La recherche en Science de la Science en Chine continentale : 40 ans d'évolution.
Une nouvelle méthode d'analyse basée sur le clustering avec maximisation des traits et graphes de contraste.

Research in Science of Sciences in Mainland China: 40 years of evolution.
A new analysis method based on clustering with line maximization and contrast graphs.

Jean-Charles Lamirel[1], Pascal Cuxac[2]

[1] SYNALP Team-LORIA, INRIA Nancy-Grand Est, Vandoeuvre-lès-Nancy, France, jean-charles.lamirel@loria.fr
[2] INIST-CNRS, 2, allée du parc de Brabois, 54519 Vandoeuvre-lès-Nancy, France, pascal.cuxac@inist.fr

RÉSUMÉ. Dans une première partie de cet article, nous mettons en lumière le contexte historique de la Science de la Science en Chine et à l'échelle mondiale. Dans une deuxième partie, en utilisant la combinaison d'un clustering GNG (gaz de neurones), des mesures de maximisation des traits et des graphes de contraste, nous effectuons une analyse du contenu d'articles de revues académiques sélectionnées dans le domaine de la Science de la Science en Chine et construisons une carte globale de la recherche au cours des 40 dernières années. De plus, nous mettons en évidence l'évolution du domaine en exploitant les dates de publication et les informations auteurs afin de clarifier le contenu des sujets. Les résultats obtenus, validés par l'expertise, montrent clairement que la Science de la Science en Chine a progressivement mûri au cours des 40 dernières années, passant de la nature générale de la discipline aux disciplines connexes et à leurs interactions potentielles, de l'analyse qualitative à l'analyse quantitative et visuelle, et de la recherche générale sur la fonction sociale de la science aux études plus spécifiques sur sa fonction économique et stratégique. La méthode originale proposée permet d'obtenir sans supervision, sans paramètres et sans connaissances externes une vision à la fois très claire et très précise du développement d'un domaine scientifique.

ABSTRACT. In a first part of this paper, we highlight the historical context of Science of Science both in China and at a world level. In a second part, based on the unsupervised combination of GNG (neural gas) clustering with feature maximization metrics and associated contrast graphs, 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 and make additional use of the author's information for the sake of clarifying topics content. The obtained results, validated by domain experts, interestingly show that the Chinese Science of Science has gradually become mature in the last 40 years, turning from the general nature of the discipline to the relative disciplines and their potential interactions, from the qualitative analysis to the quantitative and visual analysis, and from the general research on social function of science to more specific economic function and strategic function studies. Consequently, the proposed novel method 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.

MOTS-CLÉS. Science de la science, Chine, Monde, Evolution thématique, Maximisation des traits, Apprentissage non supervisé, Analyse diachronique.

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

1. Introduction

La "Science de la Science" prend pour objet de recherche l'ensemble des connaissances scientifiques et techniques et leurs activités, et explore les lois fondamentales du développement de la science et de la technologie. Dès les années 1910, elle a pris racine en Pologne : l'attitude des chercheurs polonais qui sont passés de la métaphysique à la recherche empirique et de l'analyse d'une seule discipline scientifique à l'étude globale de la science a jeté les bases théoriques de la Science de la Science comme domaine d'étude spécifique en Pologne (Yue et al., 2017). Cependant, le livre du chercheur anglais d'orientation communiste Bernal (1939) "*The Social Function of Science*" est généralement reconnu comme le symbole de la naissance réelle de la Science de la Science, l'achèvement de ce livre étant directement guidé par l'"épisode de Hessen" (Hongzhou et Guohua, 1988) qui a pris son origine profonde dans l'idéologie de Marx, Marx arguant que "l'essence des sciences est juste leur fonction sociale". Des réflexions similaires semblent être récurrentes, et plus récemment, (Hongzhou et Guohua, 1983) ont également fait valoir que science et société sont étroitement liées et ne peuvent être dissociées l'une de l'autre.

