Feature selection and complex networks methods for an analysis of collaboration evolution in science: an application to the ISTEX digital library

Nicolas Dugué, Ali Tebbakh, Pascal Cuxac, Jean-Charles Lamirel

▶ **To cite this version:**

Nicolas Dugué, Ali Tebbakh, Pascal Cuxac, Jean-Charles Lamirel. Feature selection and complex networks methods for an analysis of collaboration evolution in science: an application to the ISTEX digital library. ISKO-MAGHREB 2015, Nov 2015, Hammamet, Tunisia. hal-01231791

HAL Id: hal-01231791
https://hal.science/hal-01231791

Submitted on 20 Nov 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Feature selection and complex networks methods for an analysis of collaboration evolution in science: an application to the ISTEX digital library

Nicolas DUGUÉ, Ali TEBBAKH, Pascal CUXAC and Jean-Charles LAMIREL

[1]*Abstract*—In this paper, we aim to give insights about the self-organization of scientific collaboration. To that aim, we describe a new framework to monitor the evolution of a collaboration graph that models the co-authorship of research papers authors. We use community structure of the network as a high-level description of its self-organization and thus consider the evolution of the communities across time. To monitor this evolution, we describe a diachronic analysis method based on the extraction of prevalent nodes for each community. We apply this approach on data issued from the ISTEX project, a scientific digital library that contains so far more than 16 million documents and present some preliminary results and visualizations.

***Index Terms*— Feature selection, complex networks, diachronic analysis, communities, dynamic graphs, co-authorship evolution.**

I. Introduction

THE ISTEX project (Excellence Initiative for Scientific and Technical Information) is part of the "Investments for the Future" program initiated by the French Ministry for Higher Education and Research (MESR). The ISTEX project's main objective is to provide the whole French higher education and research community with online access to retrospective collections of scientific literature in all disciplines by setting up a national document acquisition policy covering journal archives, databases, text corpora etc. (http://www.istex.fr/). The first stage of the ISTEX project relates to a large-scale proactive policy in favor of grouped acquisitions of scientific archives under national licenses. The second stage of the ISTEX project involves setting up the ISTEX platform to host all the data. Access to document resources will be provided in 2015 via the ISTEX platform administered by INIST-CNRS. This platform will host several million of digital documents in all disciplines and will offer varied benefits for users.

On the basis of the initial platform services, we are currently working towards proposing new added-value services. One of our central concern is then to develop tools for highlighting the dynamics of the collection.

Hence, the development of dynamic information analysis methods, like incremental clustering and novelty detection techniques, is becoming a central concern in a bunch of applications whose main goal is to deal with large volume of information, such as ISTEX ones, whose content is significantly varying over time.

In this paper, we aim to give insights about the self-organization of scientific collaboration. We thus make use of collaboration graphs that model the co-authorship of research papers authors. In such a graph G=(V,E,W), the set of vertices V describe the set of authors whilst the set of edges E describe the co-authorship relations between authors. The set of weights W associated to the edges E describe the frequency of co-publication. Basically, if (v1,v2) is an edge with v1 and v2 vertices of V, then v1 and v2 published W_{v_1,v_2} papers together. This graph is actually the unipartite projection of the bipartite graph which links authors to research papers. Thus, authors linked to the same paper in the bipartite graph shapes a clique in the unipartite projection, namely a complete subgraph.

Sociologists, complex networks scientists and physicists have shown that such graphs are of interest to study scientific production [1]. Indeed, the structures of these graphs have an impact on the success of collaborations according to Uzzi and Spiro [2]. Furthermore, Burt claims that being part of several different « groups » in such a graph increases creativity [3]. Finally, groups in such graphs are often called « communities », and are described as group of authors that published more together than with the rest of the network in this context [4]. These so-called communities are proved to be efficient to model the network as a map of different knowledge domains or fields [5] [6].

