Analyse des évolutions et des interactions entre domaines scientifiques: GRAFSEL, association de la sélection de variables et de la représentation graphique

Pascal Cuxac, Jean-Charles Lamirel

Analyse des évolutions et des interactions entre domaines scientifiques: GRAFSEL, association de la sélection de variables et de la représentation graphique

Pascal CUXAC (*),**Jean-Charles LAMIREL** (**),
pascal.cuxac@inist.fr, jean-charles.lamirel@loria.fr

(*) CNRS-Inist, Vandœuvre-Lès-Nancy, France,
(**) LORIA-Synalp, Vandœuvre-Lès-Nancy, France

Mots clefs :
Veille scientifique et technologique, gestion des connaissances, classification supervisée, sélection de variables, graphe
Keywords:
Scientific and technical observation, knowledge management, classification, features selection, graph
Palabras clave :
Escudriñar científico y tecnológico, administración del conocimiento

Résumé

Cet article présente l'application d'une nouvelle méthode de sélection de variables pour l'analyse de l'évolution et des interactions entre domaines scientifiques. L'interrogation de bases de données bibliographiques fournit un corpus de publications scientifiques dans différents domaines. Chaque domaine scientifique est considéré comme une classe obtenue à partir d'un processus d'apprentissage automatique, qu'il soit supervisé ou non, et chaque document est représenté par un sac de mots. Il est alors possible de sélectionner les mots les plus significatifs de chaque classe (domaine). Nous représentons ensuite les relations mots-classes par un graphe dont les arêtes sont pondérées par une fonction de contraste. Cette méthode nous permet de discriminer entre les mots spécifiques à chaque domaine et ceux qui sont pluridisciplinaires. En outre, l'analyse conjointe de plusieurs périodes de temps nous permet également d'apprécier parallèlement l'évolution des domaines scientifiques.

1 Introduction

Le développement de méthodes d'analyse d'informations dynamiques, comme le clustering incrémental et les techniques de détection de nouveauté, devient une préoccupation centrale dans beaucoup d'applications dont l'objectif principal est de traiter un grand volume d'informations textuelles variant au fil du temps.

Le but de l'analyse et de la cartographie diachronique est de suivre, pour un domaine donné, les changements de contextes (sous-thèmes) et l'évolution des vocabulaires et des acteurs qui se matérialise en termes d'apparitions, de disparitions, de divergences ou de convergences. Les applications concernent des domaines très divers et hautement stratégiques, tels que l'exploration du Web et la veille technologique et scientifique.

Afin d'identifier et d'analyser l'émergence, ou de détecter des changements dans les données, nous avons déjà proposé deux approches différentes et complémentaires:

1. réalisation de classifications statiques à différentes périodes de temps et analyse des changements entre ces périodes (approche par pas de temps ou analyse diachronique);
2. développement de méthodes de classification qui peuvent suivre directement les évolutions : méthodes de clustering incrémentales et méthodes de détection de nouveauté (classification supervisée incrémentale).

Ces méthodes ne prennent cependant pas en compte le phénomène important, à savoir celui de la transdisciplinarité. Le concept de transdisciplinarité est souvent discuté de manière corrélative par de multiples facettes comme l'interdisciplinarité, la multidisciplinarité, la pluridisciplinarité (Zaman et Goschin 2010). En effet, comme l'a noté Alvargonzalez (Alvargonzalez, 2011), les termes multidisciplinarité, interdisciplinarité et transdisciplinarité sont souvent utilisés de manière interchangeable. Plus précisément, ces concepts peuvent être définis comme suit (Do Espirito Santo 1979):

- Interdisciplinarité: interaction entre les différentes disciplines;
- Multidisciplinarité: juxtaposition de différentes disciplines (sans lien apparent entre elles);
- Pluridisciplinarité: juxtaposition de différentes disciplines plus ou moins connexes;
- Transdisciplinarité: système commun pour un ensemble de disciplines.

