# Extracting meta-information by using Network Analysis tools

Agnieszka Stawinoga, Maria Spano and Nicole Triunfo

**Abstract** This paper has been developed in the frame of the European project BLUE-ETS (Economic and Trade Statistics). In order to obtain business information by documentary repositories, we refer to documents produced with non statistical aims. The use of textual sources is still viewed as too problematic, because of the complexity and the expensiveness of the pre-processing procedures.
The aim of this paper is to create financial-economic meta-information exploring relational structures existing among terms of the management commentary of companies listed on Italian market.

**Key words:** network analysis, statistical language resources, BLUE-ETS project

## 1 Introduction

This work has been developed in the frame of the European project BLUE-ETS, acronym for BLUE Enterprise and Trade Statistics (www.blue-ets.istat.it), funded by the European Commission (7th Framework Programme). Our peculiar task in BLUE-ETS consists in proposing new tools for collecting and analyzing data. In order to reduce response burden and, at the same time, to collect cheaper and (better) quality data, we refer to secondary sources, produced with non statistical aims. The use of secondary sources, typical of data and text mining, is an opportunity not sufficiently explored by National Statistical Institutes. NSIs aim at collecting and representing information in a usable and easy-readable way. The use of textual

Agnieszka Stawinoga
University of Naples Federico II, e-mail: agnieszka.stawinoga@unina.it

Maria Spano
University of Naples Federico II e-mail: maria.spano@unina.it

Nicole Triunfo
University of Naples Federico II, e-mail: nicole.triunfo@unina.it

data has been still viewed as too problematic, because of the complexity and the expensiveness of the pre-processing procedures and often for the lack of suitable analytical tools. In order to focus our attention on enterprise statistics we propose to extract and to analyze data by mining into the management commentaries attached to the annual reports. In this work we pay attention to the problems related to the pre-processing procedures, mainly concerning with semantic tagging. In order to perform the semantic tagging it is necessary to have a language resource appropriate for the subject of analysis. In this paper we propose a semi-automatic strategy based on network analysis tools for creating financial-economic meta-information.

## 2 Pre-processing: State of Art

In order to understand and extract the information contained in a set of documents, it is necessary to transform textual (unstructured) data in a lexical matrix which can be analyzed with statistical tools. In the literature this process is well-known as text pre-processing. A unique definition of the pre-processing steps does not exist. According with the aim of the analysis, the researcher creates an ad hoc strategy for extracting the significant information from the documents. It is obvious that the researchers' choices affect completely the results of the analysis. In order to perform an automatic text analysis, the first step consists in choosing the unit of the analysis. On the one hand, the formalists [2] consider the graphical form / type (sequence of characters delimited by two separators) as the unit of analysis. They carry out the statistical analysis regardless of the meaning of the graphical forms, which are considered language independent. On the other hand, the computational linguists consider the lemma as the unit of the analysis [5]. Electronic dictionaries, frequency lexica and automatic normalization are the practical tools adopted according with this language dependent approach. In the field of textual data analysis, Bolasco [3] considers a mixed language dependent unit (graphical form / lemma / multiword expression) named "textual form". Once the unit of analysis is chosen, the pre-processing can be summarised in the following steps:

1. cleaning of the text (definition of alphabet characters / separators);
2. normalization (recognition of particular entities such as dates, acronyms, abbreviations);
3. text annotation (introduction of meta-information by grammatical and semantic tagging, lemmatization, etc.) [4].

## 3 Methodology

In order to detect meaningful communities of closely related terms Balbi and Stawinoga [1] proposed a new Text Mining strategy for dimensionality reduction by the use of Network Analysis tools. The strategy has been applied to the management

commentary of the world leader of eyewears, Luxottica, because this company is listed on both U.S. and Italian markets and the U.S. law thoroughly explains all information that the management commentary must include (differently to the Italian law). The data matrix used for the analysis is the lexical table $\mathbf{T}$ $(n \times p)$ which indicates the occurrences of the $p$ terms in $n$ parts of a corpus. The matrix T can be easily transformed into an 1-mode co-occurrence matrix $\mathbf{W}$ $(p \times p)$ which represents the relational system of the selected terms. To avoid weak relations and to obtain less sparse matrix the authors proposed to use the strength of associations among terms instead of the simply co-occurrence frequency. According to a predefined threshold based on the actual distributions of Jaccard index, the matrix of similarities $\mathbf{S}$ $(p \times p)$ is dichotomized to obtain network representation of the relations existing among the terms. In the following step, various network analysis tools are used for identifying meaningful structures of terms.

