NETWORK TOOLS FOR THE ANALYSIS OF BRAND IMAGE

Agnieszka Stawinoga, Simona Balbi, Germana Scepi

Department of Economics and Statistics, University of Naples "Federico II", Naples, Italy

Abstract. *In Text Mining, large corpora are explored, in order to discover and synthesize their content, in an automatic, time-saving way. If the aim is to understand the similarity of the documents in the corpus, it is interesting to represent them as a network, measuring their proximity in relation to the words they use. Therefore, statistical tools, developed for the analysis of Social Networks, can be applied. In Social Network Analysis, it is often important to introduce information related to the actors (nodes), such as characteristics identifying their belonging to a specific group. In literature, among other proposals, the homophily index E-I has been introduced by Krackhardt and Stern in order to measure the tendency of actors to have relations with actors similar to themselves. Analogously, in the analysis of textual data, better results can be achieved by introducing information related to each document (e.g. characteristics of the author, or when the document has been written). Here we focus our attention on advertisements, and their textual component. In competitive markets there is a complex relation between the "image" of a brand, and the messages of the advertising campaigns over time. In the paper we propose a statistical procedure for analyzing the evolution of the brand image through the different campaigns of a famous brand, namely Coca Cola, on the basis of the E-I index. The validity of our results has been confirmed by means of simulations.*

Keywords: *homophily index, textual data analysis, advertisement campaign*

1. INTRODUCTION

One of the main tasks in Text Mining (TM) consists in exploring large unstructured textual databases, in order to automatically discover and synthesize their content (Aggarwal, Zhai, 2012). In this paper, we adopt a TM viewpoint in order to explore the language and the content of documents belonging to a specified corpus. As the aim is understanding the similarity of the documents, we represent them as a network, measuring their proximity in relation to the words they use. In recent decades there has been an enormous interest in relational data: friendship networks, co-authorship networks of scientists, the world-wide web (Broder et al., 2000), transportation networks and biological networks (Jeong et al., 2000), and so on. Therefore, in the frame of Social Science, a huge quantity of methods for analyzing

social networks has been proposed, under the label of Social Network Analysis (SNA, Wasserman, Faust, 1994). In SNA the characteristics of social actors are understood in terms of patterns, or structures, of ties among them. On the other hand, it is often important to introduce explicitly information related to the actors, as attributes identifying their belonging to a group. In SNA literature, the attitude of individuals to associate with other individuals of the same group, the so called "homophily", has been deeply investigated (Marsden, P. V., 1988; McPherson, M. et al., 2001). Different indexes have been proposed, in order to measure homophily (Krackhardt, Stern, 1988; Everett, Borgatti, 2012). In the analysis of textual data, we state that better results can be achieved by introducing contextual information related to each document (e.g. characteristics of the author, or when the document has been written). In a network of documents, homophily indexes can measure the lexical proximity of documents sharing contextual information (in terms of words in common, and consequently in the topic).

After introducing the basic notation for Network Analysis in paragraph 2, in paragraph 3 we propose our TM strategy, consisting in representing our lexical table, in a one-way network in which documents (nodes) are linked basing on the common vocabulary they share. In order to measure the influence of a contextual categorical variable, the corresponding category is associated to each node. The strength of the relations between and within the documents in the groups identified by the different categories has been computed by means of the relative homophily index *E-I*, proposed by Krackhardt and Stern (1988). In paragraph 4, the effectiveness of our proposal is showed in studying advertising campaigns during time, focusing attention on the textual component. As in competitive markets there is a complex relation between the "image" of a brand, and the temporary advertising campaigns, we apply our proposal for analyzing the evolution of the brand image through the different campaigns of Coca Cola, during the 20th century. Advertisements are labeled with the decade in which they appeared. Based on the *E-I* index, we identify the decades when the general image of the brand is dominant and the periods representing original themes. The result gives a deep insight of the different strategic choices of the brand. The validity of our results has been confirmed by means of simulations. In the last paragraph, future developments of our proposal are presented.

2. NETWORK ANALYSIS: BASIC NOTATIONS

Networks are representations of relational systems. In a network, the nodes represent individuals, and the links (ties or edges) represent a specific relationship among them. The primary object of Network Analysis are the relations among individuals. The most common way to represent a network is by using a graph, which provides a flexible abstraction for describing relationships between discrete objects.