In the framework of this paper, we thus use community structure of the network as a high-level description of its self-organization. We therefore consider the evolution of communities across time in the collaboration network of our corpus extracted from ISTEX. To detect these communities, we make use of the INFOMAP algorithm which is proved to be particularly efficient and fast to run [7].

1 Pascal CUXAC, CNRS-INIST, Vandoeuvre lès Nancy, France. (e-mail: *pascal.cuxac@inis.fr*). Jean-Charles LAMIREL and Nicolas Dugué, LORIA-Synalp, Vandoeuvre lès Nancy, France. (e-mail : first.last@loria.fr). Ali TEBBAKH, LORIA-LorExplor, Vandoeuvre lès Nancy, France. (e-mail : ali.tebbakh@loria.fr).

Once the communities detected on the several periods, we use a diachronic analysis to analyze the dynamics of these communities. The purpose of diachronic mapping here, is to track communities' appearance, disappearance, divergence or convergence across time.

In order to identify and analyze the emergences, or to detect changes in the data, we have previously proposed two different and complementary approaches:

- Performing static classifications at different periods of time and analyze changes between these periods (time step approach or diachronic analysis) [8];
- Developing methods of classification that can directly track the changes: incremental clustering methods (incremental clustering) [9] and novelty detection methods (incremental supervised classification) [10].

We present hereafter an original method relying on the first approach and using feature maximization metric [11] to monitor the evolution of collaboration graphs across time. Unlike some common approaches [12] [13], we are tackling the problem using community detection in time periods in combination with feature selection to associate salient authors with communities. In a further step, we construct a graph visualizing the interactions between salient authors and their collaboration in the different time periods.

In the following sections, we first present short states-of-the-art on evolution detection and on feature selection. In a second step we present our feature maximization metric exploited throughout our approach. In a third step, we describe the diachronic analysis used to monitor the evolution of communities. In a next step, we describe our experimental data and associated preprocessing. Lastly we highlight our preliminary results and conclusion.

II. State of the art

A. Evolution detection

One of the main objectives of the analysis of the scientific and technical information is to identify the major changes linked to developments in science. Emerging technologies play an essential role both in scientific and industrial advances. On the one hand, in the technology field, the monitoring of the evolution of patents is essential to maintain a technological leadership over its competitors. On the other hand, analysis of the results of basic research can identify scientific advances that might well end up in technological advances. Last but not least, in the activity of researchers, analyzing changes and monitoring the development of cross thematic or emerging themes allows them to ensure the innovativeness of their research topic.

Visualization of the results of the incremental classification represents an important milestone for the understanding of the corresponding analyses. Without this step, arrays of numbers and words are the only output that the user can operate, with all the difficulties that we imagine. In recent years, technological advances allowed the emergence of new methods of representation, particularly for text data.

The ThemeRiver approach [14] allows to visualize changes in counts. The topics associated with the data are constructed from occurrences of terms. If this method allows well to highlight the relations of counts over time, this representation has the disadvantage of not to reveal any structure or relationships between the data. Erten and al. [15] propose to visualize the evolution of the topics through the TGRIP system. It illustrates the evolution of the size of the topics in the form of a graph. The size of each vertex of such graph is evolving on the basis of the number of data that contains the topic represented by the said vertex. This method allows to highlight the existence of a thematic structure, the evolution being suggested by the superposition of levels. The approach proposed by the CiteSpace [16] system allows the representation of the evolution of networks of citations between bibliographic data. For that purpose, authors use two different temporal dimensions: the date of publication and the date of citation. Publication date determines the position of the data (the nodes of the graph) along a time axis. The second dimension corresponds to the year of citation: each node is characterized by different levels of colors that represent the year or the corresponding data has been cited. In such a way, this approach based on the citations reveals the dynamics of construction of networks of data on close topics, but doesn't provides any overall vision.