De nombreux auteurs se sont intéressés à la représentation graphique pour estimer l'interdisciplinarité de la science. Le but est souvent de déterminer si un document (ou un journal) est «interdisciplinaire» ou non. Utilisant les codes de classification, ou «subjects categories», des «Currents Contents», Adams et al. (Adams, Jackson et Marshall, 2007) définissent «l'indice d'interdisciplinarité» sur la base des références citées et sur l'indice de diversité de Shannon. De même, Porter et al. (Porter et Rafols, 2009), en utilisant les codes de classification du Web of Science, définissent la métrique de l'interdisciplinarité NAFKI s'appuyant sur la méthode de représentation développée par Leysdesdorff (Leydesdorff 2007). De leur côté, Leysdesdorff et al. (Leydesdorff et Rafols, 2009) utilisent les codes de classification ISI inclus le «Science Citation Index» et construisent des graphes en utilisant les dimensions citées et citantes, afin de cartographier les différents domaines scientifiques. Le calcul de la centralité d'intermédiarité (betweenness centrality) dans ces graphes leur permet de mesurer l'interdisciplinarité (Leydesdorff, 2007).

Van Raan (Van Raan, 2000) présente la nature interdisciplinaire de la science comme une interaction de problèmes socio-économiques, de problèmes scientifiquement intéressants, et de l'interdisciplinarité. Il utilise des méthodes bibliométriques pour mettre en évidence l'interdisciplinarité.

Certains travaux se concentrent sur les auteurs: Schummer (Schummer, 2004) s'intéresse à la collaboration entre les chercheurs (ou institutions) pour aborder le domaine multidisciplinaire en nanosciences. Klein souligne que l'identification d'experts est cruciale, car ils forment une communauté épistémique interdisciplinaire appropriée (Klein, 2008).

Nous présentons ici une approche originale basée sur les mots en utilisant la métrique de maximisation des variables (Lamirel et al. 2013) afin de détecter des différences significatives entre les deux périodes pour le même domaine scientifique, mais aussi de détecter les termes transdisciplinaires qui sont des marqueurs de collaborations scientifiques entre différentes disciplines. Nous montrons que notre approche est également applicable aux auteurs (acteurs) permettant de mettre rapidement en évidence ceux qui sont des «passerelles» entre les disciplines scientifiques.

Contrairement aux approches communes fondées sur l'analyse de graphes (Porter et Rafols 2009) (Sayama et Akaishi 2012), nous abordons le problème en utilisant une classification des documents (en domaines scientifiques) en combinaison avec une sélection de variables (mots-clés d'indexation) associée aux classes de documents. Seulement alors, nous construisons un graphe permettant de visualiser l'interaction entre les mots clés et les catégories à l'aide des liens pondérés par des valeurs de contraste définissant la force de la relation. La sélection des variables et l'établissement du contraste des liens sont basés sur la métrique de maximisation des variables (F-max) qui a déjà été utilisée avec succès, à la fois dans le contexte de la classification non supervisé et dans celui de la classification supervisée.

2 La sélection des variables

Depuis les années 1990, les progrès de l'informatique et des capacités de stockage permettent la manipulation de très gros volumes de données: il n'est pas rare d'avoir des espaces de description de plusieurs milliers, voire de dizaines de milliers de variables. On pourrait penser que les algorithmes de classification sont plus efficaces avec un grand nombre de variables. Toutefois, la situation n'est pas aussi simple que cela. Le premier problème qui se pose est l'augmentation du temps de calcul. En outre, le fait qu'un nombre important de variables soient redondantes ou non pertinentes pour la tâche de classification perturbe considérablement le fonctionnement des classifieurs. De plus, la plupart des algorithmes d'apprentissage exploitent des probabilités et les distributions de probabilités peuvent alors être difficiles à estimer dans le cas de la présence d'un très grand nombre de variables. L'intégration d'un processus de sélection de variables dans le cadre de la classification des données de grande dimension devient donc un enjeu central. Ceci est d'autant plus vrai qu'il est également nécessaire, pour des raisons de synthèse, de mettre en avant les variables privilégiés lors de la visualisation des résultats de classification.

