A novel approach based on clustering along with feature maximization and contrast graphs

An example of application for diachronic analysis of research

Jean-Charles LAMIREL

SYNALP Team-LORIA, INRIA Nancy-Grand Est
Vandoeuvre-lès-Nancy, France
jean-charles.lamirel@loria.fr

ABSTRACT. In this paper, we perform an analysis of the contents of selected academic journal papers in Science of Science in China and the construction of an overall map of the research topic structure during the last 40 years. Furthermore, we highlight the topic evolution by the exploitation of the publication dates information for the sake of clarifying topics changes. Consequently, the proposed novel method, based on the unsupervised combination of GNG clustering with feature maximization metrics and associated contrast graph, permits without supervision, without parameters and without help of any external knowledge to have very clear and very precise insights of the development of a scientific domain.

KEYWORDS: Science of Science; China; World; Topic evolution; Feature maximization; Unsupervised learning; Diachronic analysis.

1. Introduction

"Science of Science" refers to the research on scientific and technological knowledge and explores the fundamental laws of the development of science and technology. In this article, we exploit the research material of the last 40 years in Science of Science in China and put in place a new method to understand and monitor both more clearly and more accurately the development in this field. Our objective is therefore to provide relevant indications on the origin of Chinese Science of Science, its structure and future directions through an original data analysis method, operating in a completely unsupervised manner, without parameters and without external knowledge source. The proposed method can be considered as an efficient alternative to usual LDA approaches (Blei et al. 2003).

The next section focuses on the description of the feature maximization metric and its associated feature selection and visualization methods. Section 3 presents your data collection and preprocessing methods. Section 4 presents our experimental protocol. Section 5 is dedicated to the description of our experimental results. Lastly, section 6 draws our conclusion and perspectives.

2. Feature maximization as a global approach for data analysis

Most of our further data analysis work on the Science of Science dataset is based on a feature selection approach relying itself on feature maximization metric (Lamirel et al., 2011). Feature maximization is an unbiased metric which can be used to estimate the quality of a classification whether it is supervised or unsupervised. In unsupervised classification (i.e. clustering), this measure exploits the properties (i.e. the features) of clusters' associated data for different purposes (clustering labeling and cluster content highlighting, clustering results global visualization like the representation by contrast graph presented in this paper, optimal clustering model detection). Its main advantages are to be parameter free, to be totally independent of the clustering method and of its operating mode, to work suitably in high dimensional spaces and to represent a better compromise between discrimination and generalization than usual metrics (Euclidean, Cosine or Chi-square, …).

2.1 Feature F-measure

Consider a partition C which results from a clustering method[1] applied to a dataset D represented by a group of features F. The feature F-measure $FF_c(f)$ of a feature f associated with a cluster c is defined as the harmonic mean of the Feature Recall $FR_c(f)$ and of the Feature Predominance $FP_c(f)$, which are themselves defined as follows:

$$FR_c(f) = \frac{\Sigma_{d \in c} W_d^f}{\Sigma_{c \in C} \Sigma_{d \in c} W_d^f} \quad (1)$$

$$FP_c(f) = \frac{\Sigma_{d \in c} W_d^f}{\Sigma_{f' \in F_c, d \in c} W_d^{f'}} \quad (2)$$

with

$$FF_c(f) = 2\left(\frac{FR_c(f) \times FP_c(f)}{FR_c(f) + FP_c(f)}\right) \quad (3)$$

where W_d^f represents the weight of the feature f for the data d and F_c represents all the features present in the dataset associated with the cluster c. Feature Predominance measures the ability of f to describe cluster c. In a complementary way, Feature Recall allows to characterize f according to its ability to discriminate c from other clusters.

