Multivariate Techniques to define Semantic Spaces in Social Networks

Tecniche multivariate per la definizione di reti sociali in spazi semantic

Giuseppe Giordano¹, Michelangelo Misuraca², Maria Prosperina Vitale¹

¹ Dipartimento di Scienze Economiche e Statistiche, Università di Salerno
² Dipartimento di Economia e Statistica, Università della Calabria
e-mail: michelangelo.misuraca@unical.it

Keywords: correspondence analysis, graph, lexical table, textual data

1. Introduction

Multidimensional Data Analysis (MDA) and Social Network Analysis (SNA) represent two different approaches to deal with peculiar data structures. On one hand, the techniques of MDA focus on statistical units (individual characteristics) to obtain factorial syntheses and clusters. On the other hand, SNA focuses on ties among interacting units (dyad, triad and subgroups) to describe the patterns of different social relationships (Wasserman and Faust, 1994).

The classical Multidimensional Data Analysis approach on textual data, proposed by Lebart et al. (1998), is here considered in order to discover and extract knowledge from documents collections. In this way, the latent semantic structures are analyzed and the resulting lexical profiles are graphically explored through factorial map. The goal is to visualize the relationships among several documents and among words (Balbi and Misuraca, 2005).

The data structure can be described through a Network (e.g. web forum), where the textual data are collected on a set of actors (nodes) and it is used, in the next, either to aid the interpretation of the network (Giordano and Vitale, 2006) or as an ex-post Network (e.g. text network). In the latter case, Textual Data Analysis (TDA) allows to discover latent associations among documents useful to define new network ties.

The aim of the present paper is to bring together the theoretical frameworks of TDA and SNA in order to define an ex-post Network structure derived from a semantic space in which the documents can be seen as nodes and the ties are set according to the shared vocabulary.

Several ways to derive networks from textual data are discussed in the SNA literature (e.g. Popping, 2000; Polanco and San Juan, 2006). An overview of possible applications of SNA in the text analysis is in Batagelj et al. (2002).

In this paper, the main purpose is to define an adjacency matrix where the ties are derived by a distance measure according to the factorial approach. The proposed strategy is able to:

- define a textual data set by performing a documents pre-treatment (Normalization, Lemmatization and Filtering);
- summarize the textual data through Correspondence Analysis and Clustering;

(*) Work supported by 2007 ex-60% MIUR grants coordinated by M. Misuraca, G. Giordano, M.P. Vitale
derive an adjacency matrix from the distance matrix;
• generate an ex-post Network structure;
• explore and describe the Network through SNA and textual data information.

Let \( F(n \times k) \) be the lexical table built by juxtaposing the \( n \) document-vectors obtained by the corpus numerization, we look for few factorial axes derived by applying to the matrix \( F \) the Singular Value Decomposition (SVD):

\[
F = U \Lambda V^T
\]

\[
U^T D_n^{-1} U = V^T D_k^{-1} V = I
\]

where \( D_n \) and \( D_k \) are the weights and metric (and viceversa) of the subspaces in which the \( n \) documents and the \( k \) words are represented; \( U, V \) and \( \Lambda \) are respectively the left and right eigenvectors and the singular values of \( F \).

The factorial coordinates of the correspondence analysis allow to define a distance matrix of documents. A tie is then defined according to a cut-off threshold of the distribution of this distance measure.

The proposed strategy can be applied in different research fields. For instance, to explore the scientific collaboration worked out by researchers, involved in different disciplinary areas, starting from the research co-interests network; to deal with open-ended questions in the survey research; to evaluate the communication flows in the e-learning environment. In all these frameworks several documents can be gathered both to enrich the interpretation of the phenomenon under study and to define proximities in the network structures.

References


