Analysis of evolutions and interactions between science fields : the cooperation between feature selection and graph representation

Pascal Cuxac, Jean-Charles Lamirel

Analysis of evolutions and interactions between science fields: the cooperation between feature selection and graph representation

Pascal CUXAC[1] and Jean-Charles LAMIREL[2]

[1] *pascal.cuxac@inis.fr*
CNRS-INIST, Vandoeuvre lès Nancy, France

[2] *jean-charles.lamirel@loria.fr*
LORIA-Synalp, Vandoeuvre lès Nancy, France

Abstract

This paper presents the application of a new method of feature selection for the analysis of the evolution and interaction of scientific domains. The query of bibliographic databases provides a corpus of scientific publications in different fields. Every scientific field is considered as a class issued from a machine learning process, either supervised or unsupervised, and each document is represented by a bag of words. Therefore, it is possible to select the most significant words of every class (domain). We then represent the words-classes relationships by a graph whose edges are weighted by a function of contrast. Thus we highlight the specific words for each area and those that are multidisciplinary. In addition, the joint analysis of two periods of time allows us to appreciate the evolution of the scientific fields.

Introduction

On the one hand, 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 textual information which is varying over time.

The purpose of the analysis and diachronic mapping is to track, for a given domain, changes in contexts (sub-themes) and the evolution of vocabularies and actors that materialize these changes in terms of appearances, disappearances, divergence or convergence. The applications relate to very various and highly strategic domains, including web mining, technological and scientific survey.

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

1. performing static classifications at different periods of time and analyze changes between these periods (time step approach or diachronic analysis);
2. developing methods of classification that can directly track the changes: incremental clustering methods (incremental clustering) and novelty detection methods (incremental supervised classification).

On the other hand, the concept of transdisciplinarity is often discussed in conjunction with its facets that are interdisciplinarity, multidisciplinarity, pluridisciplinarity (Zaman and Goschin 2010). As noted by Alvargonzalez (Alvargonzalez, 2011), the terms multidisciplinarity, interdisciplinarity, and transdisciplinarity are often used interchangeably. However these concepts can be defined as follows (Do Espirito Santo 1979):

- Interdisciplinarity: interaction among different disciplines;
- Multidisciplinarity: juxtaposition of various disciplines (without apparent connection between them);
- Pluridisciplinarity: juxtaposition of various disciplines more or less related;
- Transdisciplinarity: a common system for a set of disciplines.

Many authors are interested in graph representation for estimating the interdisciplinarity of science. The goal is usually to determine whether a paper (or journal) is "interdisciplinary" or not. Using subject categories of Currents Contents, Adams et al. (Adams, Jackson and Marshall, 2007) define "interdisciplinarity index" based on the cited references and the "Shannon diversity index". Similarly, Porter et al. (Porter and Rafols, 2009), using the subject categories of Web of Science, define the NAFKI interdisciplinarity metrics and relies on the representation method developed by Leysdesdorff (Leydesdorff 2007). On their own side, Leysdesdorff and al. (Leydesdorff and Rafols, 2009) uses the ISI subject categories included in Science Citation Index and builds graphs using citing and cited dimensions, in order to map different scientific fields. The calculation of "betweenness centrality" allows to measure interdisciplinarity (Leydesdorff, 2007).

Van Raan (Van Raan, 2000) presents the interdisciplinary nature of science as an interaction of socio-economic problems, scientifically interesting problems, and interdisciplinarity. He uses bibliometric approaches to highlight the interdisciplinarity.

Some work focuses on authors: Schummer (Schummer, 2004) is interested in collaboration between researchers (or institutions) for addressing the multidisciplinary field in nanoscience. Klein emphasizes that identifying experts is crucial, because they form an interdisciplinary appropriate epistemic community (Klein, 2008).

We present hereafter an original word-based approach using feature maximization metric (Lamirel and al. 2013) in order to detect significant differences between two time periods for the same scientific field, but also to detect transdisciplinary terms that are markers of cross-domain scientific collaborations. We show that our approach is also applicable to the authors (i.e. actors) allowing quickly highlighting those that are "bridges" between scientific fields.

Unlike common approaches based on graph analysis (Porter and Rafols 2009) (Sayama and Akaishi 2012), we are tackling the problem using a classification of documents (in scientific fields) in combination with a selection of features (index keywords) associated with document classes. Only then, we construct a graph visualizing the interaction between keywords and classes using links weighted by values of contrast defining the strength of the relation between these latter. Feature selection and links contrasting are based on the feature maximization metric (F-max) that has been already successfully used in an unsupervised context.