2.2 Feature maximization

In supervised context, feature maximization measure can be exploited to generate a powerful feature selection process. In our unsupervised (clustering) context, the selection process can be used to describe or label clusters according to the most typical and representative features. This process is a parameter-free process that uses both the capacity of F-measure to discriminate between clusters ($FR_c(f)$ index) and its ability to faithfully represent the cluster data ($FP_c(f)$ index). The set S_c of features that are characteristic of a given cluster c belonging to a partition C is defined as:

$$S_c = \left\{ f \in F_c | FF_c(f) > \overline{FF}(f) \;\; and \;\; FF_c(f) > \overline{FF}_D \right\} \quad (4)$$

$$\text{with } \overline{FF}(f) = \Sigma_{c' \in C} \frac{FF_{c'(f)}}{|C_{/f}|} \;\; and \;\; \overline{FF}_D = \Sigma_{f \in F} \frac{\overline{FF}(f)}{|F|} \quad (5)$$

where $C_{/f}$ represents the subset of C in which the feature f occurs.

Finally, the set of all selected features S_C is the subset of F defined by:

$$S_C = \cup_{c \in C} S_c \quad (6)$$

In other words, the features judged relevant for a given cluster are those whose representations are (1) better in this cluster than their average representation in all the clusters, and (2) better than the average representation of all the features in the partition, in terms of Feature F-measure. Features which never respect the second condition in any cluster are discarded.

2.3 Contrast

A specific concept of contrast $G_c(f)$ can be defined to calculate the performance of a retained feature f for a given cluster c. It is an indicator value which is proportional to the ratio between the F-measure

[1] In this article, the features represent the words extracted from the title, abstract and keywords of the article, the weights of the features are the adjusted frequency information associated with them and the unsupervised classification (clustering) is based on the GNG algorithm.

$FF_c(f)$ of a feature in the cluster c and the average F-measure $\overline{FF}$ of this feature for the whole partition.Contrast of a feature f for a cluster c is expressed as:

$$G_c(f) = FF_c(f)/\overline{FF}(f) \qquad (7)$$

The active features of a cluster are those for which the contrast is greater than *1*. Moreover, the higher the contrast of a feature for one cluster, the better its performance in describing the cluster content. A simple and illustrative example of the feature maximization and feature selection operating mode can be found in (Lamirel et al. 2015).

As already mentioned before, in clustering, the active features in a cluster are selected features for which the contrast is greater than 1 in that cluster. Conversely, the passive features in a cluster are selected features present in the cluster's data for which contrast is less than unity. A simple way to exploit the features obtained is to use active selected features and their associated contrast for cluster labeling as we proposed in (Lamirel et al. 2015) and we also exploit further in the current experimental context.

2.4 Contrast graphs

In the mathematical field of graph theory, a bipartite graph (or bigraph) is a graph whose vertices can be divided into two disjoint and independent sets U and V such that every edge connects a vertex in U to one in V. Contrast graphs are bipartite graphs based on the relations between a set of features S and a set of labels L (Cuxac and Lamirel 2013). Theoretically, the set of labels L could represent any kind of information to which features can be related with and the set of features S is a subset of a global feature set F (i.e. he original feature space on which rely the data of a dataset) that has been obtained through a feature selection process, like feature maximization presented above. In the case of the use of feature maximization, the weight $c_{(u,v)}$ of an edge $(u, v), u \in S, v \in L$ represents the contrast of feature u for a label v as, it is defined by equation 7. [2]

2.5 Exploitation of complementary information through external labels

As it is defined in (Attik et al. 2006), external labels are information that is associated to data but that does not play any role in the initial data analysis process. However, this information could carry important clues for enhancing the precision of the analysis. In the case of the formerly presented clustering process, external labels can be exploited in a secondary step (i.e. after de clustering process) by evaluating their posterior distribution into the clusters through clusters' associated data for providing complementary information about those latter or about the related topics.

In the case of our Science of Science dataset we focus on one kind of external labels that is papers' publication dates. Papers' publication dates are exploited to perform a diachronic analysis of the topics' activity, highlighting the importance of each topic in each time period, either this activity is considered individually or relatively to the other topics. As it is shown in the next section related to the analysis of the results, this approach helps to precisely understand the chronology of the research activity of a global research domain, like in our specific case the Science of Science domain.