Dans la littérature, trois types d'approches pour la sélection de variables sont principalement proposés: les approches directement intégrées aux méthodes de classification, dites «embedded», les méthodes basées sur des techniques d'optimisation, dites «wrapper», et finalement, les approches de filtrage. Des états de l'art exhaustifs des différentes techniques ont été réalisés par de nombreux auteurs, comme Ladha et al. (Ladha et Deepa, 2011), Bolón-Canedo et al. (Bolón-Canedo, Sanchez-Maroño et Alonso-Betanzos, 2012), Guyon et al. (Guyon et Elisseeff, 2003) ou Daviet (Daviet, 2009). Pour avoir un aperçu de ces méthodes, le lecteur se réfèrera aux articles mentionnés, ainsi qu'à (Lamirel et al. 2013).

3 Maximisation d'étiquetage pour la sélection de variables

3.1 – Principe de la métrique de maximisation d'étiquetage en apprentissage non supervisé

La maximisation d'étiquetage (F-max) est une métrique non biaisée d'estimation de la qualité d'une classification non supervisée (clustering) qui exploite les propriétés des données associées à chaque cluster sans examen préalable des profils de clusters. Cette métrique a été initialement proposée dans (Lamirel et al 2004). Son principal avantage est d'être tout à fait indépendante des méthodes de classification et de leur mode de fonctionnement.

Considérons un ensemble de données *D* représenté par un ensemble de variables *F*, et un ensemble de clusters *C* résultant d'une méthode de clustering. La métrique de maximisation d'étiquetage favorise les clusters avec une valeur maximale de F-mesure d'étiquetage. La F-mesure d'étiquetage $FF_c(f)$ d'une variable f associée à un cluster *c* est définie comme la moyenne harmonique du rappel d'étiquetage $FR_c(f)$ et de la précision d'étiquetage $FP_c(f)$, eux-mêmes définis comme suit.

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

avec

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

Et où W_d^f représente le poids de la variable f pour les données D et F_c représentent l'ensemble des caractéristiques des données associées au cluster *c*.

Deux applications importantes de la métrique de maximisation d'étiquetage sont liées à l'estimation de la qualité du clustering et au clustering incrémental. Plus de détails sur ces applications sont donné dans la référence (Lamirel, 2012).

3.2 - Adaptation de la métrique de maximisation d'étiquetage pour la sélection de variables en apprentissage supervisé

Tenant compte de la définition de base de la métrique de maximisation d'étiquetage présentée dans la section précédente, son exploitation pour la tâche de sélection de variables dans le contexte de l'apprentissage supervisé devient un processus simple, dès lors que cette métrique générique peut s'appliquer sur des données associées à une classe aussi bien qu'à celles qui sont associées à un cluster. Le processus de sélection peut donc être défini comme un processus non paramétré basé sur les classes dans lesquelles une variable de classe est caractérisée en utilisant à la fois sa capacité à discriminer une classe donnée ($FP_c(f)$ index) et de sa capacité à représenter fidèlement les données de la classe (FR_c(f) index).

L'ensemble S_c des variables qui sont caractéristiques d'une classe donnée c appartenant à un ensemble de classe C se traduit par:

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

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

.

et $C_{/f}$ représente un sous ensemble de C aux classes dans lesquelles la variable f est représentée.

Enfin, l'ensemble de toutes les variables S_C sélectionnées est le sous-ensemble de F défini comme:

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

Les variables qui sont jugées pertinentes pour une classe donnée sont les variables dont les représentations sont meilleures que leurs représentations moyennes dans toutes les classes, et meilleures que la représentation moyenne de toutes les variables, en termes de F-mesure d'étiquetage.

Dans le cadre spécifique du processus de maximisation d'étiquetage, une étape d'amélioration par contraste peut être exploitée en complément de la première étape de sélection. Le rôle de cette étape est d'adapter la description de chaque donnée aux caractéristiques spécifiques de leurs classes associées qui ont été précédemment mises en évidence par l'étape de sélection (Lamirel et al. 2013). Dans le cas de notre métrique, cela consiste à modifier le schéma de pondération des données pour chaque classe en prenant en considération le gain d'information fournie par la F-mesure d'étiquetage des variables, localement à cette classe.

Le gain d'information est proportionnel au rapport entre la valeur de la F-mesure d'une variable dans la classe et la valeur moyenne de la F-mesure de variable dans toute la partition.