In this work we propose to extend this strategy to the analysis of more than one management commentary, in order to identify meta-information (statistical language resources) for the financial-economic field. The work focuses on the section devoted to the results of operations, which is common for all management commentaries. We go to identify topics which are mostly treated by all companies in their annual reports. In this case, the rows of the lexical table $\mathbf{T}$ represent the different companies and the columns are the terms. We analyze a sample of 50 Italian listed companies, extracted using a technique of quota sampling to ensure compliance with the composition of the sectors in which the firms in the Italian stock exchange (Borsa Italiana s.p.a.) are classified. Our reference year is 2010.

The network analysis tools we propose to use are the following. Firstly, different components of the network (maximal connected sub-graphs) are extracted. From a textual point of view, the different connected components give the possibility to individuate different topics. Figure 1 illustrates examples of some components extracted from the analyzed network. For instance, the pink one consists of 3 terms (nodes) which specify the methods (LIFO-FIFO) to calculate the value of inventories reported in the financial statements. When a component of the network is too

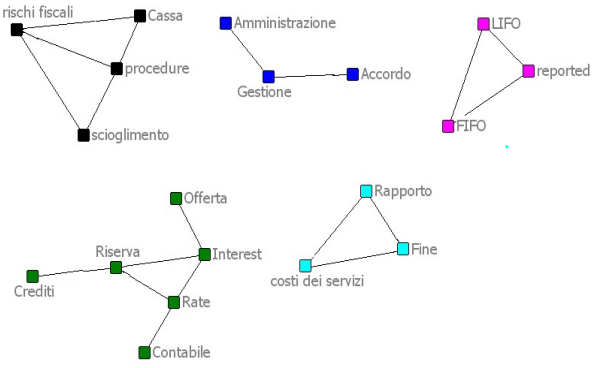


**Fig. 1** Examples of components extracted from the analyzed network

big to identify a single concept (e.g. the main component), we propose to calculate betweenness centrality [6] to individuate the most influential nodes connecting different communities within this subgraph of the network. In terms of text analysis, the higher the value, more often the term links various contexts within the text. Once the nodes with the highest value of betweenness centrality are selected, their ego networks (an ego network is a network of a focal node (ego), all actors connected to that node (alters), and all the connections among those other nodes [7]) are investigated to identify and to describe different topics. The proposed strategy allows us to pass from elementary data (terms) to higher order data (context, topics). It gives the possibility to select a subset of relevant terms, which illustrate important topics characterizing the corpus.

## 4 Conclusions and future work

In this paper we propose the use of Network Analysis tools for the construction of statistical language resources. The proposed strategy makes it possible to identify terms, which characterize the documents under analysis and to carry out the problems related to the pre-processing. Further developments of the research will be devoted to apply the proposed tool for improving the quality of the process of documents categorization.

**Acknowledgements** This work is financially supported by the European Project BLUE-ETS.

## References

1. Balbi, S., Stawinoga, A.: The use of Network Analysis tools for dimensionality reduction in Text Mining, SLDS 2012, Florence, Italy (2012) https://www.docenti.unina.it/ricerca/visualizzaAttivitaRicerca.do?idDocente=53494d4f4e4142414c4249424c42534d4e35384c35394638333944&nomeDocente=SIMONA&cognomeDocente=BALBI
2. Benzécri, J. P.: Pratique de l'Analyse Des Données, Linguistique e Lexicologie, Dunod, Paris (1981)
3. Bolasco, S.: Sur différentes stratégies dans une analyse des forms textuelles: une expérimentation à partir de données d'enquête, In: M. Bécue, L. Lebart, N. Rajadell (eds.) JADT 1990, UPC, Barcellona, 69-88 (1990)
4. Bolasco, S.: Meta-data and strategies of textual data analysis: problems and instruments, In: Hayashi et al. (eds.) Data Science, Classification and related methods, (proceedings V IFCS - Kobe, 1996), Springer-Verlag Tokio, 468-479 (1998)
5. De Mauro, T.: I vocabolari ieri e oggi, In: Il vocabolario del 2000, a cura di IBM Italia, Roma (1989)
6. Freeman, L. C.: Centrality in Social Networks Conceptual Clarification, Social Networks, 1, 215-239 (1979)
7. Hanneman, R., Riddle, M.: Introduction to social network methods. Riverside, CA: University of California, Riverside (2005) http://faculty.ucr.edu/ hanneman/