En tant que matière globale et interdisciplinaire, la Science de la Science a pour objectif principal de prendre l'ensemble des connaissances scientifiques et technologiques et leurs activités comme objet de recherche afin d'explorer les lois fondamentales du développement de la science et de la technologie. Son champ de recherche devrait donc englober la recherche historique, philosophique, sociologique et économique sur la science. Cependant, le développement de la Science de la Science sur la scène internationale ne s'est pas opéré aussi simplement. La figure 1 illustre le parcours des lauréats du prix Bernal. Il est divisé en trois axes de recherche différents, "Scientométrie (Scientometrics)", "Science, technologie et société (STS)" et "Sociologie du savoir scientifique (SSS)". Derek J. de Solla Price a hérité et développé les idées et paradigmes scientifiques de Bernal, approfondissant et élargissant la théorie et les méthodes de la Science de la Science en mettant l'accent sur les données et sur l'analyse quantitative de la science (Zeyuan et al., 2013). De son côté, le sociologue scientifique américain R.K. Merton a examiné la relation entre science, technologie et société (STS) comme objet de recherche indépendant mais en excluant la possibilité de recherche sociologique sur le contenu du savoir scientifique (Genxiang et Renkun, 1998). En conséquence, la recherche sur la sociologie de la science explorant les "perspectives sociales" et les "perspectives cognitives" de la Science de la Science a été continuellement différenciée au cours du développement du domaine (comme l'illustre par exemple la création de la SSK[1], qui s'intéresse aux domaines de l'"anthropologie" et de l'"éthique"). De fait, la Science de la Science s'est graduellement écartée du paradigme original de la théorie scientifique de Bernal[2].

Dans le contexte spécifique de la Chine, le livre de Bernal (Bernal, 1939) a suscité un grand intérêt dès sa publication. En particulier, la partie mentionnant la Chine et soulignant les limites du développement de la science moderne dans ce pays a rapidement attiré l'attention de grands scientifiques chinois tels que Kezhen Zhu (1890-1974, président de l'Université du Zhejiang), Xuezhou Wu (1902-1983, directeur de la Chinese Chemistry Institution) et Hongjun Ren (1886-1961, l'un des fondateurs de la China Science Society) (Wei et Xinxin, 2012) et les remarques qui y sont faites se sont donc rapidement propagés en Chine.

---

[1] Cf. figure 1.

[2] Il convient de noter que des contributions scientifiques très récentes sur la Science de la Science, comme celle de (Zeng et al. 2017) publiée sur Physics Reports and Science par le System Science Research Team de la Beijing Normal University en Chine, celle de (Fortunato et al. 2018), rejoignant des auteurs de l'Université de l'Indiana (États-Unis) et de l'Université de Leiden (Pays-Bas), ou même des articles récents de grande qualité publiés par la "Complex Networks Research Team" de la Northeastern University (États-Unis) (Huang et al. 2012) (Wang et Barabási 2013) (Shen et Barabási 2014) (Sinatra et al. 2017) semblent indiquer que le champ de recherche scientifique reprend son ampleur et revient au modèle original de Bernal.

La naissance formelle de la Science de la Science en Chine est cependant venue de l'initiative de Tsien Hsueshen dans le document intitulé "Science et Technologie" en 1977 qui a encouragé la création d'un nouvel espace de recherche en Chine appelé "Science de la Science" (Hsueshen, 1979). En suivant la pensée de Bernal, Tsien Hsueshen a souligné que le domaine de la Science de la Science devrait appartenir aux sciences sociales (1979, 1980). Jusqu'à présent, trois instituts de recherche spécifiques ont focalisé leurs recherches sur la Science de la Science à Pékin, Tianjin et Shanghai. En outre, les instituts non spécialisés établis à Pékin, comme la CAST (China Association for Science and Technology), la CASTED (The Chinese Academy of Science and Technology for Development du Ministère des sciences et de la technologie de la République populaire de Chine), l'Institute of Science and Development de la Chinese Academy of Sciences, le Chinese Institute of Engineering Development Strategies et de nombreux collèges et universités dans toute la Chine ont également investis des forces importantes dans la recherche fondamentale et appliquée en Science de la Science.

