The Ideal Candidate. Analysis of Professional Competences through Text Mining of Job Offers

Emilio Di Meglio, Maria Gabriella Grassia, Michelangelo Misuraca[1]

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

Summary. The aim of this paper is to propose analytical tools for identifying peculiar aspects of the job market for graduates. The main objective is to reduce the complexity of the phenomenon, both on the variable side, by transforming the collected information into latent factors, and on the unit side, by classifying observations. We propose a strategy for dealing with data that have different source and nature. The dependence structure is investigated to identify potential evolutionary paths. Moreover, symbolic objects and their graphical representation are used for identifying the peculiar characteristics required by companies operating in different economic sectors.
Keywords: Text mining; Association rules; Factor analysis; Symbolic objects; Zoom-star.

1. Text mining

The huge quantity of *on-line* documents and companies' data warehouses makes necessary tools and methods for their analysis (Manning & Schütze, 2001). *Text mining* is a way to unveil the information within verbatim documents, i.e. written in natural language, by using statistical, linguistic and information technology methods. The applications of text mining are increasing at a fast rate: search engines, email filtering and automatic delivery, market analysis based on reports and papers, automatic document classification according to different queries and criteria, and case-based reasoning.

---

[1] This paper is the result of the joint research of the three authors. However, M. Misuraca was responsible for the final editing of Sections 2 and 5, whereas M.G. Grassia was responsible for Section 4 and E. Di Meglio for Section 3.

Multivariate statistical analysis gives good results in terms of synthesis and graphical representation of the information discovered. These techniques analyse the association structure in documents and create a knowledge base of the different concepts in the text.

This knowledge base can be used in different applications. For example, we can represent proximities and oppositions of different concepts on a graphical display, automatically classify documents according to some concepts and extract the information by querying the obtained index.

To analyse real situations, we can define some rules that point out situations and behaviours of the objects detectable not in an intuitive way, but only through a deep analysis of the databases using formalized methodologies. Such behavioural rules are known as Association Rules (Agrawal *et al*., 1993).

Rules are defined as binary attributes (presence/absence,) and, if necessary, through a transformation of the data. A rule is made of a precedent and a consequent part; at the same time, it is possible to identify two distinct parts in the information contributed by a rule, called *support* and *confidence*. Support is the association strength between the considered items, confidence is the logical dependence strength expressed by the rule. The identified rules are reduced with *ad hoc* algorithms for analysing only the meaningful information.

The first order observations are generally described by classical data, while the second order observations, because of their conceptual complexity, need the use of more structured data, such as the symbolic data. The symbolic data analysis consists of a first step of collection and organisation of simple data, in a Knowledge Discovery in Database (KDD) framework.

Then, new concepts are defined in terms of complex data and analysed with statistical techniques. A *concept* is characterized by a set of properties apt to define its description, while the classes of observations that satisfy these properties, known as *objects*, represent its extent.

The objects can be created in several ways. For example, in a multidimensional dataset the categories of a variable can represent concepts to be described with the values assumed by other related variables. On the contrary, if we consider a relational database, the descriptors of the objects can be extracted with a query that expresses the properties of a set of units whose description implies the union of several relations.

Moreover, the objects are derived from the description of the classes obtained from a classification technique. In this way, it is possible to reduce considerably the first order observations. The objects are called *native* if they are the results of an expert knowledge on the phenomenon.

In this paper, we propose the joint use of multidimensional analysis techniques together with association rule building and symbolic data analysis. The aim is designing new text mining strategies, resulting in finding patterns and regularities in on-line job offer databases (Section 3), in organising the data in higher order structures and visualising them with graphic tools (Section 4), and in graphical information syntheses (Section 5).