Un exemple de l'intérêt de l'exploitation de la sélection de variables basé sur la maximisation d'étiquetage est donné par la tâche d'assistance à la validation des brevets du projet QUAERO. Cette tâche consistait à générer un aide aux experts dans leur tâche d'évaluation de la nouveauté d'un brevet fondée sur l'assignation automatique des papiers scientifiques plus pertinents liés avec les codes de la de classification des brevets. Dès lors que l'apprentissage était basé sur les citations extraites des brevets qui sont habituellement associées à une hiérarchie des codes de classification ayant différents niveaux de généralité, en premier lieu, il n'y avait aucune garantie d'une répartition homogène des citations (c.-à-d. les échantillons d'apprentissage) parmi les codes, en second lieu, il y avait de fortes chances d'avoir des citations similaires dans différentes classes. Cette tâche soulevait donc de nouveaux défis dans le domaine de la classification, en particulier celui de devoir traiter des données très déséquilibrées appartenant à des classes fortement similaires entre elles (Hajlaoui et al. 2013). Elle n'a pu fournir des résultats satisfaisant et exploitables par les experts (une faible confusion entre les classes s'avérait naturellement indispensable dans ce cadre) qu'après l'exploitation de mécanismes de sélection de variable basés sur la maximisation d'étiquetage. En effet, dans ce contexte, cette méthode a permis d'améliorer les performances de la classification de plus de 90%, alors que toute les méthodes concurrentes se sont révélées totalement inopérantes, voir néfastes aux performances (Lamirel et al. 2013).

4 Notre approche

Afin de clarifier le principe de notre approche, que nous avons nommée GRAFSEL, nous suivons les quatre étapes qui sont schématiquement présentés dans la figure 1:

- Nous interrogeons une base de données bibliographiques (PASCAL est utilisé dans les exemples suivants) afin de construire des corpus pour chacun des domaines scientifiques choisis (suivant les codes de classification ou «subjects categories», par exemple) et / ou par période de temps;
- Les notices bibliographiques de chaque corpus sont affectées à une classe qui représente le domaine scientifique et / ou la période. Ce faisant, nous construisons une classification mélangeant les thèmes et les périodes;
- Les notices sont représentées par les mots clés associés, nous sélectionnons ceux liés à chaque domaine scientifique et / ou à la période et nous calculons la force des relations (contraste) entre les mots clés et les domaines scientifiques et / ou les périodes en utilisant la métrique F-max de maximisation d'étiquetage décrite ci-avant;
- La dernière étape est la construction du graphe mettant en évidence les relations entre les domaines scientifiques et les mots clés en pondérant les liens du graphe avec les valeurs de contraste précédemment obtenus.