En 2010, à l'occasion du 30e anniversaire de la revue chinoise Science of Science and S&T Management, Liu Zeyuan a mis en évidence les frontières et les principaux domaines de la Science de la Science en Chine en cartographiant la littérature dans ce domaine avec l'outil CiteSpace[3]. Deux grands blocs de connaissances, correspondant à 2 voies de développement complémentaires de la Science de la Science en Chine ont été ainsi isolés : scientométrie, axée sur l'analyse quantitative, et, études scientifiques, axées sur l'analyse philosophique.

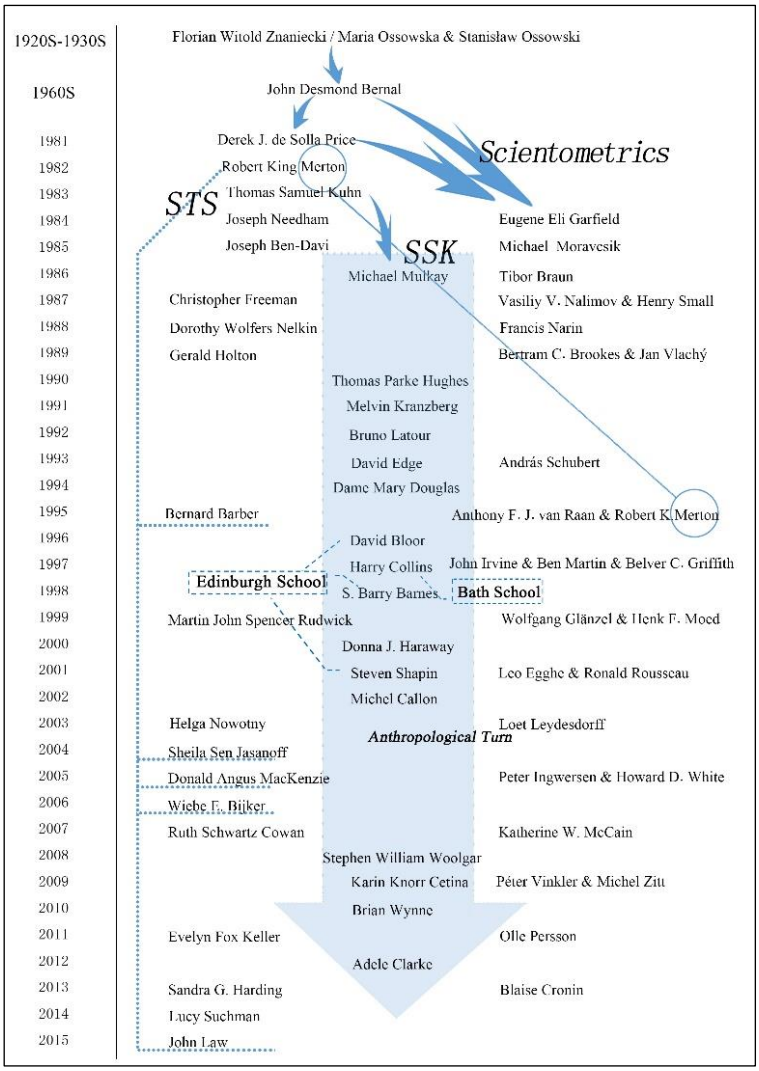

Figure 1. *La voie du développement de la Science de la Science.*

[3] http://cluster.ischool.drexel.edu/~cchen/citespace/download/

Dans cet article, nous exploitons le matériel de recherche des 40 dernières années en Science de la Science en Chine et nous mettons en place une nouvelle méthode pour comprendre et suivre à la fois plus clairement et plus précisément le développement dans ce domaine. Notre objectif est donc de donner des indications plus claires sur l'origine de la Science de la Science chinoise, sa structure et ses orientations futures par l'intermédiaire d'une méthode d'analyse données originale, fonctionnant de manière entièrement non supervisée, sans paramètres et sans source de connaissances externe.