Let graph G be a network. It consists of a set of relationally connected units (actors) and can be represented by $G = (V, E_g)$ composed of a set nodes (actors) $V = (v_1, v_2, \ldots, v_n)$ and a set E_g of edges representing the relationships between the actors. The edge $e_{ij} = (v_i, v_j)$, $e_{ij} \in E_g$ indicates that an actor v_i is linked to an actor v_j. We denote the number of nodes and edges by n and m, respectively.

In a two-mode network $N = (V, U, E_g)$ the set of vertices consists of two disjoint sets of vertices $V = (v_1, v_2, \ldots, v_n)$ and $U = (u_1, u_2, \ldots, u_n)$, and all the links from E_g have one end-vertex in V and the other in U.

A network G can be fully represented by the binary square matrix (adjacency matrix) $X_{VxV} = (x_{ij})$, $i = 1,\ldots, n, j = 1, \ldots, n$, with $x_{ij} = 1$ if $(e_i, e_j) \in E_g$ and 0 otherwise.

A two-mode network N can be fully described by a binary rectangular matrix (an affiliation matrix) $T_{VxU} = (t_{ij})$, $i = 1,\ldots, n, j = 1, \ldots, k$, with $t_{ij} = 1$ if $(v_i, u_j) \in E_g$ and 0 otherwise.

A property F_V can be associated to the nodes, either computed from the network, or input as metadata. A vector $\mathbf{f} = [f_1, f_2, \ldots, f_n]$ consists of the values assumed by F_V for each node.

The density of a graph G is a proportion of the number of edges present in a graph to the maximum possible number of edges in a graph with n vertices.

The degree deg(v) of a node $v \in V$ is the number of edges incident to v, with loops being counted twice. A vertex of degree 0 is an isolated vertex. The degree distribution is the probability distribution of degrees of vertices over the whole graph; so, the degree distribution P(q) of a graph is the fraction of vertices in the graph with degree q.

A component of a graph G is a subgraph in which any two vertices are connected to each other by paths, and which is connected to no additional vertices. A path from $v_1 \in V$ to $v_2 \in V$ is an alternating sequence of vertices and edges, beginning with v_1 and ending with v_2, such that each edge connects its preceding vertex with its succeeding one.

3. OUR PROPOSAL

Let us consider a large corpus C we wish to explore by TM tools. After the usual preprocessing step (Bolasco, 1998), we structure C in a lexical table **T** ($n \times p$), where n is the number of documents and p the number of words. As we are interested in finding documents sharing words, in order to understand the different topics in C, we choose a binary coding: the general element t_{ij} is equal to 1 if the word j occurs at least once in the i-th document, and 0, otherwise (for $i = 1$ to n and $j = 1$ to p). The (n, n) matrix **X** is computed as **X=TT'**, which represents the adjacency weighted matrix for the documents. The weights are given by the numbers of the common words. To avoid weak relations and to obtain less sparse matrix, we normalize **X**.

Therefore, **X** is transformed into the symmetric matrix **S** ($n \times n$) where the value of cell s_{ik} is the similarity of the documents i and k, measured by the Jaccard index:

$$s_{ik} = \frac{x_{ik}}{x_{ii} + x_{kk} - x_{ik}}, \tag{1}$$

where the element $x(i,k)$ represents the number of words common to the documents i and k; $x(i,i)$ is the number of words in the document i.

Basing on the actual distribution of the Jaccard index, we fix a threshold (st) as the point after which the density of the network becomes close to zero (an empty network, not interesting in practical point of view). We dichotomized the matrix **S**, as follows: for each s_{ik} with the value higher than st we set the value of the element a_{ik} equal to 1 and 0 otherwise. We obtain a binary adjacency matrix **A** ($n \times n$), representing the network of relations existing among the documents. It is often interesting introducing contextual information (meta-data) on the documents (e.g. characteristics of the author, or when the document has been written). In a content analysis frame, it is important to comprehend if this information can give an insight in understanding the different topics in the corpus. In a network analysis approach, we consider this information as an attribute associated to the network nodes, defined by a categorical variable H with h categories, stored in an n-dimensional vector **h**, partitioning the network into h mutually exclusive groups.

In order to measure the influence of the attribute on the relations among the documents, we study the so called homophily in the network, i.e. the attitude of nodes to associate with other nodes presenting the same category. We measure the strength of the relations between and within the elements in the groups described by H using the relative homophily index (Krackhardt, Stern, 1988):

$$(E-I)_{index} = \frac{E-I}{E+I'} \tag{2}$$

where E is the number of external ties (ties between documents presenting different categories of H) and I is the number of internal ties (ties between documents presenting the same value for H).

