| *No. clusters* | *Correct classification rate* | *No. clusters\** | *Correct classification rate* |
| CRAN + MED (10) | 2 | CRAN: 100%  MED: 100% | 8 | CRAN: 100%  MED: 90% |
| CRAN + MED | 4 | CRAN: 99%  MED: 98% | 4 | CRAN: 97%  MED: 57% |
| CRAN + CISI | 3 | CRAN: 98%  CISI: 99% | 2 | CRAN: 99%  CISI: 90% |
| MED + CISI | 3 | CISI: 97%  MED: 99% | 2 | CISI: 99%  MED: 99% |
| CRAN + MED + CISI | 3 | CRAN: 96%  MED: 95%  CISI: 99% | 2 | CRAN: 99%  MED: 63%  CISI: 90% |
| PAPERS | 2 | PAPERS\_1: 100%  PAPERS\_2: 100% | 9 | PAPERS\_1: 100%  PAPERS\_2: 100% |
| TWITTER1 | 35 | \*\* see the notes | 3 | TWITTER1\_W: 28%  TWITTER1\_B: 74% |
| TWITTER2 | 32 | TWITTER2\_W: 90%  TWITTER2\_B: 83% | 2 | TWITTER2\_W: 83%  TWITTER2\_B: 93% |
| TWITTER3 | 2 | TWITTER3\_W: 100%  TWITTER3\_B: 100% | 2 | TWITTER3\_W: 72%  TWITTER3\_B: 66% |

\* The no. of clusters for the blockcluster method has been calculated using the Calinski-Harabasz index.

\*\* The results showed that the algorithm classified most of the tweets in one of the 35 groups, while the other clusters contained just few texts.

\*\*\* A continuous distribution follows the term frequency.

As shown in the table, the procedure using the Louvain method gives good results, except for corpora with a short average length, like twitter messages. In fact, on long tweet documents dataset (TWITTER3) the results are very satisfactory, so the method performs less on tweets due to the short length of the texts, and not to the general word use in tweets. For the other corpora the results show good classification rates, particularly when compared with the blockcluster method. Moreover, the network approach solves the problem of the identification of the number of groups, because the Louvain algorithm identifies by itself the number of communities; for the blockcluster method we had to know before how many clusters we needed. Another issue this procedure solves is the way to show the most important words representing each group: we decided to use the degree centrality score, because it is a measure that shows the importance of the terms on the basis on the connection with the other words, as opposed to simple term frequency.

***6. Discussion***

*6.1 Theoretical contribution*

In this article we presented a new approach for content classification of textual data, using techniques developed for social network analysis. In particular, we proposed a way for classifying both words and documents, showing the strengths of this method and the weaknesses of other approaches. Moreover, we extended the current knowledge in terms of network text analysis by comparing the different methodologies where text and network analysis are merged. The theoretical contribution of the procedure presented in this paper lies in its major capability when treating textual data; in fact, the results showed the potentiality of the method in analysing and extracting the content from the text, in connection with the different documents. In terms of methodological advancements, this procedure presents several strong points (see section 3) in comparison with previous methods. However, while this procedure has demonstrated its ability in treating documents with medium and long length, on the other hand it presents several problems for the analysis of short texts. This criticism of the procedure was not really unexpected, because when using word co-occurrence short texts result in a limited number of relevant terms (for each document) and so few edges in the network.