2. Collecte et prétraitement des données

Etant donné les contours flous et la vaste étendue du domaine de la Science de la Science (cf. section 1), il n'est pas facile de faire une extraction complète et précise de la littérature relative à ce domaine. C'est pourquoi nous avons choisi, dans le présent document, de nous concentrer sur l'évolution du contenu-cœur du domaine plutôt que d'essayer d'être exhaustifs.

Nous avons interrogé la base de données de la China National Knowledge Infrastructure (CNKI) en utilisant "Science of Science" comme terme thématique[4]. Nous avons ainsi extrait 2401 articles publiés dans les revues principales de l'Université de Pékin et les revues du CSSCI (couvrant une période de recherche allant du démarrage de la discipline en Chine jusqu'au 22-10-2017). Un nettoyage des données a été effectué dans une seconde phase afin de supprimer les éléments qui ne correspondaient pas à des documents de recherche (par exemple les avis de réunion, les présentations ou les éditoriaux de journaux). Cette phase nous a permis de conserver 1334 articles de revues. En repartant de ces derniers, nous avons ensuite récupéré 2677 articles cités (après en avoir supprimé les doublons), dont 1539 ont été publiés dans des revues spécialisées. Nous avons ajouté ces 1539 documents à nos 1334 documents de base pour former notre ensemble de données expérimentales : 2873 articles de revues au total.

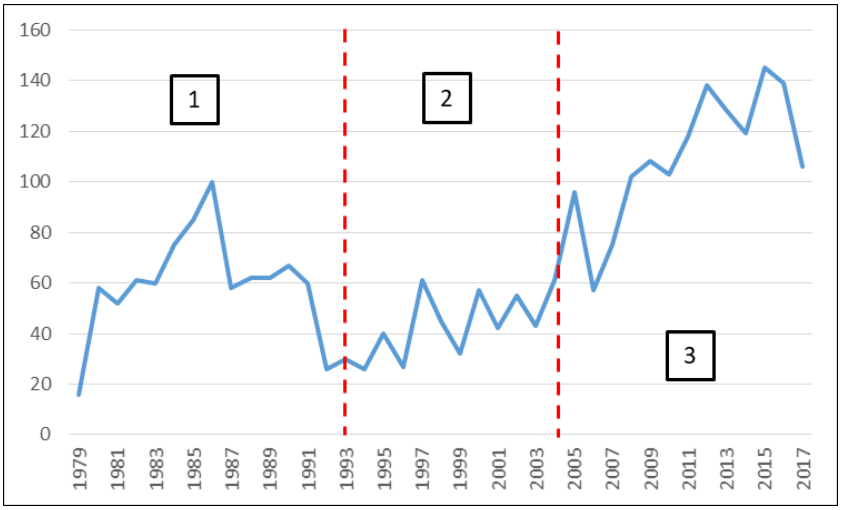

Figure 2. *Distribution du nombre d'articles par ans et correspondance avec les périodes historiques pour « Science de la Science en Chine ».*

Nous avons ensuite d'abord cherché un moyen indirect de valider notre processus de collecte de données. Pour cela, nous avons tracé la distribution des articles en fonction de leurs dates de publication. Il s'est avéré que la tendance que nous avons pu observer (Fig. 2) s'accorde parfaitement avec les observations récentes de Liu Zeyuan basées sur l'expérience du domaine (Zeyuan, 2017)

[4] Par le fait que l'interrogation de la base de données CNKI se fait en chinois, nous utilisons deux termes différents car "Science de la Science" est décrit par 2 termes différents dans cette langue : " 科学学 " et " 科学的科学 ".

montrant que la recherche en Science de la Science en Chine, a connu trois étapes, à savoir, une période de croissance rapide (1977-1990) (1), une période de développement tortueux (1992-2003) (2) et une période de régénération (2003-) (3).