In our proposal, the $(E\text{-}I)_{index}$ measures the tendency for documents to share words with documents similar with respect to H. It ranges from -1 (all ties are internal, showing homophily) to +1 (all ties are external, showing heterophily), but for a given network density and group sizes, its range may be restricted. A permutation test is performed in order to check whether the index is significantly different from the value, we would expect by chance (Everett, Borgatti, 2012). Krackhardt and Stern (1988) propose the index to the network as a whole. Furthermore, the value of the index can be calculated for each group and each individual, as in UCINET (Borgatti et al., 2002).

In order to investigate the strength of homophily in the different groups, the following ratio has been proposed (D'Amore, et al., 2013):

$$(E-I)_{ratio} = \frac{(E-I)_{actual}+1}{(E+I)_{\exp ected}+1'} \tag{3}$$

where $(E\text{-}I)_{actual}$ is the observed value, and $(E\text{-}I)_{expected}$ is obtained by the following simulation procedure.

We keep the observed adjacency matrix and generate a large number of random permutations of the elements of the vector **h**, which consists of the values assumed by the variable H for each node. We compute the value of $(E\text{-}I)_{expected}$ as the average value of simulated $(E\text{-}I)_{index}$ calculated on the permutated networks.

The $(E\text{-}I)_{ratio}$ has 0, as minimum value. The maximum depends on the network structure. In details, the value of the numerator varies from 0 to 2, according to the definition of $(E\text{-}I)_{index}$. The denominator ($(E\text{-}I)_{expected}$ + 1) assumes positive values, as $(E–I)_{expected}$ is an average of many $(E\text{-}I)_{index}$ values. The denominator can be equal to 0 only when the value of $(E–I)_{expected}$ is equal to -1. It happens only when all the values composing the sum are equal to -1 and this is possible only in the case in which all the nodes present the same category of H (and this is not realistic). Table 1 gives the details of the proposed strategy after a usual pre-processing step performed on the documents of a large corpus C.

Table 1: The steps of our proposal

INPUT: a lexical table **T** (*n*, *p*), an *n*-dimensional vector **h** with contextual information *H,* the number of permutations *n_permut*

1) computation of the adjacency weighted matrix for the documents, **X=TT'**;
2) transformation of the matrix **X** in the similarity matrix **S** with general element S_{ik} (Jaccard index);
3) representation of the actual distributions of the Jaccard index and detection of the threshold (*st*);
4) computation of the matrix **A** as follows: $a_{ik} = \begin{cases} 1 & \forall s_{ik} \geq st \\ 0 & otherwise \end{cases}$
5) computation of $(E\text{-}I)_{actual}$ on **A** for the whole network and for the groups defined by the variable *H*
6) computation of $(E\text{–}I)_{expected}$, as follows: for i = 1 to *n_permut*

6.1 computation of $\mathbf{h_i}$ by random permutation of the elements of the vector **h**;

6.2 association of $\mathbf{h_i}$ to **A**

6.3 computation of $(E\text{-}I)_{index}$ for **A**

7) computation of the average value of the *n_permut* simulated $(E\text{-}I)_{index}$ values;
8) calculation of $(E\text{-}I)_{ratio}$

OUTPUT: $(E\text{-}I)_{actual}$, $(E\text{-}I)_{expected}$, $(E\text{-}I)_{ratio}$

4. COCA COLA DURING THE 20TH CENTURY

As in competitive markets there is a complex relation between the "image" of a brand and the advertising campaigns over time, we apply our proposal for analyzing the evolution of the brand image of Coca Cola, during the 20th century (Figure 1), by exploring the textual components. On the Web, we collect a set of 164 American advertisements, distributed as illustrated in Table 2.

It is worth noting that this distribution depends both on the availability on the Web of the advertisements and on the different role of the textual component during the century. At the beginning, there were very long texts and little by little the texts become shorter until the '80s and the '90s, with advertisements based most of all on images accompanied only by slogans. For a brief history of Coca Cola Advertising Slogans, see Coca Cola Journey (2012).