*6.2 Implications for practice*

Our results show that in order to extract meaningful information from text it is important not only to examine the algorithm applied on the corpus but also the capacity of the researcher to look at the entire process of data analysis. Texts differ significantly in their characteristics on the basis of the data source, so it is crucial to treat them in the right way depending on the data type. The implications for practice in adopting this procedure are numerous; in fact, text mining tools are now experiencing a huge growth in multiple domains – e.g. business, computer science, psychology, learning context, and so on. As companies’ interest in monitoring and getting insights from online textual data (reviews, comments, opinions, etc.) continues to increase, the research findings provide new instruments for analysing this amount of information. Especially in the business sector, the extraction and the analysis of textual data has become for organizations a strategic expertise to increase and maintain a competitive advantage over others and to support their decision-making processes (Fronzetti Colladon & Gloor, 2018; Gloor et al., 2017). Suffice it to say that most of an organization’s information is contained in text documents, such as emails, memos, customer correspondence, and reports (Tan, 1999). Moreover, the development of the Internet has generated a wealth of virtual communities (Antonacci et al., 2017) producing textual data, which contain hidden knowledge for businesses to leverage for a competitive edge (He et al., 2013; Fronzetti Colladon, 2018). In this sense, online reviews represent for firms new knowledge (e.g., brand popularity) and interesting patterns, which can be used for example to analyse how they can impact the sales of the product for e-commerce platforms (Kaushik et al., 2018; Singh et al., 2017) or to identify the reasons for positive or negative sentiments on social media (Shirdastian et al., 2017). Social media data can be also investigated to derive insights related to theoretical frameworks within political science (Grover et al., 2018) or to identify victims asking for help in a disaster situation (Singh et al., 2017). In particular, these data can help companies to realize strengths and weaknesses, enhance business effectiveness, improve customer satisfaction (Lau et al., 2005; Gloor et al., 2017) and reduce marketing costs by better understanding customer preferences (Xu et al., 2017; Wixom & Watson, 2010). For instance, financial sector news articles have been used to analyse company movements and market changes in order to extract and categorize Business Intelligent factors (Chung, 2014). In the banking industry online reviews have been matched with annual financial performance of 68 banks, with the aim of testing patterns of the bi-directional relations between eWOM indicators and banks’ profitability over time (Tang et al., 2016).

***7. Conclusions***

Information extraction from documents can be a challenging task; in the course of time, more and new methods have been proposed in order to discover and analyse the content of the texts. In this sense, text clustering allows us to analyse and synthesize a collection of documents, detecting groups of similar entities within the corpus; the aim could be to classify texts, words or both. The third case is known as two-way clustering, where groups of terms and groups of documents are identified. As already shown, text co-clustering presents several advantages and strengths in treating lexical tables, even when dealing with big and sparse matrices.

In this study, we proposed a novel procedure for content classification using a network approach; starting from the simultaneous classification of words and documents, the idea is to identify the contents of the corpus through the linguistic specificities. To do that, we implemented a relational model based on word co-occurrences where a community detection algorithm is carried on the affiliation matrix. In order to verify the validity, we collected several corpora with different characteristics – both in terms of size than in terms of average length of the documents. On these corpora we applied the co-clustering procedure, comparing the results with other models.

*7.1 Limitations and future research direction*

This study has some limitations that offer opportunities for further research. We showed in this paper a new procedure for content classification that takes into consideration the relational structure of the texts. Our objective was to construct a process of text analysis that could be fast, scalable and applicable to all the type of documents. So, we applied it on several corpora in order to verify the validity of the technique; what we found is that the procedure has shown good results for most of the corpora, except for the texts with a short average length, like tweets. In fact, comparing the results with another method of co-clustering we discover that our technique produces better results for all the document mixtures excluding those composed of short texts. This limitation lies in the fact that the method is based on the relational structure of the texts, and short documents present less links between words and documents. Future studies could try to adapt the procedure in order to find the way to analyse also the content of short texts, maintaining the relational approach to the documents.

***References***

Aggarwal, C. C., & Zhai, C. (2012). A survey of text clustering algorithms. In C. C. Aggarwal, & C. Zhai (Eds.), *Mining text data* (pp. 77-128). Berlin: Springer Science & Business Media.

Antonacci, G., Fronzetti Colladon, A., Stefanini, A., & Gloor, P. (2017). It is rotating leaders who build the swarm: social network determinants of growth for healthcare virtual communities of practice. Journal of Knowledge Management, 21(5), 1218–1239. https://doi.org/10.1108/JKM-11-2016-0504.