Dans une deuxième phase, les titres, résumés et mots-clés des 2790 articles ont été extraits[5].

Le processus d'indexation s'est avéré assez complexe. Il a commencé avec un dictionnaire initial de 9679 mots-clés rassemblés à partir du champ des mots-clés des 2790 articles. Pour la segmentation des mots et le balisage des titres et des résumés en plein texte des articles, nous avons utilisé NLPIR-ICTCLAS [6], une boîte à outils spécifique pour le traitement de la langue chinoise. En raison des particularités de la Science of Science, les logiciels ne peuvent pas segmenter avec précision certains termes de domaine exprimés par des mots complexes à plusieurs caractères, tels que " **科学学** " (Science de la Science), " **科学学研究** " (Recherche sur la Science), " **科学**逻辑学 " (Logistique sur la Science), " **科学的社会功能** " (Fonction sociale des Sciences). Nous avons donc réalisé une reconstruction postérieure de ces mots.

Parmi les mots extraits, nous avons ensuite, à l'aide d'un programme Python ad-hoc, filtré les éléments étiquetés comme noms avec et supprimé les quantités (nombres, dates,...). Puis, nous avons opéré une deuxième phase de nettoyage des noms pour supprimer les mots vides et ceux couvrant l'ensemble contexte du corpus (par ex. "recherche", "analyse", "année") et aussi pour fusionner des noms ayant une signification similaire (par ex. " **著者分布** " et " **作者分布** " : distribution des auteurs, " **作者合作网** " et "**作者合作网**络 " : réseau de coauteurs, " **知**识图谱 " et " **知**识图谱分析 " : cartographie des connaissances). Une fois fusionné avec le dictionnaire initial de mots-clés, il en est résulté un dictionnaire de 13442 noms chinois.

La traduction anglaise a ensuite été appliquée au dictionnaire des noms. En raison d'un vocabulaire plus pauvre en anglais qu'en chinois, la traduction était susceptible de générer de nouveaux mots équivalents (tels que " **知**识地图 " : "knowledge geography" et " **可**视化底图 " : "basic visualization map" , " **科技**评估 " : "S&T evaluation" et "-**科研**评价 " : "research evaluation") qui devaient être refusionnés. Après ce processus, nous avons obtenu un dictionnaire de 11931 noms anglais. Des étiquettes de catégorie (resp. "nom", "ville", "pays") ont finalement été attachées aux noms représentant les entités correspondantes (resp. personnes, lieux, pays).

Pour éliminer le bruit restant nous avons appliqué une passe supplémentaire de nettoyage détaillée dans le tableau 1. Premièrement, nous avons fusionné les mots équivalents restants en une seule entrée (par exemple, un auteur peut apparaître avec ou sans son prénom comme "Merton" et "R.K. Merton" - une institution peut apparaître avec son acronyme ou sous une forme développée comme "NSF" et "National Science Foundation"), deuxièmement, nous avons supprimé les mots ou expressions dont le sens est peu clair en anglais, et, nous avons corrigé quelques erreurs de traduction. Ce dernier processus a permis de supprimer 360 entrées dans le lexique (235 entrées fusionnées et 125 entrées supprimées). Un seuil de fréquence de 6 a finalement été appliqué pour supprimer les mots de basse fréquence[7]. Il en est résulté un dictionnaire final de 1576 termes avec lesquels les articles ont été réindexés.

---

[5] Pour les articles antérieurs à 1997 qui ne contenaient ni résumé ni mots-clés, nous n'avons utilisé que l'information présente dans le titre.

[6] http://ictclas.nlpir.org/

[7] Le seuil de fréquence de 6 est trouvé empiriquement : il permet à la fois de réduire l'espace de description de façon significative (qui sinon serait beaucoup trop important) tout en permettant une classification de qualité (estimée à la fois par les experts et par nos paramètres de qualité présentés à la section 4.1). Aucun document n'est supprimé par ce processus.