As it is also shown in more recent works as those based on dynamic trees [17], the visualization of the results of incremental classification remains, and still to this day, an important, even vital, field of investigation towards end-users. It is likely that after having explored various tracks, the ideal solution is not a single type of visualization, but rather a combination of approaches.

B. Feature selection

Since the 1990s, advances in computing and storage capacity allow the manipulation of very large data: it is not uncommon to have description space of several thousand or even tens of thousands of variables. One might think that classification algorithms are more efficient if there are a large number of variables. However, the situation is not as simple as this. The first problem that arises is the increase in computation time. Moreover, the fact that a significant number of variables are redundant or irrelevant to the task of classification significantly perturbs the operation of the classifiers. In addition, as soon as most learning algorithms exploit probabilities, probability distributions can be difficult to estimate in the case of the presence of a very high number of variables. The integration of a variable selection process in the framework of the classification of high dimensional data is a central challenge.

In the literature, three types of approaches for variable selection are mainly proposed: the integrated (embedded) approaches, the "wrapper" methods and the filter approaches. An exhaustive overview of the state-of-the-art techniques in this domain has been achieved by many authors, like Ladha and al. [18], Bolón-Canedo and al [19], Guyon and al [20] or Daviet [21]. For an overview of these methods, you might refer to the previous articles, as well as to [11].

III. Feature maximization for feature selection

A. Feature maximization principles in unsupervised learning

Feature maximization (F-max) is an unbiased cluster quality metric that exploits the features of the data associated to each cluster without prior consideration of clusters profiles. This metrics has been initially proposed in Lamirel and al [22]. Its main advantage is to be independent altogether of the clustering methods and of their operating mode. This metric was previously used in a data clustering context. We adapt it and describe it in a graph context to fit with our application. Indeed, by using Feature maximization, we aim to associate salient authors to communities of authors that are highly connected in the collaboration graphs.

Consider a weighted undirected graph G=(V, E, W) where V is the set of vertices, E the set of edges between pairs of vertices of V and W, the set of weights associated to the edges of E. We also consider the set of communities C, which is a partition of the set of vertices V into clusters of highly connected nodes.

The *Feature F-measure* $FF_c(v)$ of a *vertex v of V* associated to a community *c* of C is defined as the harmonic mean of *Feature Recall* $FR_c(v)$ and *Feature Precision* $FP_c(v)$ indexes which in turn are defined as:

$$FR_c(f) = \frac{d_c(v)}{\sum_{c_i \in C} d_{c_i}(v)}, FP_c(f) = \frac{d_c(v)}{d_c}$$

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

where $d_c = \sum_{v_1 \in c} \sum_{v_2 \in c} W_{v_1,v_2}$ is the sum of weights of the community c, $d_c(v) = \sum_{u \in c} W_{u,v}$ is the degree of a vertex v related to the community c.

B. Adaptation of feature maximization metric for feature selection in supervised learning

Taking into consideration the basic definition of feature maximization metric presented in the former section, its exploitation for the task of node selection in the context of supervised learning becomes a straightforward process. The feature maximization-based selection process can thus be defined as a parameter-free community based process in which a node v is characterized using both its capacity to discriminate a given community from the others (Feature Precision index) and its capacity to accurately represent the community (Feature Recall index).

The set S_c of nodes that are characteristic of a given community *c* belonging to *C* results in:

$$S_c = \{v \in V_C | FF_c(v) > \overline{FF}(v) \wedge FF_c(v) > \overline{FF}_V\}$$

where $\overline{FF}(v) = \sum_{c \in C} FF_c(v)/|C_v|$ and

$$\overline{FF}_V = \frac{1}{|V|}\sum_{v \in V} \overline{\mathrm{FF}}(v)$$

and C_v represents the restriction of the set *C* to the communities in which the node v is represented.

Finally, the set of all the selected nodes S_C is the subset of *V* defined as:

$$S_C = \cup_{c \in C} S_c$$

Nodes that are judged relevant for a given community are the nodes whose representation is altogether better than their average representation in all the communities and better than the average representation of all the nodes, as regard to the feature F-measure metric.