Aswani, R., Kar, A. K., Ilavarasan, P. V., & Dwivedi, Y. K. (2018). Search engine marketing is not all gold: Insights from Twitter and SEOClerks. International Journal of Information Management, 38(1), 107-116.

Batagelj, V., Doreian, P., Ferligoj, A., & Kejzar, N. (2014). *Understanding large temporal networks and spatial networks: Exploration, pattern searching, visualization and network evolution*. Hoboken: John Wiley & Sons.

Beckett, S. J. (2016). Improved community detection in weighted bipartite networks. *Royal Society open science*, 3(1).

Bhatia, P., Iovleff, S., & Govaert, G. Blockcluster: An R package for model based co-clustering. (2014). https://www.jstatsoft.org/index.php/jss/article/view/v076i09/v76i09.pdf/ Accessed 26 February 2017.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *10*, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

Brier, A., & Hopp, B. (2011). Computer assisted text analysis in the social sciences. *Quality & Quantity*, *45*(1), 103-128. https://doi.org/10.1007/s11135-010-9350-8.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, *39*(3), 510-526. https://doi.org/10.3758/BF03193020.

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior research methods*, *44*(3), 890-907. https://doi.org/10.3758/BF03193020.

Celardo, L. Classifying textual data: a two-way approach. (2017). http://www.diss.uniroma1.it/sites/default/files/allegati/Livia%20Celardo.pdf/ Accessed 25 September 2017.

Choudhury, M., Chatterjee, D., & Mukherjee, A. (2010, August). Global topology of word co-occurrence networks: Beyond the two-regime power-law. Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 162-170). Association for Computational Linguistics.

Chung, W. (2014). BizPro: Extracting and categorizing business intelligence factors from textual news articles. *International Journal of Information Management*, *34*(2), 272-284. https://doi.org/10.1016/j.ijinfomgt.2014.01.001.

Colladon, A. F., & Gloor, P. A. (in press). Measuring the impact of spammers on e-mail and Twitter networks. *International Journal of Information Management*. https://doi.org/10.1016/j.ijinfomgt.2018.09.009.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5), 1-9.

Dagan, I., Marcus, S., & Markovitch, S. (1995). Contextual word similarity and estimation from sparse data. *Computer Speech & Language*, *2*(9), 123-152. https://doi.org/10.1006/csla.1995.0008.

Dhillon, I. S. (2001, August). Co-clustering documents and words using bipartite spectral graph partitioning. Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 269-274). ACM.

Dhillon, I. S., Mallela, S., & Modha, D. S. (2003, August). Information-theoretic co-clustering. Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 89-98). ACM.

Diesner, J. (2012). Uncovering and managing the impact of methodological choices for the computational construction of socio-technical networks from texts. Technical Report CMU-ISR-12-101, Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Pittsburgh, PA.

Diesner, J. & Carley, K. M. (2004). Automap1.2 - extract, analyze, represent, and compare mental models from texts. Technical Report CMU-ISRI-04-100, Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Pittsburgh, PA.

Doreian, P., Batagelj, V., & Ferligoj, A. (2004). Generalized blockmodeling of two-mode network data. *Social networks*, *26*(1), 29-53. https://doi.org/10.1016/j.socnet.2004.01.002.

Du, N., Wang, B., Wu, B., & Wang, Y. (2008, December). Overlapping community detection in bipartite networks. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (Vol. 1, pp. 176-179). IEEE.

Everett, M. G., & Borgatti, S. P. (Forthcoming). Partitioning Multimode Networks. In Doreian, P., Batagelj, V., & Ferligoj, A. (Eds.), *Advances in Network Clustering and Blockmodeling*. Hoboken: John Wiley and Sons.