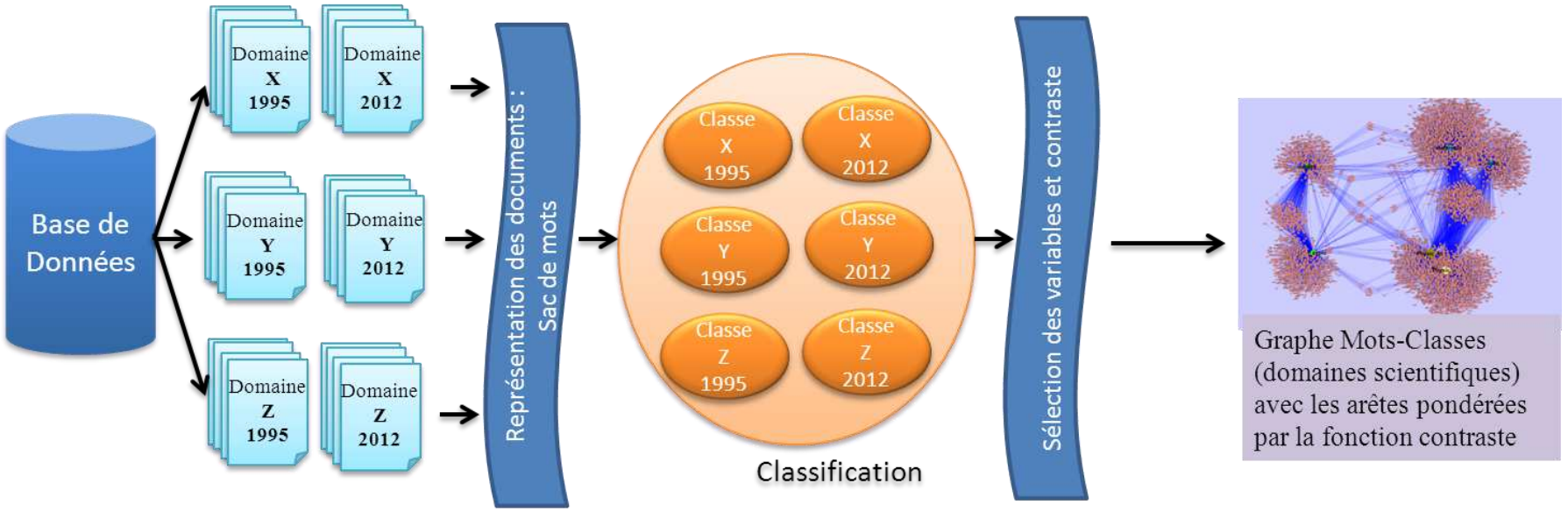

Fig. 1: Principe de l'approche GRAFSEL.

5 Résultats expérimentaux

5.1 - Les ensembles de données

Nous présentons les premiers résultats obtenus à partir d'un corpus de notices bibliographiques extraites de la base de données PASCAL. Notre corpus expérimental comprend des documents provenant des six domaines scientifiques suivants: physique, géologie, électronique, médecine (techniques de diagnostic), sciences de l'Information et linguistique, pour les années 1995 et 2012 (respectivement 61 109 et 64 036 articles scientifiques), répartis comme indiqué dans Tableau 1.

Tab. 1: Nombre de notices par domaines scientifiques et années

Domaine	**1995**	**2012**
Electronique	11906	10414
Géologie	16549	17467
Sciences de l'Information	2747	3211
Linguistique	2871	4441
Médecine (techniques de diagnostic)	11336	10673
Physique	21700	17830

5.2 - Les résultats

La méthode de sélection de variables F-max appliquée ici permet de réduire considérablement le nombre de variables (mots) afin de ne garder que les mots les plus importants (représentant) de chaque catégorie (Tableau 2). Bien sûr, nous éliminons ainsi certains des mots fréquents qui pourraient être trouvés dans plusieurs catégories (domaines), mais notre objectif est de se concentrer sur les «mots de spécialités» en oubliant les mots plus généraux tels que par exemple «analyse», «étude», «méthode» ou «modèle».

La figure 2 montre les résultats obtenus en tenant compte de l'ensemble du corpus et des deux périodes considérées. Pour des raisons de lisibilité, nous séparons les groupes principaux. Tous les graphiques suivants sont obtenus avec un algorithme basé sur les forces («*Force*-based» ou «*Force*-directed algorithms»).

Après avoir calculé un graphe global montrant les interactions entre les domaines, nous avons ensuite séparé les domaines en deux groupes principaux dans lesquels une forte interaction et/ou d'évolution sont plus susceptibles de se produire.

La figure 3 montre le graphe obtenu avec tous les domaines sélectionnés du second groupe qui comprend la physique, l'électronique et la médecine.

Tab. 2: Nombre de mots clés avant et après la sélection de fonction F-max

Domaine	MC Initial	MC sélectionnés	(%) sélectionnés
Electronique_1995 (E15)	12813	1273	9.94
Electronique_2012 (E12)	11706	2193	18.73
Géologie_1995 (G95)	16124	1613	10.00
Géologie_2012 (G12)	13768	1856	13.48
Science Information_1995 (S95)	4915	1163	23.66
Science Information_2012 (S12)	2338	365	15.61
Linguistique_1995 (L95)	20186	322	1.59
Linguistique_2012 (L12)	29093	322	2.72
Medecine_1995 (M95)	10138	886	8.74
Médecine_2012 (M12)	10326	1037	10.04
Physique_1995 (P95)	12268	1051	8.57
Physique_2012 (P12)	15397	1894	12.30