	1- Fusion de termes	2- Suppression des termes vides	3- Seuil de fréquence >5
Taille initiale du vocabulaire	11931	11696	11571
Mots supprimés	--	125	9995
Mots fusionnés	235	--	--
Taille du vocabulaire obtenu	11696	11571	1576

Tableau 1. *Résumé des étapes de traitement du lexique.*

3. La maximisation des traits en tant qu'approche globale pour l'analyse des données

Notre méthode d'analyse des données de Science de la Science repose principalement sur un processus de sélection de variables (traits) qui s'appuie elle-même sur la mesure de maximisation des traits (Lamirel et al., 2011). Nous présentons d'abord cette mesure avant de présenter l'ensemble du processus d'analyse des données. La maximisation des traits est une mesure non biaisée qui peut être utilisée pour estimer la qualité d'une classification, qu'elle soit supervisée ou non supervisée. Dans le cas de la classification non supervisée (i.e. clustering), cette mesure exploite les propriétés (i.e. les caractéristiques) des données associées aux clusters à des fins multiples (étiquetage de clusters, visualisation globale des résultats de clustering telle que la représentation par graphe de contraste présentée dans la suite de ce travail, détection du modèle de clustering optimal). Ses principaux avantages sont d'être sans paramètres, d'être totalement indépendante de la méthode de clustering et de son mode de fonctionnement, de supporter des espaces de grandes dimensions et de présenter un meilleur compromis entre discrimination et généralisation que les métriques usuelles (Euclidienne, Cosinus ou Chi carré, etc.).

3.1. *F-mesure de traits*

Considérons un ensemble de données D représenté par un ensemble de variables, ou traits, F, et un ensemble de clusters C résultant d'une méthode de clustering[8]. La métrique de maximisation des traits favorise les traits saillants attachés aux données des clusters. La F-mesure de traits d'un trait f associée à un cluster c est définie comme la moyenne harmonique du rappel de traits et de la prédominance de traits, eux-mêmes définis comme suit :

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

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

avec

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

où W_d^f représente le poids du trait f pour les données et F_c représente l'ensemble des traits présents dans l'ensemble des données associées au cluster. La prédominance de traits mesure la capacité de f à décrire le cluster c. De manière complémentaire, le Rappel de traits met en avant la capacité de f à discriminer c des autres clusters.

[8] Dans cet article, les traits représentent les mots extraits du titre, du résumé et des mots clés de l'article, les poids des traits sont les informations fréquentielles ajustées qui leur sont associées et la classification non supervisée (clustering) est basée sur l'algorithme GNG.

Le rappel de traits est une mesure indépendante de l'échelle, mais la prédominance des traits ne l'est pas. Nous avons cependant montré expérimentalement (Lamirel et al. 2015) que la F-mesure, qui est une combinaison de ces deux mesures, n'est que faiblement influencée par l'échelle des traits. Néanmoins, pour garantir un comportement indépendant de l'échelle, les données doivent être normalisées. Le choix du schéma de pondération des données n'est pas vraiment limité par l'approche, mais il est nécessaire de traiter des valeurs positives. Un tel schéma est censé déterminer la signification (c'est-à-dire la sémantique et l'importance) du trait pour les données[9].

3.2. *Maximisation des traits*

Dans un contexte supervisé, la mesure de maximisation des traits peut être exploitée pour générer un puissant processus de sélection de variables. Dans notre contexte de regroupement non supervisé, le processus de sélection peut être utilisé a posteriori pour décrire ou étiqueter les clusters selon leurs caractéristiques les plus typiques et les plus représentatives. Il s'agit d'un processus non paramétré qui utilise à la fois la capacité de la F-mesure à discriminer entre eux les clusters (mesure $FR_c(f)$) et sa capacité à représenter fidèlement les données des clusters (mesure $FP_c(f)$).

L'ensemble S_c des traits caractéristiques d'un cluster c donné appartenant à une partition C est défini comme :

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

où

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

et où $C_{/f}$ représente le sous-ensemble de C dans lequel le trait apparait.