Everett, M. G., & Borgatti, S. P. (2013). The dual-projection approach for two-mode networks. *Social Networks*, *35*(2), 204-210. https://doi.org/10.1016/j.socnet.2012.05.004.

Fronzetti Colladon, A. (2018). The Semantic Brand Score. Journal of Business Research, 88, 150–160. https://doi.org/10.1016/j.jbusres.2018.03.026.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

Gloor, P., Fronzetti Colladon, A., Giacomelli, G., Saran, T., & Grippa, F. (2017). The impact of virtual mirroring on customer satisfaction. Journal of Business Research, 75, 67–76. https://doi.org/10.1016/j.jbusres.2017.02.010.

Gloor, P. A., Colladon, A. F., Grippa, F., & Giacomelli, G. (2017). Forecasting managerial turnover through e-mail based social network analysis. *Computers in Human Behavior*, *71*, 343-352. https://doi.org/10.1016/j.chb.2017.02.017.

Govaert, G., & Nadif, M. (2003). Clustering with block mixture models. *Pattern Recognition*, *36*(2), 463-473. https://doi.org/10.1016/S0031-3203(02)00074-2.

Grover, P., Kar, A.K., Dwivedi, Y.K. & Janssen, M. (2018). Polarization and acculturation in the 2016 US presidential election: Can Twitter analytics predict changes in voting preferences? Technological Forecasting & Social Change. https://doi.org/10.1016/j.techfore.2018.09.009.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33*(3), 464-472. https://doi.org/10.1016/j.ijinfomgt.2013.01.001.

Hillard, D., Purpura, S., & Wilkerson, J. (2007, April). An active learning framework for classifying political text. Annual Meeting of the Midwest Political Science Association, Chicago, IL.

Hogenraad, R., McKenzie, D. P., & Péladeau, N. (2003). Force and influence in content analysis: The production of new social knowledge. *Quality & Quantity*, *37*(3), 221-238. https://doi.org/10.1023/A:1024401325472.

Iezzi, D. F. (2010). Topic connections and clustering in text mining: an analysis of the JADT network. *LEXICOMETRICA*, *2*, 719-730.

Iezzi, D. F. (2012). Centrality measures for text clustering. *Communications in Statistics-Theory and Methods*, *41*, 3179-3197. https://doi.org/10.1080/03610926.2011.633729.

Iezzi, D. F., Mastrangelo, M., & Sarlo, S. (2012). Text clustering based on centrality measures: an application on job advertisements. Proceedings of the 11th International Conference on Statistical Analysis of Textual Data (pp.515-524).

Iezzi, D. F., & Mastrangelo, M. (2014). Fuzzy c-means for web mining: the Italian tourist forum case. In D. Vicari, A. Okada, G. Ragozini, & C. Weihs (Eds.), *Analysis and Modeling of Complex Data in Behavioral and Social Sciences* (pp. 153-160). Cham: Springer.

Kaushik, K., Mishra, R., Rana, N. P., & Dwivedi, Y. K. (2018). Exploring reviews and review sequences on e-commerce platform: A study of helpful reviews on Amazon. *Journal of Retailing and Consumer Services*, *45*, 21-32. https://doi.org/10.1016/j.jretconser.2018.08.002.

Krippendorff, K. (2004). *Content analysis: an introduction to its methodology*. (2nd ed.). Thousand Oaks: Sage Publications.

Lau, K., Lee, K., & Ho, Y. (2005). Text Mining for the Hotel Industry. *Cornell Hotel and*

*Restaurant Administration Quarterly*, *46*(3), 344–362. https://doi.org/10.1177/0010880405275966.

Liu, H., & Cong, J. (2013). Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin*, *58*(10), 1139-1144. https://doi.org/10.1007/s11434-013-5711-8.

Liu, X., & Murata, T. (2010). Community detection in large-scale bipartite networks. *Transactions of the Japanese Society for Artificial Intelligence*, 25(1), 16-24.