In the context of our experiment, our external label analysis is based on two different measures that are the label frequency. Label frequency F_c^l of a label l of a type t in a cluster c can be defined as:

$$F_c^l = Card\{d \in D \mid af(d) = c \wedge l \in Extlab_t(d)\} \qquad (8)$$

where Card is the set cardinal function, D is the whole set of exploited data, af the function defined at eq. 10 (that provides the cluster associated to data d) and $Extlab_t(d)$ a function that provides the list of external labels of type t associated to data d.

[2] In equation 7, labels materialize in this case categories or clusters to which data are associated.

3. Data collecting process and preprocessing

For our experiment, we queried the China National Knowledge Infrastructure (CNKI) database using "Science of Science" as the thematic term. We extracted 2401 articles published in the main journals of Beijing University and CSSCI3 (covering a research period until 2017-10-22). The indexation process has been quite complex. It started with an initial dictionary of 9679 keywords gathered from the keyword field of the 2790 articles. For word segmentation and tagging of titles and article summaries, we used NLPIR-ICTCLAS, a specific toolbox for Chinese language processing. The English translation was then applied to the dictionary of obtained keywords. A frequency threshold of 6 was finally applied to remove low frequency words. It issued in a final dictionary of 1576 terms with which the articles were re-indexed.

4. Data analysis process

We present the global architecture of our experimental process in figure 1. After light preprocessing steps (see section 3), the process exploits clustering in combination with feature maximization to extract the main topics of research from the Science of Science dataset we are dealing with. We have recently shown (Lamirel et al. 2015) that the combination of a suitable clustering approach, like neural clustering (Fritzke 1995), with feature maximization offers superior performance as compared to alternative approaches for topic extraction, like LDA (Blei et al. 2003), at the condition to be able to properly identify an optimal clustering model (i.e. suitable number of clusters) from the analyzed data. We thus propose to exploit one of your recent and efficient approach also based on feature maximization for the optimal model detection task (Lamirel et al. 2016). The management of the clustering results with a graph approach based of contrast is an original method presented in this paper. It permits both to reduce the cognitive overload that should result of the representation of interactions in large datasets and to accurately figuring out the dependencies between extracted topics through shared features with high contrast. The last part of our approach exploit external labels of data associated to clusters. Publication date is used to perform a diachronic analysis of clusters (i.e. topics) activity. Date information is also reported on the contrast graph.

