Co-word Analysis Revisited: Modelling Co-word Clusters in Terms of Graph Theory

Xavier Polanco

*xavier.polanco@*inist.fr
Unité de Recherche et Innovation, Institut de l'Information Scientifique et Technique, Centre National de la Recherche Scientifique, 1 allée du Parc de Brabois, 54514 Vandoeuvre-lès-Nancy, France.

Introduction

The idea that graph theory would be useful in co-word analysis was suggested to us by social network analysis, where both graph theory and matrix operations have been widely used (Wasserman & Faust, 1999; Degenne & Forsé, 2000). In co-word analysis tradition, Courtial (1986) had started to explain the co-word clusters in terms of graphs. However, he did not take this idea to its logical next step, which we will attempt here. We distinguish two analytical units: clusters and the network of clusters. In the network each cluster represents a node. In addition, we will consider each cluster as a "super vertex" (according to the "reduced model approach" in Everett & Borgatti, 1999)

Co-word Analysis Components

Recall the main components of the standard co-word analysis (cf. Callon et al. 1993, c. VII). The input dataset in co-word analysis is a matrix D (n, m) where n is the number of documents, and m the number of keywords. From this matrix two other matrices are derived, initially a matrix of occurrence of the keywords (columns) in the document collection (rows), and then a co-occurrence matrix making it possible to constitute pairs of terms (or dyads). A normalised weight is attributed to the co-word associations. This is calculated by an association coefficient. In this case, it is called "equivalence coefficient" (Michelet, 1988) and defined as:

$$E_{ij} = \frac{\left[C_{(ij)}\right]^2}{C_{(i)} \times C_{(j)}}$$, where $C(ij)$ is the total number of co-occurrences of words i and j, and $C(i)$ is the total number of occurrences of the word i. The huge normalised and weighted co-occurrence network is submitted to a cluster analysis with the objective to constitute readable sub-sets (or clusters). The clusters are then disposed on a map, which essentially is a bi-dimensional plan. The members of a co-word cluster are also normalised weighted elements. For each internal or external item a of the cluster Cl we calculate this weight w as follows:

$$w_{Cl}(a) = \frac{k_{Cl}(a)}{n_{Clin} + n_{Clex}},$$

with $0 < k_{Cl}(a) \leq n_{Clin} + n_{Clex}$, and $0 < w_{Cl}(a) \leq 1$, where m_{Cl} = the number of its internal and external terms, n_{Clin} = the number of its internal associations, n_{Clex} = the number of its external associations, $k_{Cl}(a)$ = the number of occurrences of internal or external term a (a = 1, m_{Cl}) in the internal or external associations of Cl. The term with most weight serves to label clusters.

Using Graph Theory

A graph G consists of two sets of information: a set of nodes, $N = \{n_1, n_2, ..., n_N\}$, and a set of undirected relations, $R = \{r_1, r_2, \ldots, r_R\}$ between pair of nodes, denoted G (N,R). A weighted graph, denoted by G_W (N, R, W), consists of three sets of information: a set of nodes, $N = \{n_1, n_2, \ldots, n_N\}$, a set of relations, $R = \{r_1, r_2, \ldots, r_R\}$, and a set of weights, $W = \{w_1, w_2, \ldots, w_R\}$, attached to the relations. A directed graph or digraph, G_D (N, R), consists of two sets of information: a set of nodes, $N = \{n_1, n_2, \ldots, n_N\}$, and a set of directed relations (or arcs), $R = \{r_1, r_2, \ldots, r_R\}$ between pairs of nodes. Each arc is an ordered pair of distinct nodes, $r_k = <n_i, n_j>$. The arc $<n_i, n_j>$ is directed from n_i (the origin or sender) to n_j (the terminus or receiver).

Co-word clusters are considered as graphs, G (N,R), because the intra-cluster relations between two internal terms (internal associations) constitute non-directional relations. We use the weighted graphs, G_W (N, R, W), for representing co-word clusters, since cluster relations always are valued by a normalised weight. Following this, we use a directed graphs, G_D (N, R), and we then propose the weighted directed graphs as a model for representing inter-cluster relations (external associations) between two clusters (in the network of clusters), since the external associations can be analysed like sending or receiving relations between two clusters. This is a consequence of the clustering algorithm that we applied (cf. Grivel et al., 1995).

Types of Relations between Clusters in the Network

The external associations can be analysed by pairs of clusters and the possible arcs between them. The classes of ties between the pairs of clusters can be null, asymmetric, and mutual. A pair of clusters (i.e. a dyad) is *null*, that is, when neither arc is present, when neither of the arcs $<Cl_i, Cl_j>$ nor $<Cl_j, Cl_i>$ is contained in the set of arcs. An *asymmetric* pair of clusters has an arc between the two clusters going in one direction or the other, but not both, that is, one of the arcs $<Cl_i, Cl_j>$ or $<Cl_j, Cl_i>$, but not both, is contained in the set of arcs. Asymmetric pairs of clusters are represented by one-way arcs, $<Cl_i \rightarrow Cl_j>$ or $<Cl_j \rightarrow Cl_i>$. *Mutual* or reciprocal pairs of clusters have two arcs between the nodes representing them, one going in one direction and the other going in the opposite direction. Both arcs $<Cl_i, Cl_j>$ and $<Cl_j, Cl_i>$ are contained in the set of arcs. The arc with a double-headed arrow between two nodes indicates a mutual pair of clusters, $<Cl_i \leftrightarrow Cl_j>$ (cf. Wasserman & Faust, 1999, p.124).

