Session: Informetric and Text based Procedures (Ohly/Stempfhuber)

Mining document contents in order to analyse a scientific domain

Josiane Mothe, Bernard Dousset, Désiré Kompaore
Institut de Recherche en Informatique de Toulouse, 118 Route de Narbonne, F-31062 Toulouse CEDEX 04, France
mothe@irit.fr

Large-scale analysis to understand the state of the art and the evolution of scientific communities and research topics become possible thanks to the availability of large sources of scientific literature. Number of methods are used to achieve this type of analysis. Co-citation analysis is among the most popular as it allows many different types of highly informative analysis to discovery the structure of a domain. A co-citation can be extracted when two references appear in the same published paper. When Journal co-citation is analysed, information on how two journals are co-cited and the relationships among journals can be extracted. On the other hand, co-authoring analysis can also be used to identify groups of scientists and their inner structure. Free text analysis provides complementary information. Mining the document content and combining it with other meta data such as author, affiliation, date of publication informs on the topics of interest and their evolution. Such domain mapping is possible when using various data analysis methods. Classification and factorial analysis are among the ones we use when analysing a domain.
In this paper we will present the approach we developed as well as the system based on this approach (named Tétralogie). It is composed of several modules that implement a knowledge discovery process. Such a process begins by selecting data or documents from one or several sources on a targeted domain. Then the documents are pre-treated in order to extract the useful information (meta data and concepts from the free text). This information is then mined using different data analysis methods including Hierarchical Agglomerative Clustering and Correspondence Factorial Analysis in order to find hidden information or knowledge. The results are presented under the form of multi-dimensional and interactive views. In this paper, the overall process will be presented through a case study.

Keywords: information mining, data analysis, domain knowledge