IV. Diachronic analysis

We now describe the method that allows us to monitor the communities' evolutions between time periods. We consider here two periods with their own collaboration graph, the source period and the target period. *S is* the set of communities detected on the graph of the source period, and *T* the sets of communities detected on the graph of the target period.

To compute the probability of matching between communities belonging to these two periods, we slightly modify the standard computation of the Bayesian inference provided by the original MVDA model [23]. The new computation is expressed as:

$$P(t|s) = \frac{\sum_{v \in L_s \cap L_t} FF_t(v)}{\sum_{v \in L_t} FF_t(v)}$$

where *s* represents a community of the source period, *t* a community of the target period, L_x represents the set of nodes that are salient and thus associated to the community *x u*sing the cluster feature maximization approach defined in the previous section, and $L_x \cap L_y$ represents the common salient nodes, which can be called the **nodes matching kernel** between the community *x* and the community *y*.

The average matching probability $P_A(S)$ of a source period community can be defined as the average probability of activity generated on all the communities of the target period by its associated salient nodes:

$$P_A(S) = \frac{1}{|Env(s)|} \sum_{t \in Env(s)} P(t|s)$$

where *Env(s)* represents the set of target period communities activated by the salient nodes of the source period community *s*.

The global average activity A_s generated by a source period model *S* on a target period model *T* can be defined as:

$$A_S = \frac{1}{|S|} \sum_{s \in S} P_A(s)$$

Its standard deviation can be defined as σ_s.

The **similarity** between a community s of the source period and a community *t* of the target period is established if the 2 following similarity rules are verified:

$$P(t|s) > P_A(s) \text{ and } P(t|s) > A_s + \sigma_s$$

$$P(s|t) > P_A(t) \text{ and } P(s|t) > A_t + \sigma_t$$

Community splitting is verified if there is more than one community of the target period which verifies the previous similarity rules with a community of the source period. Conversely, **community merging** is verified if there is more than one community of the source period which verifies the similarity rules with a cluster of the target period.

Communities of the source period that do not have similar communities on the target period are considered as **vanishing communities**. Conversely, communities of the target period that do not have similar community on the source period are considered as **appearing communities**.

V. DATA

Our experimental data is a collection of 7903 scientific papers in English language related to gerontology domain published between 2000 and 2010. This collection has been extracted from ISTEX database by INIST documentary engineers specialized in the medical domain.

As soon as the full-text extracted documents are formatted in an XML format which is specific to each publisher (most represented publishers are: Elsevier, Oxford University Press, Nature, Institute of Physics, Royal Society of Chemistry), we had first to check the structure of the extracted documents based on their related DTD to retrieve the year and authors. This task is performed through SPARKL-like queries using our own XML management toolkit [24].

VI. EXPERIMENTAL RESULTS

To clarify the principle of our approach, we named CGEM[2] (Collaboration Graph Evolution Monitoring), we follow four steps that are schematically presented in Figure 1:

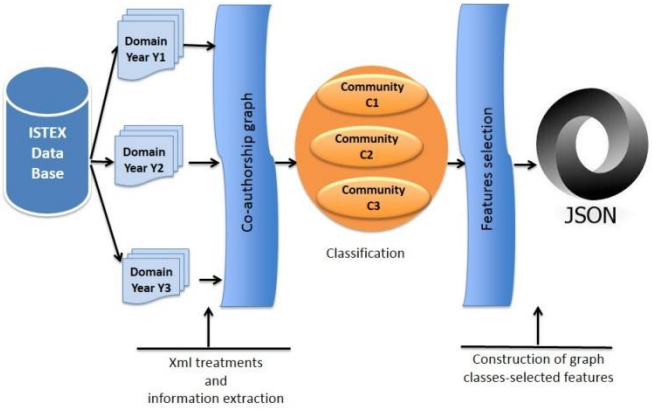

Figure 1 : The CGEM approach

1) *We query a ISTEX database to produce an initial corpus;*
2) *The documents are split into sub corpora that represent different publishing periods;*
3) *Python Script are used to create the weighted undirected collaboration graphs of each period;*
4) *Community detection is made using the INFOMAP[3] algorithm;*
5) *Salient authors are extracted for each community of each period using feature maximization metric;*
6) *Diachronic analysis is applied to monitor community visualizations evolution between periods. JSON report and Gephi visualizations are generated in this step.*