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 becomes thus 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. (Ladha and Deepa, 2011), Bolón-Canedo and al. (Bolón-

Canedo, Sanchez-Maroño and Alonso-Betanzos, 2012), Guyon and al. (Guyon and Elisseeff, 2003) or Daviet (Daviet, 2009). For an overview of these methods, you might refer to the previous articles, as well as to (Lamirel and al. 2013).

Our approach

To clarify the principle of our approach, which we named GRAFSEL, we follow four steps that are schematically presented in Figure 1:

- We query a bibliographic database (PASCAL is used in our context) to produce corpora for each selected scientific fields (following classification codes or subject category for example) and/or time period;
- The bibliographic records of each corpus are assigned to a class that represents the scientific field and/or the period. Doing so, we build up a classification mixing topics and time periods;
- The records being represented their associated keywords, we select keywords related to each scientific field and/or period and compute the strength of the relations (i.e. contrast) between keywords and scientific fields and/or periods exploiting the feature maximization metric (F max) shortly described after;
- The last step is to build the graph highlighting the relationships between the fields and the keywords by weighting the links of the graph with the formerly obtained contrast values.

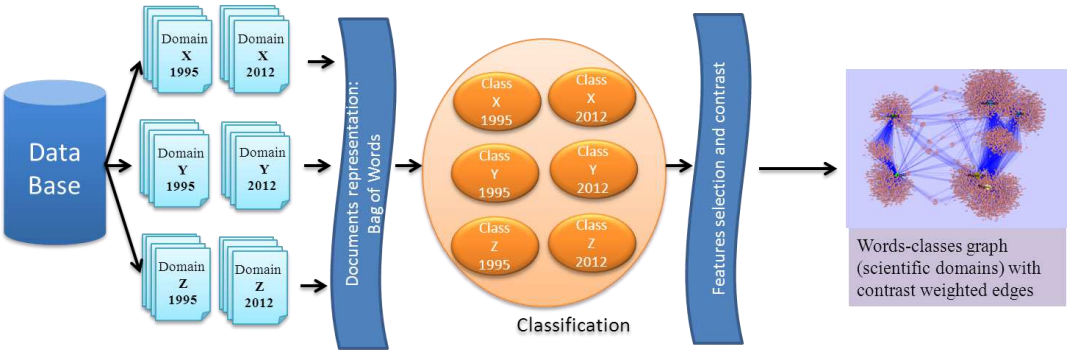

Fig. 1: Principle of the GRAFSEL approach.

Feature maximization for variable selection

- Feature maximization metric principles in unsupervised learning

Feature maximization (F-max) is an unbiased cluster quality metrics that exploits the properties of the data associated to each cluster without prior consideration of clusters profiles. This metrics has been initially proposed in (Lamirel and al 2004). Its main advantage is to be independent altogether of the clustering methods and of their operating mode.

Consider a set of data *D* represented with a set of features *F*, and a set of clusters *C* resulting from a clustering method. Feature maximization promotes clusters with maximum *Feature F-measure*. The *Feature F-measure* $FF_c(f)$ of a feature *f* associated to a cluster *c* is defined as the harmonic mean of *Feature Recall* $FR_c(f)$ and *Feature Precision* $FP_c(f)$ indexes which in turn are defined as:

$$FR_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{c' \in C} \sum_{d \in c'} W_d^f}, FP_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{f' \in F^c, d \in c} W_d^{f'}} \quad (1)$$

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

where W_d^f represents the weight of the feature *f* for data *d* and F_c represent the set of features occurring in the data associated to the cluster *c*.

Two important application of the feature maximization metric are related to the estimation of the overall clustering quality and to incremental clustering (Lamirel, 2012).

- *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 feature selection in the context of supervised learning becomes a straightforward process, as soon as this generic metric can apply on data associated to a class as well as to those associated to a cluster. The feature maximization-based selection process can thus be defined as a parameter-free class-based process in which a class feature is characterized using both its capacity to discriminate a given class from the others ($FP_c(f)$ index) and its capacity to accurately represent the class data (FR_c(f) index).