Types of Clusters in the Network

According to in- and out-degrees, we can distinguish four types of nodes in a directed graph. We can use this information for analysing the network of clusters in terms of *isolate* nodes (if $d_{in\text{-}degree}(n_i) = d_{out\text{-}degree}(n_i) = 0$), *transmitter* nodes (if $d_{in\text{-}degree}(n_i) = 0$ and $d_{out\text{-}degree}(n_i) > 0$), *receiver* nodes (if $d_{in\text{-}degree}(n_i) > 0$ and $d_{out\text{-}degree}(n_i) = 0$), and *carrier* nodes (if $d_{in\text{-}degree}(n_i) > 0$ and $d_{out\text{-}degree}(n_i) > 0$). Transmitter is the node, which only has arcs originating from it; receiver is the node that only has arcs terminating at it; carrier is the node that has arcs both to and from it (cf. Wasserman & Faust, 1999, p.128).

Conclusion

We try in this article to translate the standard co-word analysis in graph language, following the example of the social network analysis. Graph theory provides us the analytical tools and indicators for analysing co-word clusters as non-directional weighted graphs. In contrast, the network of clusters is defined as a directional weighted graph.

A real co-word analysis application on a bibliographic dataset has been used for illustrating this translation. We have used as example a data set of 228 publications indexed by 164 keywords on rough set theory, and its applications in the field of information science, through 1999-2004. This dataset was collected from PASCAL database.

The issue that remains to be considered is the alternative to apply directly a graph algorithm to the weighted co-occurrence matrix. Thus, a clustering process will be done on the graph structure that has been generated from the weighted co-occurrence matrix. This is in progress using CPCL algorithm introduced by Ibekwe-SanJuan (1998), and revisited by Berry et al. (2004). We will demonstrate in this case that the model only is an undirected weighted graph in conformity with the non-directional co-occurrence relation between keywords.

We have not yet revisited the cluster centrality and density structural properties, nor have we addressed the issue of the different class of centrality (degree, closeness, and betweenness), and density (in a graph, a directed graph or a weighted graph). We will address these issues in a following article.

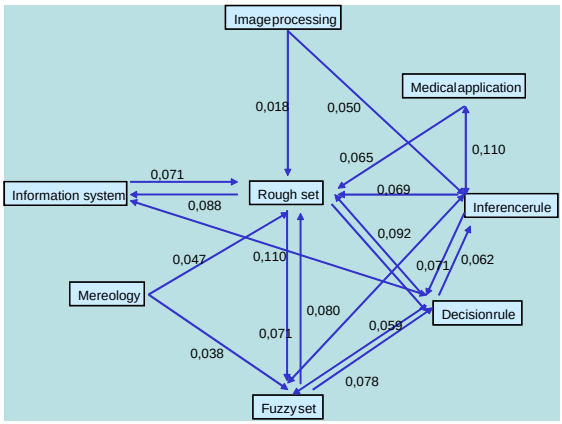

Figure 1: The network of clusters represented by a directed weighted graph

References

Berry A., Kaba B., Nadif M., SanJuan E., Sigayret A. (2004). Classification et désarticulation de graphes de termes, *7èmes Journées Internationa-les d'Analyse Statistique des Données Textuelles*, JADT 2004, Louvain-la-Neuve, Belgique, 10-12 March, vol. 1, p. 160-170.

Callon M., Courtial J-P., Penan H. (1993). *La Scientométrie.* Paris, Presses Universitaires de France, (coll. Que sais-je? Vol. 2727).

Courtial J-P.(1986). Technical issues and developments in methodology. In M. Callon, J. Law & A. Rip (Eds), *Mapping the Dynamics of Science and Technology*. London: The Macmillan Press, c.11, p.189-210.

Degenne A. & Forsé M. (2001). *Les réseaux sociaux*. Paris: Armand Colin.

Everett M.G. & Borgatti S.P. (1999). The Centrality of Groups and Classes, *Journal of Mathematical Sociology*, vol. 23, num. 3, p. 181-201.

Grivel L., Mutschke P., Polanco X. (1995). Thematic Mapping on Bibliographic Databases by Cluster Analysis: A Description of the SDOC Environment with SOLIS, *Knowledge Organization*, vol. 22, num. 2, p. 70-77.

Ibekwe-SanJuan F. (1998). Terminological variati-on, as a means of identifying research topics from texts, *Proceedings of the Joint International Conference on Computational Linguistics*, COLING-ACL'98, Montréal, 10-14 August, p. 564-570.

Michelet B. (1988). *L'Analyse des Associations*. Thèse de Doctorat, Université de Paris 7.

Wasserman S. & Faust K. (1999). *Social Network Analysis. Methods and Applications*. Cambridge: Cambridge University Press.