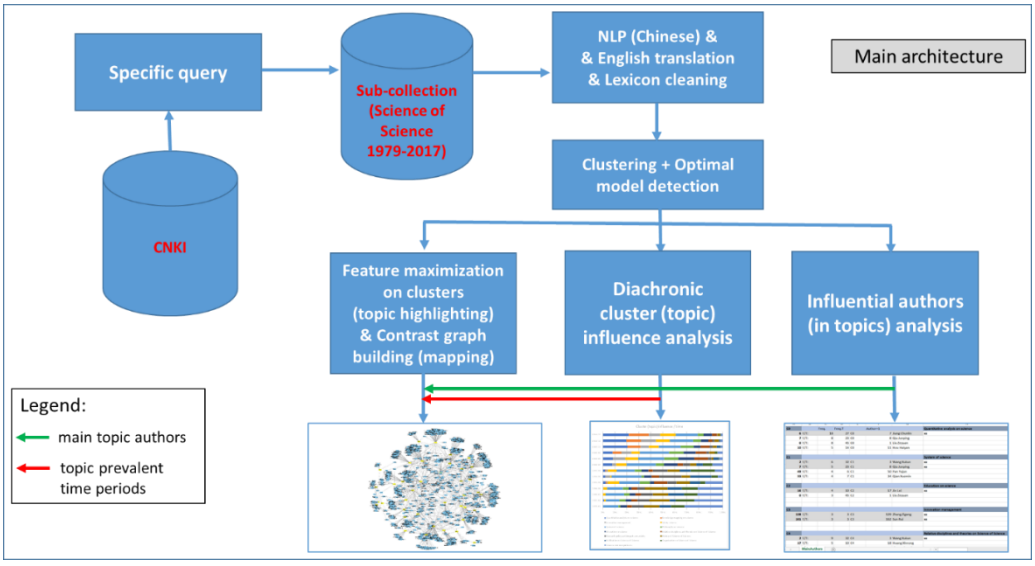

***Figure 1**: Global data analysis process.*

3 Chinese Social Sciences Citation Index.

As part of our experience, we vary the number of clusters in a range of up to 1/50 of the number of data using GNG clustering algorithm (Fritzke 1995) with standard Fritske parameters and coupling it with our optimal model detection approach. This strategy makes it possible to obtain the relevant number of clusters highlighting the main science research topics during the period under consideration. Expert analysis of the results obtained confirms that the clustering model chosen as optimal by our approach (13 clusters) is consistent to accurately represent all the main research topics in the Science of Science field. Figure 2 presents a description of clusters based on their most contrasted characteristics and Table 1 presents the list of cluster titles that the expert has characterized by exploiting the most contrasted elements.

C. 9# : Knowledge mapping on science	
5.376770 theme, 5.030978 research hot topics, 4.827424 literature, 4.734794 software, 4.697236 frontier, 4.595268 development trend, 4.401170 research topic, 4.342141 hotspot, 4.159228 both at home and abroad, 3.989917 science knowledge mapping,	3.873852 international, 3.801943 expectation, 3.778949 data, 3.721473 knowledge map, 3.648744 visualization analysis, 3.641972 tool, 3.557082 research situation, 3.495185 trend, 3.411639 representative figure, 3.327669 research direction, ……

Figure 2*: Example of description of a cluster through the list of its most contrasted features (here words). The cluster's related topic is knowledge mapping.*

	Label (expert)	**Content summary**
Cluster 0#	Quantitative analysis on science	Bibliometrics, citation analysis, journal, indicator, quantity, impact factor, statistics analysis, data, SNA
Cluster 1#	Research evaluation	Efficiency, systems engineering, decision making, forecast, evaluation, administration, input and output, efficiency, sustainable development
Cluster 2#	Education on science and talent cultivation	Higher education, Ministry of Education, planning, talent cultivation, university
Cluster 3#	Innovation management	Enterprise, knowledge management, collaborative innovation, performance, competitive advantage, technological innovation, integration
Cluster 4#	Domain structure and peripheral disciplines on Science of Science	S&T studies, theory of science of science, technology theory, technology philosophy, dialectics of nature, library science, knowledge-based economy, history of science of science, discipline structure
Cluster 5#	Philosophical foundation on Science of Science	Philosophy, Marxist doctrine, reality, criticism, ontology, dialectics, human society, materialism, humanism
Cluster 6#	Discipline system	Definition, connotation, discipline system, research method, concept, principle, comparative research, system science, safety, safety principle, safety system
Cluster 7#	Research policy and impacts on society	Scientifilization, S&T development, modern management, productivity, nation, world, emancipation of mind, socialism, social economic development

Cluster 8#	Subject attributes on Science of Science	Natural science, social science, modern science, regular pattern, development principle, edge, interdisciplinary research
Cluster 9#	Knowledge mapping on science	Research hot topics, software, hotspot, theme, frontier, development trend, knowledge map, data, visualization analysis
Cluster 10#	History on Science of Science	History of science, creator, JD.Bernard, Price, big science, Zhao Hongzhou, scientometrics, Soviet Union, world science, sociology of science
Cluster 11#	Publication on Science of Science	Journal, publication, S&T management, S&T system reform, S&T circle, editorial office, institute, S&T policy
Cluster 12#	Organization on Science of Science	Committee, leadership, Chinese Association for Science of Science, conference, symposium, academic exchange, Liu Zeyuan

***Table 1**: List and summary description of the obtained clusters.*
For better clarity, cluster labels are added by the expert of the domain.