The set S_c of features that are characteristic of a given class *c* belonging to an overall class set *C* results in:

$$S_c = \{f \in F_C \mid FF_c(f) > \overline{FF}(f) \; and \; FF_c(f) > \overline{FF}_D\} \quad (3)$$

where $\overline{FF}(f) = \sum_{c' \in C} FF_{c'}(f) / \left|C_{/f}\right|$ and $\overline{FF}_D = \sum_{f \in F} \overline{FF}(f) / |F|$.

and $C_{/f}$ represents the restriction of the set *C* to the classes in which the feature *f* is represented.

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

$$S_C = \bigcup_{c \in C} S_c \quad (4)$$

Features that are judged relevant for a given class are the features whose representation is altogether better than their average representation in all the classes including those features and better than the average representation of all the features, as regard to the feature F-measure metric.

In the specific framework of the feature maximization process, a contrast enhancement step can be exploited complementary to the former feature selection step. The role of this step is to fit the description of each data to the specific characteristic of its associated class which have been formerly highlighted by the feature selection step (Lamirel and al. 2013). In the case of our metric, it consists in modifying the weighting scheme of the data specifically to each class by taking into consideration the information gain provided by the *Feature F-measures* of the features, locally to that class.

Thanks to the former strategy, the information gain provided by a feature in a given class is proportional to the ratio between the value of the *Feature F-measure* of this feature in the class and the average value of the *Feature F-measure* of the said feature on all the partition.

Experimental results

- *The datasets*

We present the first results obtained from a corpus of bibliographic records extracted from the PASCAL database. Our experimental corpus included documents from the following six scientific fields: Physics, Geology, Electronics, Medicine (diagnosis techniques), Information Science and Linguistics, for the years 1995 and 2012 (respectively 61 109 and 64 036 scientific papers), distributed as shown in table 1.

Tab. 1: Number of records by scientific fields and years

Domain	**1995**	**2012**
Electronics	11906	10414
Geology	16549	17467
Information science	2747	3211
Linguistic	2871	4441
Medicine (diagnostic techniques)	11336	10673
Physics	21700	17830

- *The results*

The F-max feature selection method implemented here allows significantly reducing the number of features (words) in order to keep only the most important words (representative) of each class (Tab. 2). Of course, we thus eliminate words that could be found in several classes (domains), but our goal is to focus on the "words of specialties" by forgetting the more general words such as "analysis", "study", "method" or "model".

Tab. 2: Number of keywords before and after F-max feature selection

Domain	**Original keywords**	**Selected keywords**	**(%) selected**
Electronic_1995 (E15)	12813	1273	9.94
Electronic_2012 (E12)	11706	2193	18.73
Geology_1995 (G95)	16124	1613	10.00
Geology_2012 (G12)	13768	1856	13.48
Information science_1995 (S95)	4915	1163	23.66
Information science_2012 (S12)	2338	365	15.61
Linguistic_1995 (L95)	20186	322	1.59
Linguistic_2012 (L12)	29093	322	2.72
Medicine_1995 (M95)	10138	886	8.74
Medicine_2012 (M12)	10326	1037	10.04
Physics_1995 (P95)	12268	1051	8.57
Physics_2012 (P12)	15397	1894	12.30

Figure 2 shows the results obtained by taking into account altogether the whole corpus and the two time periods considered. For reasons of readability, we separate the main groups. All of the following graphs are obtained with a force-directed algorithm (spring algorithm).

After having computed a global graph showing general interactions between the fields, we have then separated the fields into two main groups in which strong interaction and/or evolution are more liable to occur.

Figure 3 shows the graph obtained with all the selected fields of the second group which includes Physics, Electronics and Medicine.