Figure 1: Images of some collected advertisements

Advertisements have been labeled with the decade related to the period in which they appeared (in ascending order). The "Decade 1" actually refers to a period comprising two decades; in this case, a longer period has been chosen because of the small number of advertisements present in this particular historical period. The decade is the contextual information we consider for understanding if the brand identity dominates the messages or occasional themes influence the advertisement wording. Based on the $(E–I)_{index}$, we identify the decades when the identity of the brand is dominant and the periods representing original themes instead. The result gives a deep insight of the different strategic choices of the brand over the century.

Table 2: Distribution of the collected advertisements

Label	Years	Nr of adv
Decade1	1890 – 1909	6
Decade2	1910 – 1919	5
Decade3	1920 – 1929	17
Decade4	1930 – 1939	30
Decade5	1940 – 1949	34
Decade6	1950 – 1959	16
Decade7	1960 – 1969	18
Decade8	1970 – 1979	16
Decade9	1980 – 1989	6
Decade10	1990 – 1999	9

In order to understand and extract the information contained in our set of advertisements, we transform textual (unstructured) data in a lexical matrix, which can be analyzed with statistical tools. This transformation concerns the selection of the terms, which are able to represent the semantic structure of the advertisements. In literature, this process is well known as text pre-processing. A unique definition of the pre-processing step does not exist. According with the aim of the analysis, we create an ad hoc strategy. First, the corpus is normalized, and cleaned by usual English stop words, showing 1,267 types. After lemmatization, we obtain 1,060 forms which are re-examined by removing hapax (terms with occurrence equal to 1) and the most frequent terms (*Coca Cola*, *Coke*, *refresh*, *pause*, *ice-cold*, *taste*, *delicious*, *drink*). After the pre-processing procedures, the corpus consists of 3,276 occurrences and 460 graphical forms. As the final step, we create a lexical table **T** document-by-terms of dimension (164×460), in binary coding. The previous steps are performed by using a function written in R based on two packages "tm" (Feinerer, 2014) and "koRpus" (Michalke, 2014).

Subsequently, we build the adjacency matrix **A** (164×164), which represents the relations among advertisements, according to the threshold value of the Jaccard index ($st = 0.1$). A vector **h**, whose general element h_i ($i = 1, \ldots, 164$) indicates the decade of the *i*-th advertisement, is associated to **A**.

The observed network is characterized by the density equal to 0.071. It consists of a huge component of 162 nodes and two isolate nodes indicating two advertisements which do not share terms with any advertisements. According to the degree centrality index, which detects the number of adjacent edges of each node, we observe that on average, the advertisements have words in common with almost 12 other advertisements (the mean value of the degree centrality equals 11.63 (6.77)).

With the aim of identifying which decades are connected with the general image of the brand and which ones represent original temporal themes, we compute $(E\text{-}I)_{actual}$ for the whole network and for each decade separately. High positive values of the index identify decades connected with the image as they use words present in all the considered periods (high heterogeneity). High negative values identify peculiar campaign ideas characterized by words related to specific themes (war in the '40s, domestic consumption in the '50s). We observe that the proximity of the brand image increases if the ratio of heterogeneous linkages (links between advertisements of different decades) increases. We have performed the randomization test based on 5000 permutation. We have maintained fixed the density and degree distribution of the network. We have randomized the assignment of advertisements to the decades, by reshuffling the *h* elements. For each run of the procedure, the value of the seed has been changed by fixing the random number generator to avoid

repetition of runs. We have obtained $(E\text{-}I)_{expected}$ by means of average of the 5000 $(E\text{-}I)_{index}$ values and, subsequently we have computed the values of $(E\text{-}I)_{ratio}$. All the values of $(E\text{-}I)_{ratio}$ lower than 1 indicate a signal of lexical proximity of advertisements based on the decade in which they appear. All the values higher than 1 indicate linguistic distance above and beyond what would be expected if the vocabularies of advertisements do not depend on the decade of their appearance. We have calculated the number of times the random test has a value greater than or equal to the observed. For all the cases, we have obtained that the probability that it occurs is equal 1, which means that $(E\text{-}I)_{actual}$ value is always lower than $(E\text{-}I)_{index}$ for each permutation run. This result indicates that the structure of the observed network is characterized by higher homophily than in the case in which the decades do not influence the vocabularies of the advertisements. For investigating the strength of homophily in the different decades, we have computed the value of $(E\text{-}I)_{ratio}$. Finally, in order to identify the most typical words of the decades, we have visualized the two-mode network of the advertisements and the words.