5. Data analysis and visualization results

5.1 General topic structure of the Science of Science domain

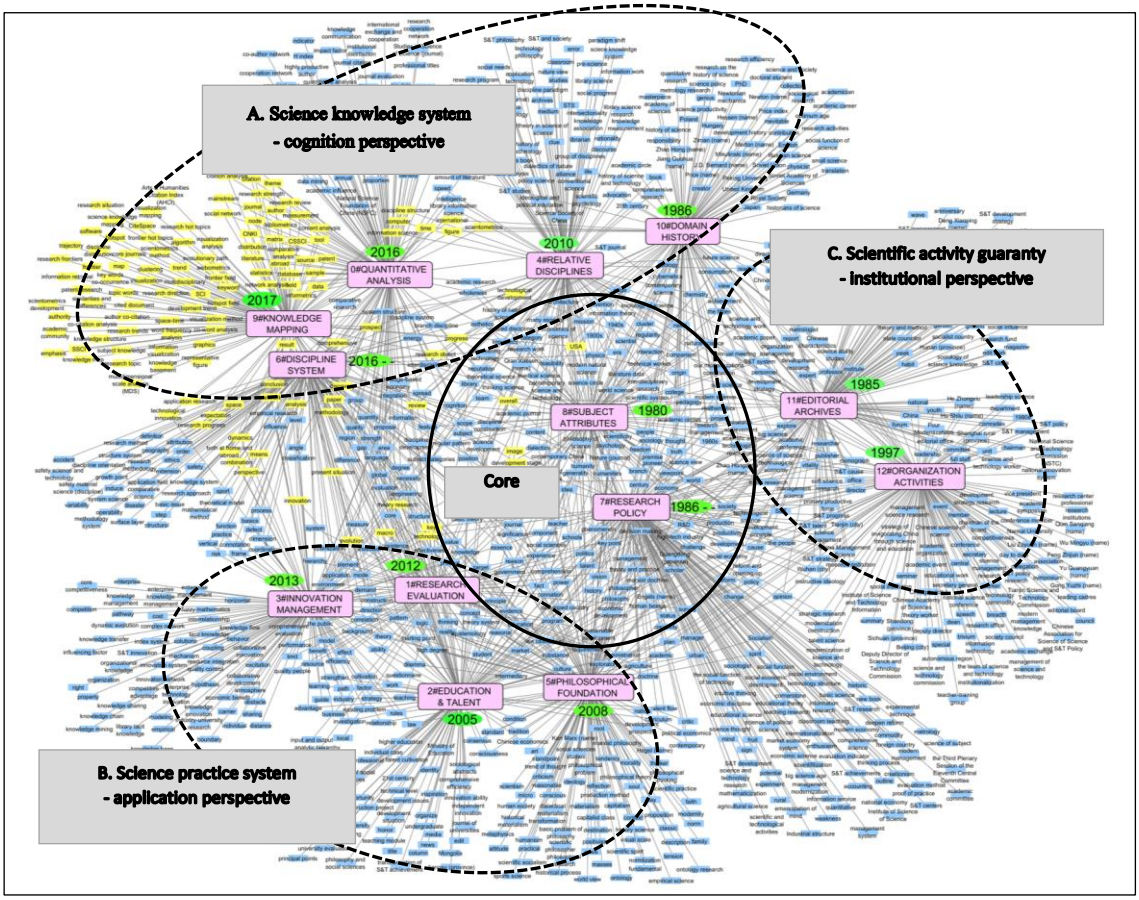

***Figure 3**: Global contrast graph representing main topics and domain structure in Science of Science in China.*

(Cluster 9# is highlighted, and the detailed information on clusters is shown in Table 2).