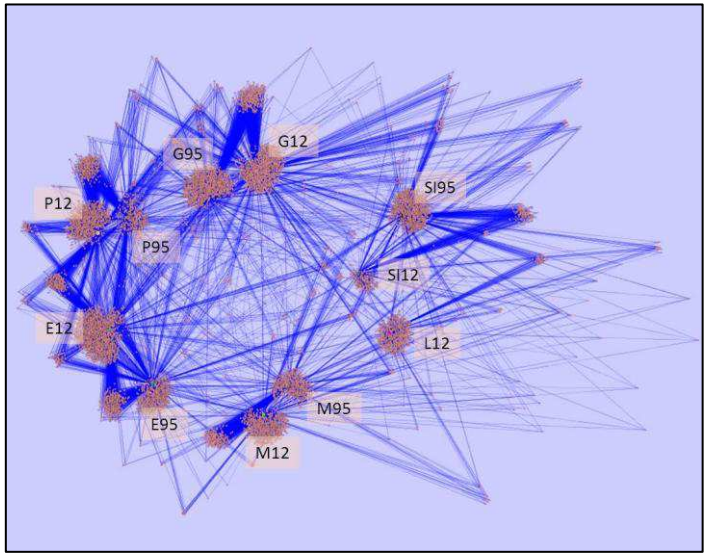

Fig. 2: Graphe mots-classes

(G=Géologie; P=Physique, E=Electronique M=Médecine; S=Science Information, L=linguistique; 12=2012; 95=1995).

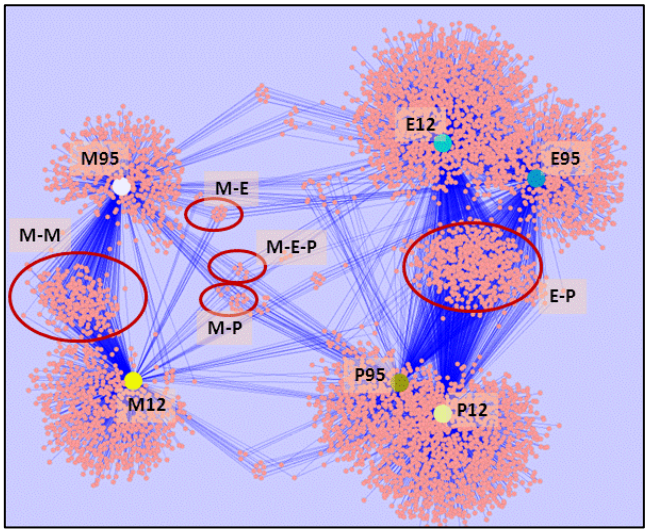

Fig. 3: Graphe mots-classes pour un sous-groupe de 3 domaines scientifiques

(P=Physique ; E=Electronique ; M = Médecine; 12=2012; 95=1995).

Nous devons garder à l'esprit que tous les phénomènes que nous visualisons décrivent le corpus dans son ensemble et ne peuvent donc pas être interprétés d'une manière absolue, en isolant une classe par rapport aux autres, ceci parce que les nœuds (mots-clés) du graphe sont calculés à partir de la classification de l'ensemble des données (il en est de même pour les poids des liens). Se référant à la figure 2, on peut cependant observer plusieurs scénarios intéressants qui se produisent de façon similaire dans nos données expérimentales:

- Evolution de domaines scientifiques: le graphe met en évidence que, comparé à d'autres domaines, le domaine de la médecine (et surtout la discipline de «techniques de diagnostic») s'individualise bien en deux groupes pour chaque période analysée. En effet, d'un côté, nous pouvons observer un nuage dense de mots-clés (M - M) qui sont communs aux deux périodes, mais ont un lien relativement faible avec les classes M95 et M12. De l'autre côté, des mots clés comme «troubles du sommeil», «hypercholestérolémie», «aspect médico-légal», «radiochirurgie» sont des termes spécifiques en 2012 (appartiennent à la classe M12), tandis que des mots clés comme «artériographie», «angine de poitrine», «valve cardiaque», «leucémie lymphoïde» appartiennent à 1995 (classe M15). D'autre part, les nuages de mots-clés liés à la physique (et dans une moindre mesure, ceux liés à l'électronique) restent homogène, quelle que soit la période considérée, indiquant des changements temporels moins importants;

- La transdisciplinarité: il y a des petits groupes de mots-clés avec des relations partagées avec plusieurs nuages principaux. Ce sont des vocabulaires transdisciplinaires reflétant la coopération entre les domaines scientifiques ou des applications pratiques de nouvelles technologies: M-E mots-clés reliant médecine à l'électronique, M-P mots-clés reliant médecine et physique, E-P mots-clés reliant l'électronique et la physique, et enfin, M-E-P mots-clés qui relient les trois domaines.