Table 3 shows the results of the permutation test for both the whole network and the group levels according to the decade. The '10s show the lowest values of $(E\text{-}I)_{actual}$ (-0.6) and $(E\text{-}I)_{ratio}$ (0.21). This indicates the advertisements of these years are based on peculiar campaign ideas using specific words. For Coca-Cola it was the moment of highlighting the brand image and its origin because the market had become national. By analyzing the two mode network, we can notice (Figure 2) that the advertisements create only one component and the terms they have in common are *atlanta, genuine, nickname, name, substitution, full, demand, encourage*. Coca Cola tries to consolidate the uniqueness of its product.

Table 3: E–I index for whole network and group levels

		$(E\text{-}I)_{actual}$	$(E\text{-}I)_{expected}$	$(E\text{-}I)_{ratio}$
Whole network		0.15	0.73	0.67 (prox)
Group levels	Decade1	0.26	0.94	0.65 (prox)
	Decade2	-0.60	0.95	0.21 (prox)
	Decade3	-0.02	0.81	0.50 (prox)
	Decade4	0.10	0.65	0.57 (prox)
	Decade5	0.24	0.53	0.64 (prox)
	Decade6	0.57	0.82	0.81 (prox)
	Decade7	0.17	0.78	0.60 (prox)
	Decade8	-0.06	0.82	0.48 (prox)
	Decade9	0.75	0.94	0.90 (prox)
	Decade10	0.20	0.91	0.62 (prox)

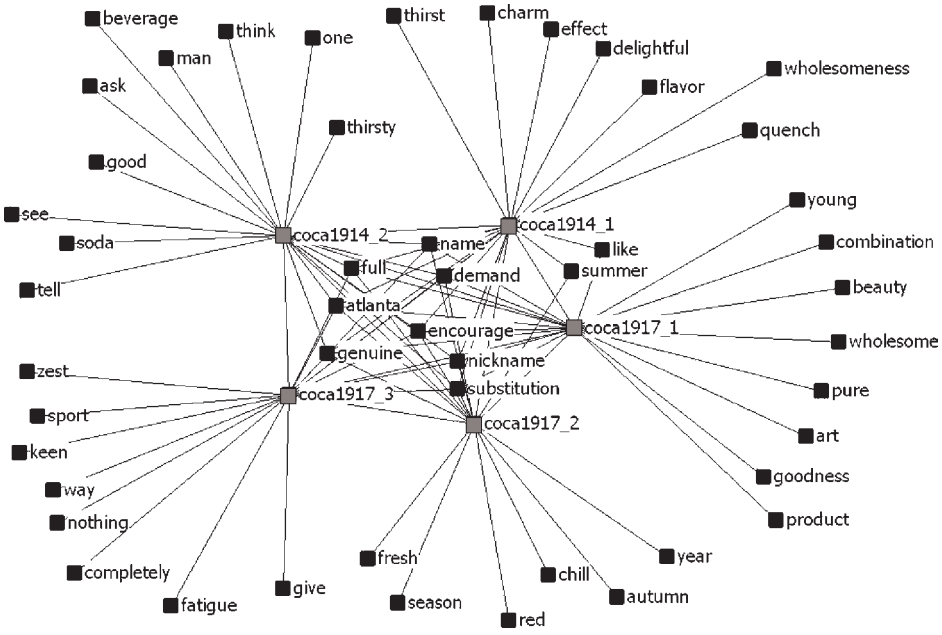

Figure 2: Component of the advertisements from of the '10s

The '80s (Decade 9) are interesting because they are characterized by high values of $(E\text{-}I)_{actual}$ (0.75) and $(E\text{-}I)_{ratio}$ (0.90). Therefore, advertisements are strictly connected with the image of Coca Cola (*refreshment*, *classic*, *summer*, *soft*). It is worth noting that the advertisements of those years are very short. Consequently, there are few internal links. They are usually slogans with few words referring to the brand image of Coca Cola (Figure 3).