In the specific case of our experiment on Science of Science data, we propose to build a contrast graph between a set of clusters (set L) representing the main research topics of the domain that have been extracted by the clustering process and the most contrasted features (set S) issued from the cluster descriptions. This approach, which combines clustering and contrast graphs in an original way, is particularly useful for Science experts to understand the construction of their domain, highlighting the most central topics in the domain (domain generators) as well as those that are most connected. In the resulting graph, only the edges with a contrast greater than 1.4 are kept to participate in the representation (1074 of the 1576 features (or words) used for clustering are kept).

The spatial distribution of the 13 topics is shown in the resulting graph presented in Figure 3. According to experts, this graph highlights a very clearly interpretable structure of the Science of Science field in China. In such a model, highly interconnected topics will tend to appear at the center of the representation (see section 2.5). In our case, this information on the core domains is represented by two complementary topics: "8# Attributes of the Science of Science domain" and "7# Research policy and impacts on society"

5.2 The evolution of Science de la Science

To materialize the evolution of the 13 research topics (see section 4), we exploit the publication dates of the analyzed articles. Hence, as presented in Figure 4, a global representation of the influence of each cluster (i.e. topic) in different periods (using 3-year blocks) can be derived from the distributions of the publication dates in the different clusters. The said representation can then be used to better understand the laws of Science of Science development in China. This point of view can especially help to distinguish between important but accidental topics that have a certain chance of developing in the short term and rational important topics that play a major role in the construction of the domain in the long term.