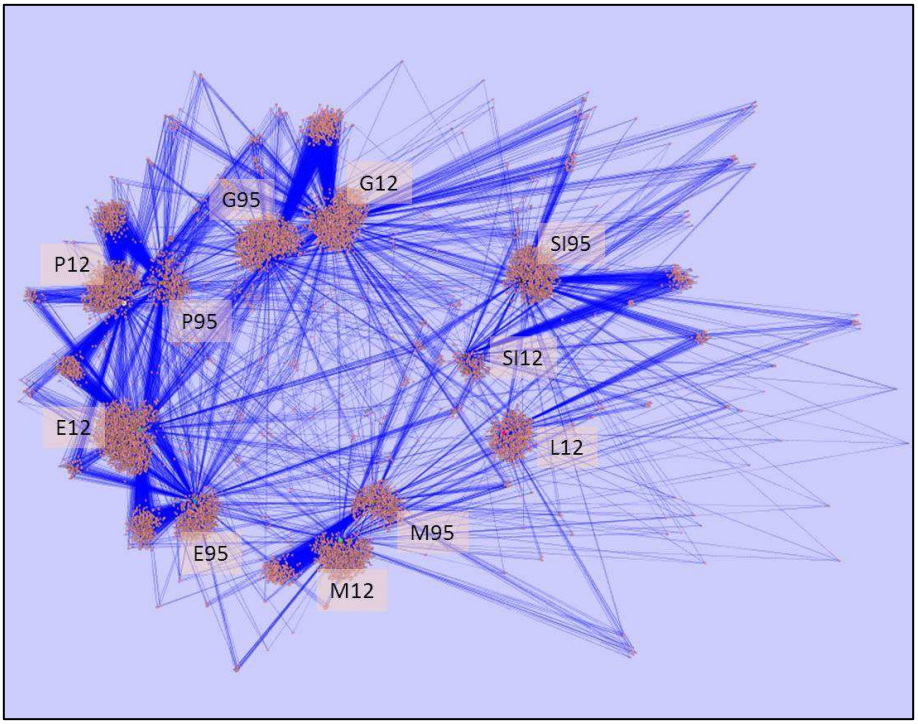

Fig. 2: Global words-classes graph (G=Geology;P=Physics;E=Electronics;M=Medicine;S=Information-Science; L=Linguistics;12=2012;95=1995).

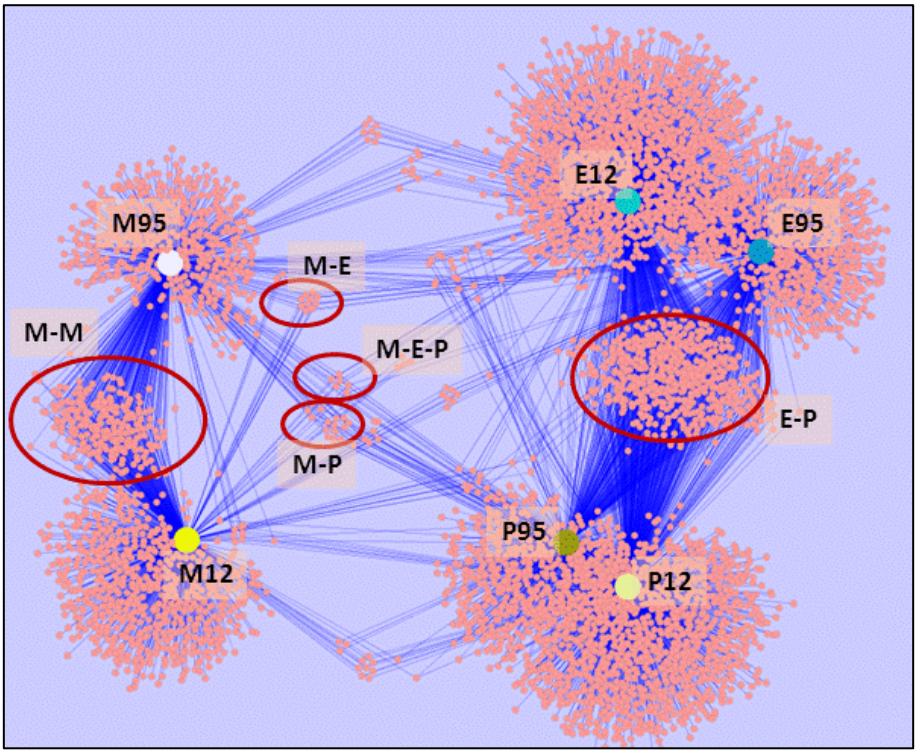

Fig. 3: Words-classes graph for a subgroup of 3 scientific fields (P=Physics; E=Electronics; M=Medicine; 12=2012; 95=1995).