Malrieu, J. P. (1994). Coloured semantic networks for content analysis. *Quality & Quantity*, *28*(1), 55-81. https://doi.org/10.1007/BF01098726.

Melamed, D. (2014). Community structures in bipartite networks: A dual-projection approach. *PloS one*, *9*(5), e97823. https://doi.org/10.1371/journal.pone.0097823.

Nasir, S. (2005). The development, change, and transformation of Management Information systems (MIS): A content analysis of articles published in business and marketing journals. *International Journal of Information Management*, *25*(5), 442-457. https://doi.org/10.1016/j.ijinfomgt.2005.06.003.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, *103*(23), 8577-8582. https://doi.org/10.1073/pnas.0601602103.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*(2), 026113. https://doi.org/10.1103/PhysRevE.69.026113.

Popping, R. (2000). *Computer-assisted text analysis*. Newcastle: Sage.

Rekik, R., Kallel, I., Casillas, J., & Alimi, A. M. (2018). Assessing web sites quality: A systematic literature review by text and association rules mining. International Journal of Information Management, 38(1), 201-216.

Roberts, C. W. (2000). A conceptual framework for quantitative text analysis. *Quality & Quantity*, *34*(3), 259-274. https://doi.org/10.1023/A:1004780007748.

Roberts, C. W. (Ed.). (1997). *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*. Mahwah: Lawrence Erlbaum Associates.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613-620.

Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality & Quantity*, *47*(2), 761-773. https://doi.org/10.1007/s11135-011-9545-7.

Shirdastian, H., Laroche, M., & Richard, M. O. (2017). Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitter. International Journal of Information Management. https://doi.org/10.1016/j.ijinfomgt.2017.09.007.

Singh, J. P., Dwivedi, Y. K., Rana, N. P., Kumar, A., & Kapoor, K. K. (2017). Event classification and location prediction from tweets during disasters. Annals of Operations Research, 1-21. https://doi.org/10.1007/s10479-017-2522-3.

Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., & Roy, P. K. (2017). Predicting the “helpfulness” of online consumer reviews. Journal of Business Research, 70, 346-355.

Sowa, J. F. (1984). *Conceptual structures: Information processing in mind and machine*. Boston: Addison-Wesley.

Sowa, J. F. (Ed.). (2014). *Principles of semantic networks: Explorations in the representation of knowledge*. Burlington: Morgan Kaufmann.

Tan, A. H. (1999). Text Mining: promises and challenges. South East Asia Regional Computer Confederation, Singapore.

Tang, C., Mehl, M. R., Eastlick, M. A., He, W., & Card, N. A. (2016). A longitudinal exploration of the relations between electronic word-of-mouth indicators and firms’ profitability: Findings from the banking industry. *International Journal of Information Management*, *36*(6), 1124-1132. https://doi.org/10.1016/j.ijinfomgt.2016.03.015.

van Atteveldt, W. H. Semantic network analysis: Techniques for extracting, representing, and querying media content. (2008). http://dare.ubvu.vu.nl/bitstream/handle/1871/15964/complete%20dissertation.pdf?sequence=5/ Accessed 01 December 2016.

Van Cuilenburg, J. J., Kleinnijenhuis, J., & De Ridder, J. A. (1988). Artificial intelligence and content analysis. *Quality & Quantity*, *22*(1), 65-97. https://doi.org/10.1007/BF00430638.

Weber, R. P. (1983). Measurement models for content analysis. *Quality & Quantity*, *17*(2), 127-149. https://doi.org/10.1007/BF00143616.

Wixom, B., & Watson, H. (2010). The BI-based organization. *International Journal of Business Intelligence Research (IJBIR)*, *1*(1), 13-28. https://doi.org/10.4018/jbir.2010071702.

Xu, X., Wang, X., Li, Y., & Haghighi, M. (2017). Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors. *International Journal of Information Management*, *37*(6), 673-683. https://doi.org/10.1016/j.ijinfomgt.2017.06.004.