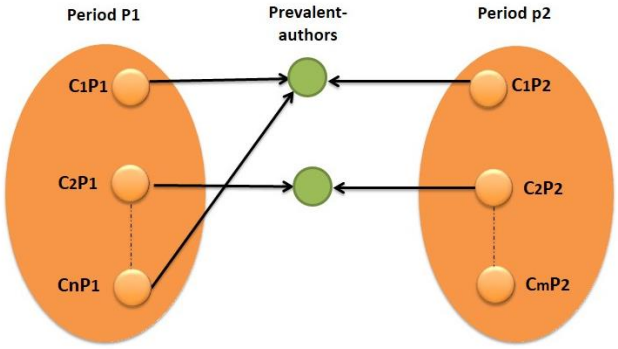

Figure 2 : Diachronic analysis between both time periods using salient authors

For each year, the number of distinct authors varies from

2 Java code that we developed can be found at https://github.com/nicolasdugue/istex/

3 Infomap : http://www.mapequation.org/code.html

more than 1000 to almost 1900. It increases over the year, especially after 2005. Graphs are thus quite small as we can see on Figure 3. Furthermore, authors of two consecutive years are very different. Only 10 to 15 % of authors of one year can be found in the previous or in the next year, which is a low rate if we want to monitor the graph evolution. We thus choose to split our corpora into two sub-corpora, one from 2000 to 2005 and another from 2006 to 2010. Indeed, these two sub-corpora respectively contain 2538 and 5961 distinct authors, and they share 709 authors, which is more than 25 % of the smallest set. We expect these common authors to help us monitoring the evolution of the graphs.

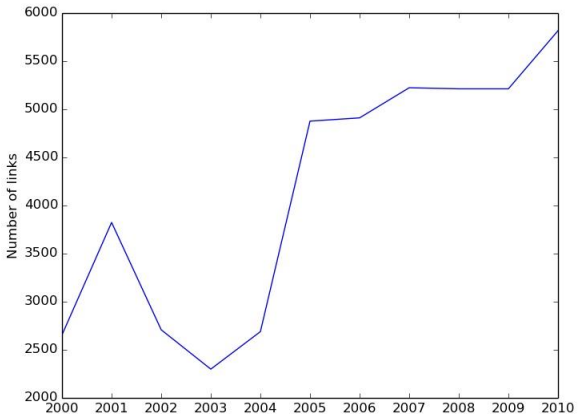

Figure 3 : Number of links of graphs obtained from the sub-corpora of each year period

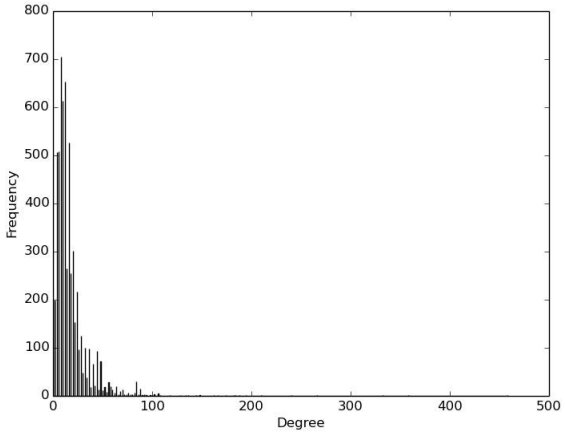

Figure 4 : Degree distribution of graph from the second sub-corpora

The degree distribution of these graphs seem to follow a power-law (Figure 4) with a high number of authors co-publishing one or a few papers, and a few authors co-publishing a high number of papers. We can see a peak around three and four. That is due to the projection of the bipartite author-paper graphs to get the author-author unipartite graph. Authors that wrote a paper with three collaborators automatically have a degree of four.