Looking to such graph, we must keep in mind that all phenomena that we visualize describe the corpus as a whole and cannot thus be interpreted in an absolute way by isolating a class from the others, because the nodes (keywords) of the graph are issued from the overall classification of the data (it is the same for the weights of the links). Referring to figure 2, we can however observe several interesting scenarios occurring in a parallel way in our experimental data:

- **Evolution of scientific fields**: the resulting graph highlights that, compared to other fields, the field of Medicine (and especially the "technical diagnosis" discipline) individualizes well into two groups for each analyzed period. Indeed, on the one side, we can observe a dense cloud of keywords (M-M) that are common to both periods but have a relatively weak link with the M95 and M12 classes. On the other side, keywords like "Sleep Disorder", "Hypercholesterolemia", "Forensic aspect", "Radiosurgery" are specific terms in 2012 (i.e. belong to M12 class), while keywords like "Arteriography", "Angina pectoris", "Heart valve", "Lymphocytic leukemia" likely belong to 1995 (M15 class). In an alternative way, keywords clouds related to Physics (and to a lesser extent the ones related to Electronics) remains homogeneous, whatever is the considered period, indicating less important temporal changes;
- **Transdisciplinarity**: there are small groups of keywords with share relationship with several different main clouds. These are transdisciplinary vocabularies reflecting cooperation between scientific fields or practical applications of new technologies: M-E keywords linking Medicine to Electronics, M-P keywords linking Medicine and Physics, E-P keywords linking Electronics and Physics, and finally, M-E-P keywords linking the three fields.

Conclusion

We have shown from a simple example that our fully unsupervised method allows a non-expert user to view the terms used in various disciplines and their temporal evolution. In a complementary way, with such method, it is also easy for the said user to distinguish the terms common to several topics. Depending on the case, the original graph might be too dense for clear viewing, but it is then easy to select a subgraph on which the analysis can be conducted. The originality of our GRAFSEL approach comes from the fact that the nodes of the graph result of the combination of a classification and a feature selection processes, which are applied in a first step, and the links of the graph result from a feature contrasting process, which is applied in a second step on the selected features (i.e. nodes). Hence, unlike the commonly used methods, we do not build here a graph of words but a graph of relationships between words and classes, each class representing a main domain under consideration. In such a way, another interesting application of our approach may be the detection of the authors which represent "bridges" between scientific fields. In figure 4, we see that in 18 years the landscape of the authors has been considerably renewed and, in addition, we can identify groups of people who are "**Transmitters of Knowledge**" between scientific domains or between periods of the same domains.

Last but not least, supervised as well as unsupervised classification methods can be used in this process.