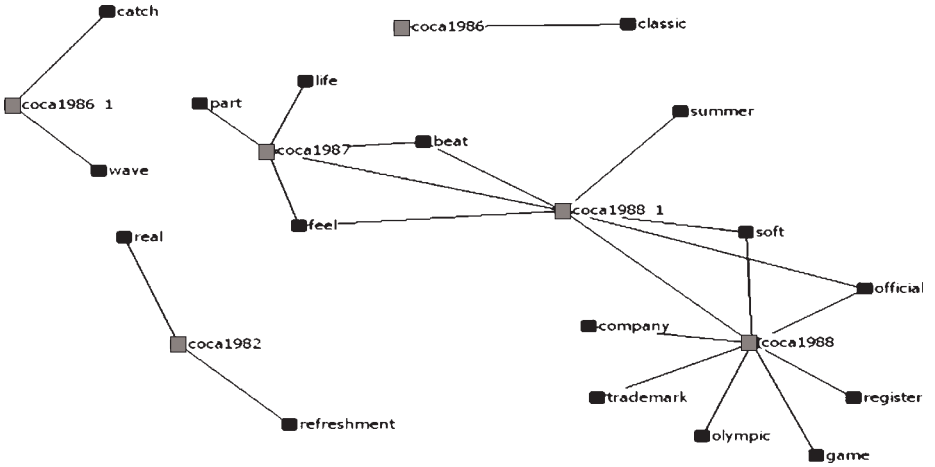

Figure 3: Components of the advertisements of the '80s

In the case of the '50s (Decade 6), we observe a high value of '$(E\text{-}I)_{actual}$ (0.57) and '$(E\text{-}I)_{ratio}$ (0.81). The advertisements of this decade are longer and treat different topics together with general themes of the brand. Therefore, the sub-graph is divided into three components (Figure 4 and Figure 5). The peculiar topic is the domestic consumption (*hospitality*, *home*, *grocer*, *family*). The most typical Coca Cola words are *flavour*, *sparkling*, *real*, *favourite*, *delight*.