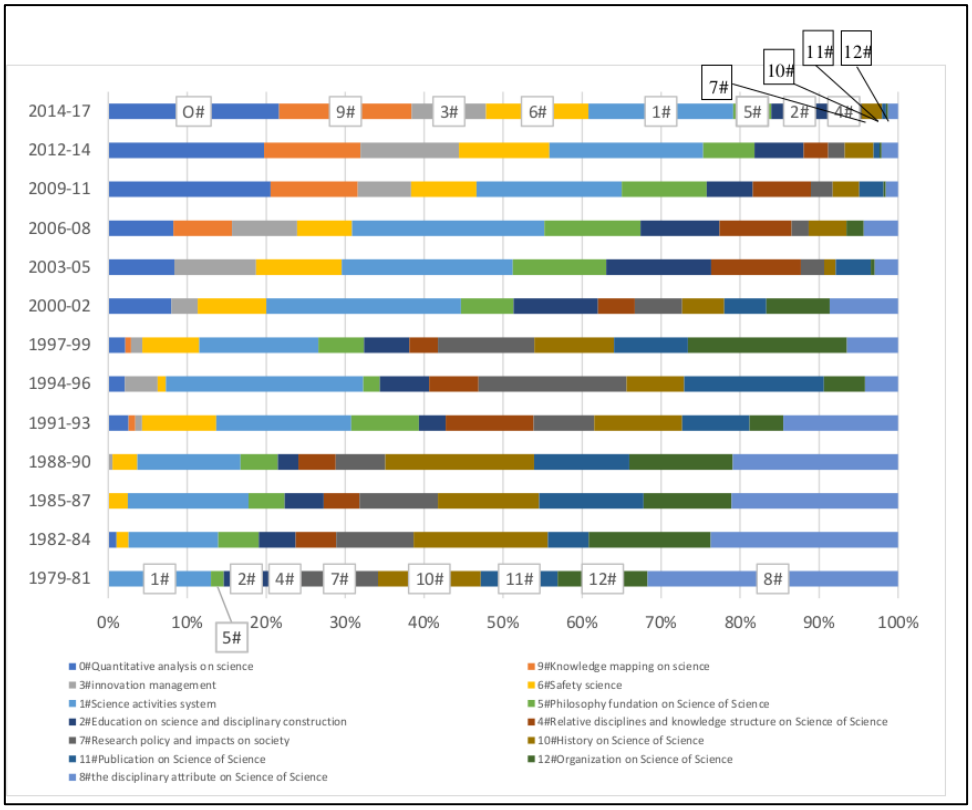

***Figure 4**: Coordinated influence of research topics in Science of Science in China.*

The topics "0# Quantitative Analysis of Science", "9# Mapping of Science Knowledge" and "3# Innovation Management" did not appear at the beginning of scientific research in the field of Science of Science in China, but only in recent years, the status of these topics becoming more and more important. The establishment of the dominant position of the topic "0# Quantitative Analysis of Science" shows that Science of Science has reached maturity as a subject. The importance of the topic "9# Science knowledge mapping" indicates that Science of Science has become an open subject, integrating computational approaches and information visualization technologies. The growing prosperity of the topic "3# Innovation management" shows that Science of Science is an increasingly practice-oriented domain that emphasizes the economic value of science and technology, and shows its strategic position in China today. In comparison, the research topics on the attributes of the field (8#), the construction of scientific organization and publication processes (#11) and the management of scientific research results (#12) have gradually weakened, which also indicates that scientific research in Science of Science is gradually becoming mature and standardized in China. This evolutions trends have also been validated by the experts of the domain.

6. Conclusions and discussion

The combination of feature maximization measure and unsupervised learning and the joint use of contrast graphs for visualization is an original approach that we have proposed in this work. Our full-scale experiments, which were validated by experts in the field, showed that this method could, without supervision, without parameters and without the support of any external source of knowledge, reveal very effectively the research topics, their interactions and changes in a very complex research field such as Science of Science in China. In this article, we propose in particular a method for visualizing the analysis results using feature maximization. This method is very suitable for large-scale data analysis in large dimensions. It also tolerates the integration of a wide range of additional information that can enrich analytical results and provide clarity and precision of results that current competing methods cannot provide. For example, methods such as LDA, which could replace the proposed approach, for the part concerning the topic extraction, suffer too much from the dependence on parameters that are very difficult to control and working hypotheses that are difficult to verify on the distribution of words, as these problems severely limit the quality of the results (level of generality, precision).

Bibliography

Attik M., Lamirel J.-C., Al Shehabi S. (2006) Clustering analysis for data with multiple labels, *Proceedings of IASTED International Conference on Databases and Applications (DBA)*, Innsbruck, Austria, February 2006.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993–1022.

Cuxac, P., & Lamirel, J.-C. (2013). Analysis of evolutions and interactions between science fields: the cooperation between feature selection and graph representation. In 14th *COLLNET Meeting*.

Fritzke, B. (1995). A growing neural gas network learns topologies. *In Advances in neural information processing systems* (pp. 625–632).

Kassab, R., & Lamirel, J.-C. (2008). Feature-based cluster validation for high-dimensional data. I*n Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications* (pp. 232–239). ACTA Press.

Kobourov, S. G. (2012). *Spring embedders and force directed graph drawing algorithms.* arXiv preprint arXiv:1201.3011.

Lamirel, J.-C., Mall, R., Cuxac, P., & Safi, G. (2011). Variations to incremental growing neural gas algorithm based on label maximization. *In The 2011 International Joint Conference on Neural Networks (IJCNN),,pp. 956–965 IEEE.*

Lamirel, J.-C., Cuxac, P., Chivukula, A. S., & Hajlaoui, K. (2015). Optimizing text classification through efficient feature selection based on quality metric. *Journal of Intelligent Information Systems*, 45(3), 379–396. doi:10.1007/s10844-014-0317-4

Lamirel, J.-C., Dugué, N., & Cuxac, P. (2015). Performing and visualizing temporal analysis of large text data issued for open sources: past and future methods. *In Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery* (pp. 56–76). Springer.

Lamirel, J.-C., Dugué, N., & Cuxac, P. (2016). New efficient clustering quality indexes. In 2016 *International Joint Conference on Neural Networks (IJCNN)*, pp. 3649–3657. IEEE.