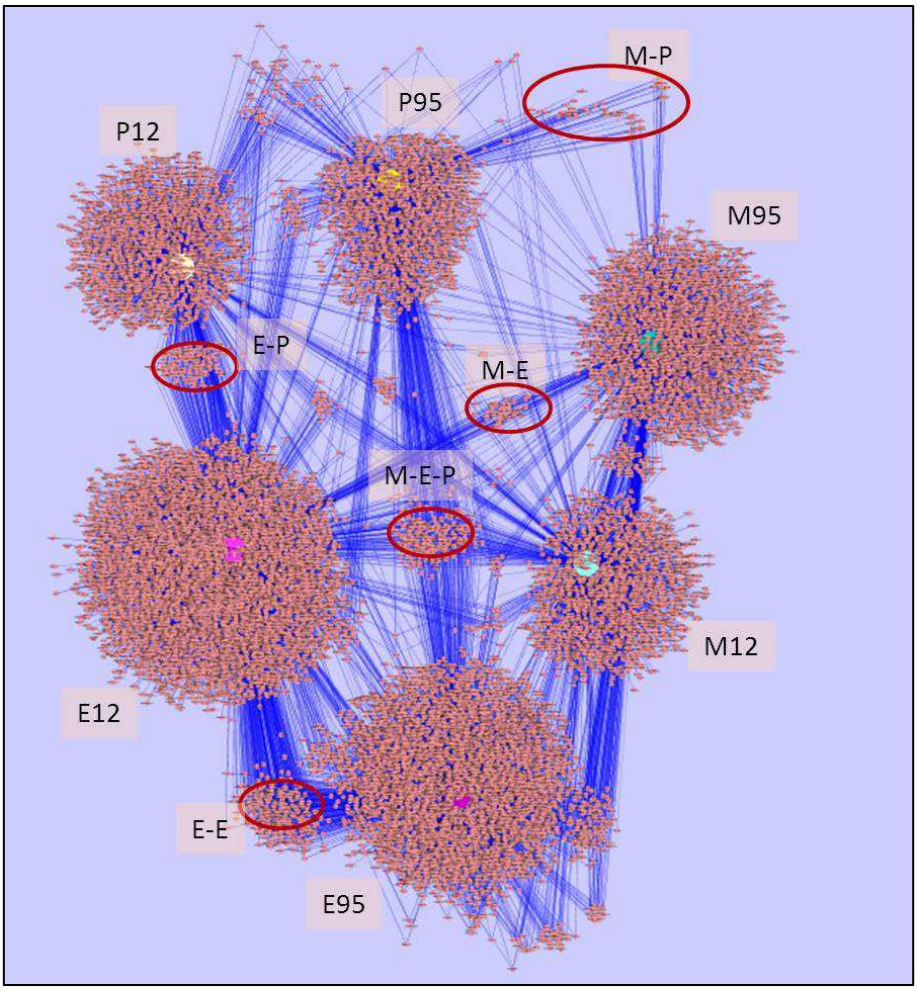

Fig. 4: Authors-classes graph for a subgroup of 3 scientific fields (P=Physics; E=Electronics; M=Medicine; 12=2012; 95=1995).

References

Adams, J., Jackson, L., Marshall, S., & Evidence Ltd. (2007). *Bibliometric analysis of interdisciplinary research: Report to the Higher Education Funding Council for England*. Leeds: Evidence.

Alvargonzález D. (2011): Multidisciplinarity, Interdisciplinarity,Transdisciplinarity, and the Sciences, *International Studies in the Philosophy of Science*, 25:4, 387-403

Bolón-Canedo, V., Sánchez-Maroño N. & Alonso-Betanzos A. (2012). A Review of Feature Selection Methods on Synthetic Data. Knowledge and Information Systems (mars 1, 2012): 1-37

Daviet, H. (2009). 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 Université de Nantes, France, 2009.

Do Espirito Santo, D. (1979) Contemporary concepts of interdisciplinarity Contemporary concepts of interdisciplinarity , *Semina Ciências Agrárias*, vol. 1, n° 3, 1979

Guyon, I. & Elisseeff A.(2003). An introduction to variable and feature selection. The Journal of Machine Learning Research 3 (2003): 1157–1182.

Klein J. T. (2008) Evaluation of Interdisciplinary and Transdisciplinary Research: A Literature Review. American journal of preventive medicine, August 2008, vol.35, issue 2 Pages S116-S123

Ladha, L. & Deepa T.(2011). Feature selection methods and algorithms. International Journal on Computer Science and Engineering, 3, nº 5 (2011): 1787–1797.

Leydesdorff, L. (2007) Betweenness centrality as an indicator of the interdisciplinarity of scientific journals, *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, nº 9, p. 1303–1319, 2007

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. Journal of the American Society for Information Science and Technology, 60(2), 348-362.

Lamirel, J.-C. (2012). A new approach for automatizing the analysis of research topics dynamics: application to optoelectronics research Scientometrics (2012) 93: 151-166 , October 01, 2012

Lamirel, J.-C., Al Shehabi, S., Francois, C. & Hoffmann, M. (2004). New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping, *Scientometrics*, vol. 60, n° 3, 2004.

Lamirel J.C., Cuxac P., Hajlaoui K. Chivukula A.S. (2013) A new feature selection and feature contrasting approach based on quality metric: application to efficient classification of complex textual data. Proceedings of *PAKDD 2013,International Workshop on Quality Issues, Measures of Interestingness and Evaluation of Data Mining Models (QIMIE)*, GoldCoast, Australia, April 2013

Porter, A. L. & Rafols, I. (2009) Is science becoming more interdisciplinary? Measuring and mapping six research fields over time, *Scientometrics*, vol. 81, nº 3, p. 719-745, 2009

Sayama, H. &Akaishi, J. (2012) Characterizing Interdisciplinarity of Researchers and Research Topics Using Web Search Engines, *Plos One*, vol. 7, nº 6, p. e38747, 2012

Schummer, J. (2004) Multidisciplinarity, Interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology. Scientometrics 59, 425-465.

Van Raan, A.F.J. The interdisciplinary nature of science: theoretical framework and bibliometric-empirical approach. In: P.Weingart and N.Stehr (eds.). pp.66-78. Practising Interdisciplinarity. Toronto: University of Toronto Press, 2000.

Zaman, G. & Goschin, Z. (2010) Multidisciplinarity, Interdisciplinarity and Transdisciplinarity: Theoretical Approaches and Implications for the Strategy of Post-Crisis Sustainable Development, *Theor. Appl. Econ.*, vol. XVII, nº 12(553), p. 5-20, 2010