Enfin, l'ensemble S_C des traits sélectionnées est le sous-ensemble de F défini par :

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

En d'autres termes, les traits jugés pertinents pour un cluster donné sont ceux dont les représentations sont (1) meilleures dans ce cluster que leur représentation moyenne dans l'ensemble des clusters, et (2) meilleures que la représentation moyenne de tous les traits dans la partition, en termes de F-mesure du trait. Les traits qui ne respectent jamais la deuxième condition dans un quelconque cluster sont éliminés, ce qui induit un processus de sélection de variables.

3.3. *Contraste*

Un concept spécifique de contraste $G_c(f)$ peut être défini pour calculer la performance d'un trait conservé *f* pour un cluster *c* donné. Il s'agit d'une valeur d'indicateur proportionnelle au rapport entre la F-mesure $FF_c(f)$ d'un trait pour le cluster et la F-mesure moyenne $\overline{FF}$ de ce trait pour l'ensemble de la partition. Le contraste d'un trait pour un cluster s'exprime par :

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

Les traits actifs d'un cluster sont ceux pour lesquels le contraste est supérieur à *1*. De plus, plus le contraste d'un trait est élevé pour un cluster, meilleures sont ses performances dans la description du contenu du cluster.

[9] Un trait ayant des valeurs négatives peut être séparé sans perte d'information en 2 sous-traits positifs différents, le premier représentant la partie positive du trait original, et le second, sa partie négative.

Un exemple-jouet du fonctionnement des indices présentés, y compris de celui du contraste est donné dans la référence (Lamirel et al. 2016).

Comme nous l'avons déjà mentionné précédemment, dans le clustering, les traits actifs d'un cluster sont des traits sélectionnés pour lesquels le contraste est supérieur à *1* dans ce cluster. Inversement, les traits passifs d'un cluster sont des traits sélectionnés présents dans les données du cluster pour lesquels le contraste est inférieur à l'unité. Relativement au principe de la méthode, chaque trait sélectionné présente inévitablement un contraste supérieur à 1 dans un ou plusieurs clusters (voir l'équation 7 pour plus de détails). Une façon simple d'exploiter les traits obtenus est donc d'utiliser des traits actifs sélectionnés et leur contraste associé pour l'étiquetage des clusters comme nous l'avons proposé dans (Lamirel et al. 2015).