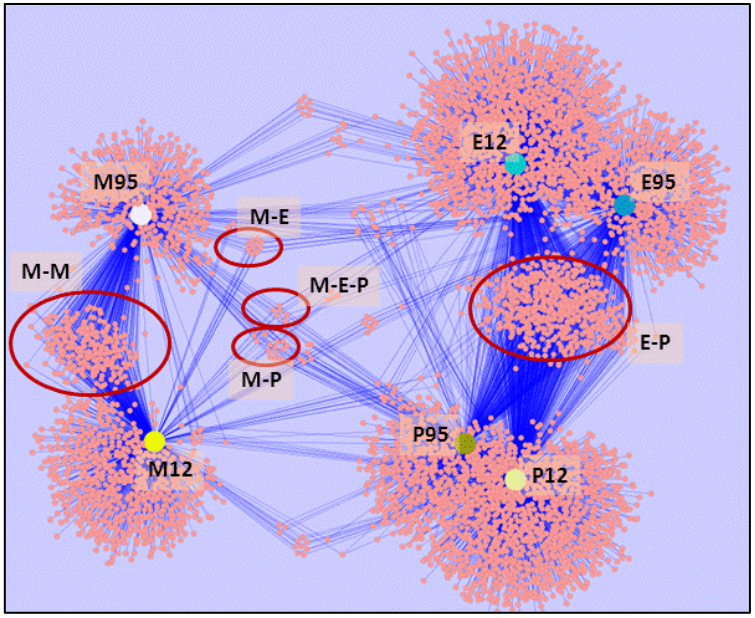

Fig. 4: Graphe auteurs-classes pour un sous-groupe de 3 domaines scientifiques (P=Physique ; E=Electronique ; M=Médecine; 12=2012; 95=1995).

6 Conclusion

Nous avons montré à partir d'un exemple simple que notre méthode totalement non supervisée permet à un utilisateur non-expert de visualiser les termes utilisés dans diverses disciplines et leur évolution temporelle. De façon complémentaire, avec cette méthode, il est également facile pour l'utilisateur de distinguer les termes communs à plusieurs sujets. Selon le cas, le graphe original peut-être trop dense pour une visualisation claire, mais il est alors facile de sélectionner un sous-graphe dont l'analyse peut être effectuée. L'originalité de notre approche GRAFSEL vient du fait que les nœuds du graphe résultent de la combinaison d'une classification et d'un processus de sélection de variables, qui sont appliqués dans un premier temps, et les liens du graphe résultent d'une fonction de contraste, qui est appliquée dans un deuxième temps sur les variables sélectionnées (les nœuds). Ainsi, à la différence des méthodes couramment utilisées, nous ne construisons pas ici un graphe de mots, mais un graphe de relations entre les mots et les classes, chaque classe représentant un domaine principal de l'étude. De cette manière, une autre application intéressante de notre approche peut être la détection des auteurs qui représentent des «passerelles» entre les disciplines scientifiques. Dans la figure 4, l'on voit, qu'en 18 ans, le paysage des auteurs a été considérablement renouvelé et, en plus, qu'il est possible d'identifier les groupes de personnes qui sont des «passeurs de savoirs» entre les domaines scientifiques ou entre les périodes d'un même domaine.

Enfin, des méthodes de classification non supervisées comme supervisées peuvent être utilisés indifféremment dans ce processus.

7 Bibliographie

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, no 3, 1979

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

Hajlaoui K., Cuxac P., Lamirel J.-C, François C., Aide à l'expertise des brevets par alignement avec les publications scientifiques, Document numérique, Vol. 16, 2013/1, Lavoisier Ed.

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