We then detect communities on both the graphs representing the sub-corpora: the one from 2000 to 2005 and the other from 2006 to 2010. To that aim, we use the INFOMAP algorithm which runs fast and produces high-quality results. INFOMAP produces communities which size follow a power-law like distribution (Figure 5) with a high number of small communities, and a few large ones. The communities constitute a high-level description of the collaboration graphs that is representative of the self-organization of research. To monitor the evolution of these communities, we apply our method.

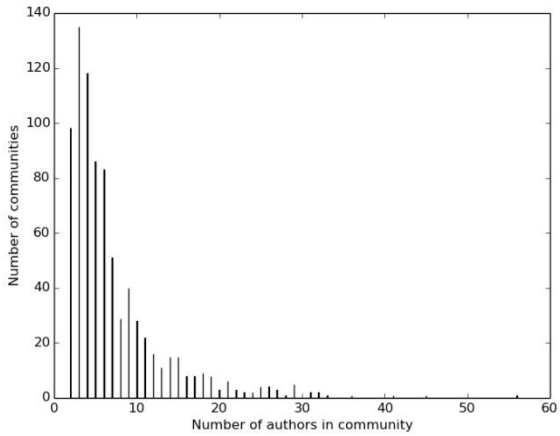

Figure 5 : Communities sizes distribution from the second sub-corpora

Our first main results are about the **nodes matching kernel**. We recall that the **nodes matching kernel** of a community s of the source period and a community t of the target period are the common salient nodes of both communities s and t. In the context of our application, the nodes matching kernel are the salient authors of collaboration communities of both periods. It is thus particularly interesting to consider these authors. As we can see in Figure 6, these authors are essentially contained in the biggest communities detected. Indeed, when the community sizes seems to follow a power-law and thus be made of mostly small communities, the community size of nodes in the matching kernels is much higher in average.

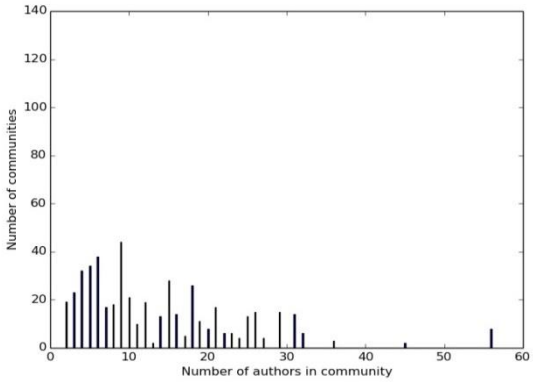

Figure 6 : Sizes distribution of the nodes matching kernel communities

Furthermore, as stated previously, the two periods share 709 authors. Among these 709 authors, 512 (i.e. 72%) of them are part of **nodes matching kernel.** It seems to indicate that salient authors in communities are particularly interesting to monitor. Indeed, they seem to form the backbone of knowledge production across time, linking different periods of time.

Our other results are included in the visualization process of collaboration graphs evolution. Our application produces JSON reports (Figure 8) that are automatically displayed in HTML using Javascript. These JSON visualizations are efficient to monitor the community evolutions through salient nodes. They are complementary from Gephi visualizations that show the graphs collaboration evolution.

In the following example, we can see the salient authors of source community 2 and target community 48 with their associated Feature F-measure in the JSON report. The Gephi visualization (Figure 7) presents in green the kernel nodes, in blue the nodes from the source period and in red, nodes from the target period. We can observe the strength of links that exists between the kernel nodes. They seem to constitute the backbone of the community.