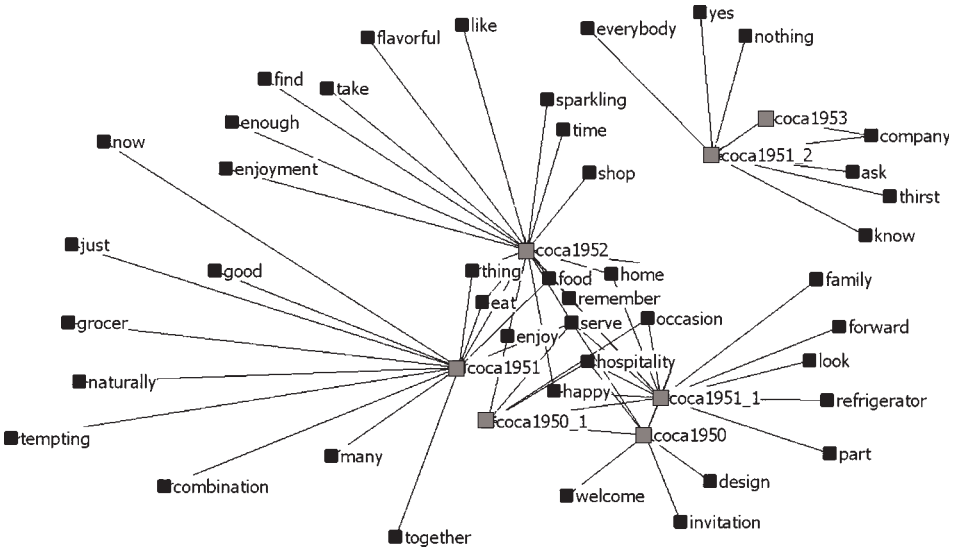

Figure 4: Two of three components of the advertisements of the '50s

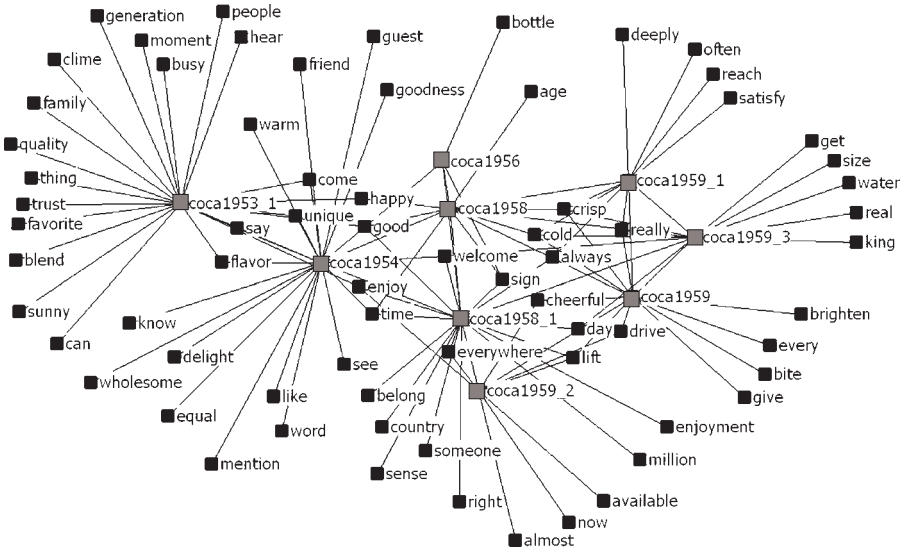

Figure 5: Third component of the advertisements of the '50s

Back to Table 3, we observe that that the homophily, as a measure of linguistic proximity, is high in the analyzed phenomenon. The values of $(E\text{-}I)_{ratio}$ is always less than one, also in the decades more characterized by words common to the general image of Coca Cola (the '80s and the '50s). This is confirmed also for the whole network ($(E\text{-}I)_{ratio}$ equal to 0.67).

We can conclude that against the perception of a strong image of Coca Cola, given by the permanent visual component (the *red-white* colors, the font), the textual component of the advertisements is strongly influenced by topical events, during the 20th Century.

5. COMMENTS AND FUTURE DEVELOPMENTS

In this paper, we propose a statistical procedure for analyzing the evolution of the brand image through the different campaigns of a famous brand, basing on network analysis tools and, in particular, on the homophily index whose properties we intend to study in detail in the future.

From a methodological viewpoint, we believe that the representation of a corpus by the network of its documents (taking into account the terms they share) is very promising, beyond the traditional Textual Network Analysis approach (Carley, 1997; Popping, 2000) more focused on networks of words and concepts. Methods developed in the frame of Social Network Analysis can offer new opportunities in exploring large unstructured databases, both in an Information Retrieval and in a broader Text Mining context. In the specific question of analyzing advertisement campaigns, we are proposing strategies, not only for understanding the communication strategy of a brand, but also for comparing it with competitors, as in the case of Coca Cola and Pepsi Cola.

REFERENCES

Aggarwal, C.C. and Zhai, C. (2012). *Mining Text Data*, Springer.

Batagelj, V., Mrvar, A. and Zaversnik, M. (2002). Network analysis of texts. Paper online: http://nl.ijs.si/isjt02/zbornik/sdjt02-24bbatagelj.pdf

Bolasco, S. (1998). Metadata and strategies of textual data analysis: problems and instruments, in Hayashi, C. et al. (eds.), *Data Science, Classification and related methods.* Springer Verlag: 468-479.

Borgatti, S.P., Everett, M.G. and Freeman, L.C., (2002). *Ucinet for Windows: Software for Social Network Analysis*. Analytic Technologies, Harvard, MA.

Broder A., Kumar R., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomkins A. and Wiener J. (2000). Graph structure in the Web: experiments and models. *Computer Networks*, 33: 309–320.

Carley, K.M. (1997). Network Text Analysis: the network position of concepts. In Carl W. Roberts (Ed.), *Text analysis for the social sciences*. Mahwah, NJ: Lawrence Erlbaum Associates: 79-102.

CocaCola Journey (2012). Paper online: http://www.coca-colacompany.com/stories/coke-lore-slogans.

D'Amore, R., Iorio R., Labory S. and Stawinoga A. (2013). Research collaboration networks in biotechnology: exploring the trade-off between institutional and geographic distance. *Industry and Innovation*, 20(3): 261-276.

Everett, M.G. and Borgatti, S.B. (2012). Categorical attribute based centrality: E–I and G–F centrality. *Social Networks*, 34(4): 562–569.

Feinerer, I. (2014). tm: Text Mining Package (Version 0.6).

Jeong H., Tombor B., Albert R., Oltavi, Z.N. and Barabási, A.L. (2000). The large - scale organization of metabolic networks. *Nature*, 407: 651-654.

Krackhardt, D. and Stern, N. (1988). Informal networks and organizational crisis: an experimental simulation. *Social Psychology Quarterly*, 51(2): 123-140.

Michalke, M. (2014). koRpus: An R Package for Text Analysis (Version 0.05-5).

Marsden, P.V. (1988). Homogeneity in confiding relations. *Social Networks*, 10: 57-76.

McPherson, M., Smith-Lovin, L. and Cook. J.M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27: 415-44.

Popping, R. (2000). *Computer Assisted Text Analysis*, Sage.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: methods and applications*. Cambridge University Press.