VII. Conclusion

We have presented an original approach for the diachronic analysis of collaboration graphs to monitor research self-organization evolution. The originality of our CGEM approach comes from the fact that the analysis is the combination of a community detection method and a feature selection process. The preliminary results highlight the relevance of using the feature selection process in this graph context. Indeed, salient authors in the communities of the collaboration graphs seem to constitute the backbone of the communities, and they are also bridges between communities of the different periods. The efficient Javascript and graph-oriented visualization solutions allow non-experts to identify easily salient authors and to observe the network organizations around them across time.

This preliminary study has interesting perspectives. First, it would be interesting to confirm the observations about kernel nodes in other datasets. Second, these observations emphasize the relevance of feature selection, even in a graph context. It would be very interesting to investigate more the graph properties of salient nodes, in terms of graph and community centralities, information diffusion, and community roles.

Acknowledgment

ISTEX receives assistance from the French state managed by the National Research Agency under the program "Future Investments" bearing the reference ANR-10-IDEX-0004-12.

We thank R. Loth, G. Guibon and E. Morale for their help during the constitution of the corpus.

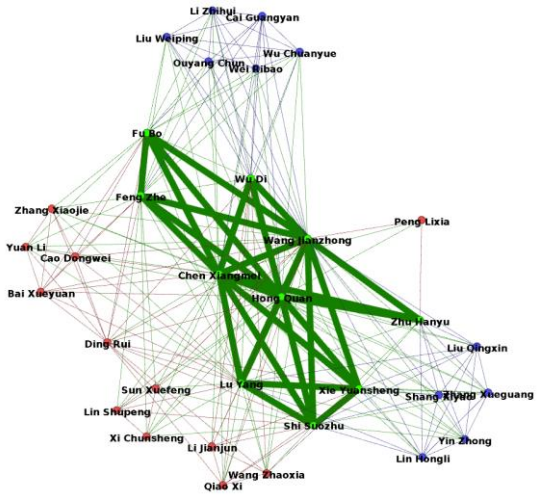

Figure 7 : A graph showing the evolution of a community: the nodes matching kernel is in green, the nodes from the source period in red and the nodes from the target period in blue

Cluster Source	2
Cluster Target	48
Kernel Labels	
label	Chen Xiangmei
fSource	0.18604651
fTarget	0.15384616
label	Feng Zhe
fSource	0.10909091
fTarget	0.080000006
label	Fu Bo
fSource	0.062111802
fTarget	0.080000006
label	Hong Quan
fSource	0.18604651
fTarget	0.15384616
label	Shi Suozhu
fSource	0.10909091
fTarget	0.080000006
label	Wang Jianzhong
fSource	0.18604651
fTarget	0.15384616
label	Wu Di
fSource	0.062111802
fTarget	0.080000006
label	Xie Yuansheng
fSource	0.062111802
fTarget	0.080000006
Common Labels prevalent in Source	
label	Lu Yang
fSource	0.16470589
fTarget	0.080000006
Common Labels prevalent in Target	
label	Zhu Hanyu
fSource	0.025316456
fTarget	0.080000006

Figure 8 : JSON report displayed with Javascript – A community matching between community 2 in source period and community 48 in target period

References

[1] R., Lambiotte and P., Panzarasa, "Communities, knowledge creation, and information diffusion", *Journal of Informetrics 3*, pp. 180–190, 2009.

[2] B., Uzzi and S., Jarrett, "Collaboration and creativity: The small world Problem", *American journal of sociology, 111.2*, pp. 447–504, 2005.

[3] R.S, Burt, "Structural holes and good ideas", *American journal of sociology, 110.2*, pp. 349–399, 2004.

[4] M,, Girvan and M.E.J., Newman, "Community structure in social and biological networks", *Proceedings of the national academy of sciences*, vol. 99, no. 12, pp. 7821-7826, 2002.

[5] K., Borner and C., Chen and K..W., Boyack, "Visualizing knowledge domain", *Annual review of information science and technology, 37.1*, pp. 179–255, 2003.

[6] K.W., Boyack and R.K., Lavans and K., Borner, "Mapping the backbone of science", *Scientometrics, 64.3*, pp. 351–374, 2005.

[7] A., Lancichinetti and S., Fortunato, "Community detection algorithms: a comparative analysis", *Physical review E*, vol. 80, no. 5, 2009.

[8] J.-C., Lamirel, "A new approach for automatizing the analysis of research topics dynamics: application to optoelectronics research", Scientometrics, pp. 151-166, October 1st, 2012.

[9] J.-C., Lamirel and R, Mall and P., Cuxac and G., Safi, "Variations to incremental growing neural gas algorithm", *Proceeding Of International Joint Conference On Neural Networks*, 2011.

[10] A.S., Chivukula and J.-C., Lamirel, "Incremental Novelty Detection applied to Complex Text Classification", *EGC 2013 CIDN Workshop*, 2013.

[11] J.-C., Lamirel and P. Cuxac and K., Hajlaoui and A.S., Chivukula, "A new feature selection and feature contrasting approach based on quality metric: application to efficient classification of complex textual data", *Proceedings of PAKDD* , 2013.

[12] A.L, Porter and I., Rafols, "Is science becoming more interdisciplinary? Measuring and mapping six research fields over time". *Scientometrics*, vol. 81, no. 3, pp. 719-745, 2009.

[13] H., Sayama and J., Akaishi, "Characterizing Interdisciplinarity of Researchers and Research Topics Using Web Search Engines", *Plos One*, vol. 7, no. 6, 2012.

[14] S., Havre and al., "ThemeRiver: visualizing thematic changes in large document collections", IEEE Transactions on Visualization and Computer Graphics, vol. 8, n° 1, 2002.

[15] C., Erten and P.J., Harding and G., Kobourov and K., Wampler and G., Yee, "Exploring the computing literature using temporal graph visualization", *Report, Department of Computer Science, University of Arizona*, 2003.

[16] C.C., Chen and Y.T, Chen and Y.S., Sun and M.C., Chen, "Life Cycle Modeling of News Events Using Aging Theory", *ECML*, pp. 47-59, 2003.

[17] A., Dubey and Q., Ho and S., Williamson and E.P., Xing, "Dependent nonparametric trees for dynamic hierarchical clustering", *NIPS 2014: 1152-1160.*, 2014.

[18] L., Ladha and T., Deepa, "Feature selection methods and algorithms", *International Journal on Computer Science and Engineering*, vol. 3, no. 5, pp. 1787–1797, 2011.

[19] V., Bolón-Canedo and N., Sánchez-Maroño and A., Betan, "A Review of Feature Selection Methods on Synthetic Data", *Knowledge and Information Systems*, pp. 1-37, mars 1, 2012.

[20] I., Guyon and A., Elisseeff, "An introduction to variable and feature selection.", *The Journal of Machine Learning Research 3*, vol. 3, pp. 1157–1182, 2003.

[21] H., Daviet, "Class-Add, une procédure de sélection de variables basée sur une troncature k-additive de l'information mutuelle et sur une classification ascendante hiérarchique en pré-traitement.", *PhD*, 2009.

[22] J.-C., Lamirel and S., Al Shehabi and C., François and M., Hoffmann, "New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping", Scientometrics, vol. 60, no. 3, 2004.

[23] S., Al Shehabi and J.-C., Lamirel., "Inference Bayesian Network for Multi-topographic neural network communication: a case study in documentary data", *International Conference on Information and Communication Technologies: from Theory to Applications - ICTTA*, 2004.

[24] A., Tebbakh, "Network of semantic wikis (Wicri) and data curation", *4th International Conference ISKO Maghreb*, 2014, Algeria.