4. Processus d'analyse des données expérimentales

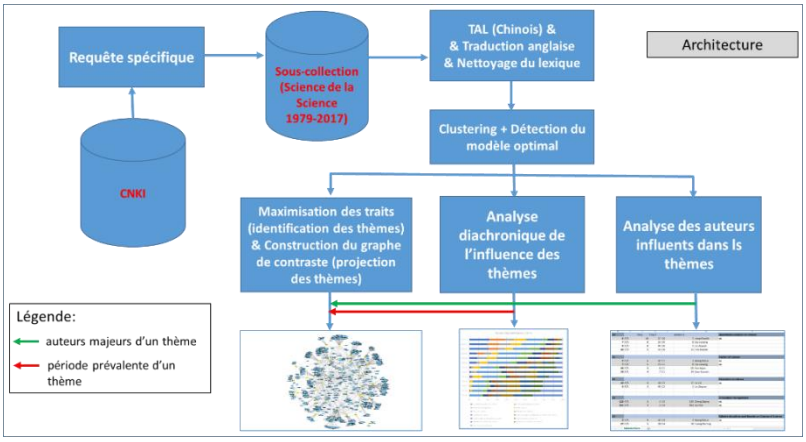

Figure 3. *Processus global d'analyse des données.*

Nous présentons l'architecture globale de notre processus expérimental à la figure 3. Après des étapes de prétraitement, le processus exploite le clustering en combinaison avec la maximisation des traits pour extraire les principaux sujets de recherche de l'ensemble de données du corpus. Nous avons récemment montré (Lamirel et al., 2015) que la combinaison d'une approche de clustering appropriée, comme le clustering neuronal à base de gaz croissants GNG (Fritzke, 1995), avec la maximisation des traits offre des performances supérieures par rapport aux approches alternatives pour l'extraction de sujets, comme LDA (Blei et al., 2003), à la condition de pouvoir identifier correctement un modèle optimal de clustering (c'est-à-dire un nombre approprié de clusters) à partir des données analysées. Nous proposons donc d'exploiter une de nos approches récentes, également basée sur la maximisation des traits, pour la tâche de détection du modèle de clustering optimal (Lamirel et al., 2016). La gestion des résultats de clustering avec une approche graphique basée sur le contraste est une méthode originale présentée dans cet article. Elle permet à la fois de réduire la surcharge cognitive qui devrait résulter de la représentation des interactions dans de grands ensembles de données et de déterminer avec précision les dépendances entre les sujets extraits grâce à des traits partagés à contraste élevé. La dernière partie de notre approche exploite des étiquettes externes des données associées aux clusters. Premièrement, la date de publication est utilisée pour effectuer une analyse diachronique de l'activité des clusters (c.-à-d. de celle des sujets) et deuxièmement, l'information sur les auteurs est utilisée pour mettre en évidence les auteurs les plus influents dans les différents sujets. Les informations relatives à la date et aux auteurs sont également indiquées sur le graphe de contraste. Les détails sur les différentes étapes de l'approche sont donnés dans les sections suivantes.

4.1. *Clustering et détection du modèle optimal*

Nous exploitons 2 méthodes de clustering différentes, à savoir les K-means (MacQueen, 1967), une méthode compétitive sans coopération entre les prototypes, et GNG (Fritzke, 1995), une méthode neuronale compétitive coopérative incluant l'apprentissage Hebbien. Dans tous les cas, la méthode GNG s'est avérée supérieure à la méthode K-means en raison du processus de coopération entre prototypes lors de l'apprentissage et du processus Hebbien complémentaire. Elle fournit notamment une meilleure indépendance par rapport aux conditions initiales et évite le plus souvent de produire des résultats de regroupement dégénérés. Ce type de résultats a également été observé dans plusieurs de nos expériences antérieures (Lamirel et al., 2011).

Le choix du modèle optimal repose également sur les mesures de maximisation des traits présentées à la section précédente. Nos expériences antérieures sur des ensembles de données de référence montrent en effet que la plupart des estimateurs de qualité habituels[10] ne produisent pas de résultats satisfaisants dans un contexte de données réalistes, ils sont très sensibles au bruit et fonctionnent mal avec des données de grandes dimensions (Kassab et Lamirel, 2008). Une méthode plus judicieuse consiste donc à s'affranchir des problèmes des distances usuelles, comme la distance euclidienne, utilisées par ces indices en exploitant la maximisation des traits et des informations plus spécifiques liées à l'activité et à la passivité de certains traits dans les clusters pour identifier une partition optimale. On s'attend à ce que ce type de partition maximise le contraste décrit à l'équation 7. En effet, plus les traits sont contrastés, plus les clusters sont compacts et séparés. Cette approche nous a amenés à définir trois nouveaux indices : PC, EC et CB.

Nous donnons ci-après à titre d'exemple l'expression des indices des indices PC et EC. L'indice CB représentant une combinaison pondérée des deux autres. Une description plus précise de cette approche ainsi que les expérimentations de comparaison avec le autres indices sur des cas de données réelles, de simples à complexes, peut être trouvée dans la référence (Lamirel et al., 2016).

L'indice PC, dont le principe correspond par analogie à celui de l'inertie intra-cluster dans les modèles habituels, est une macro-mesure basée sur la maximisation du contraste moyen pondéré des caractéristiques actives pour une partition optimale.

Pour une partition comprenant plusieurs clusters il peut s'exprimer sous la forme :

$$PC = arg\max_{k}\left(\frac{1}{k}\sum_{i=1}^{k}\frac{1}{|s_i|}\sum_{f\in S_i} G_i(f)\right) \qquad (8)$$

L'indice EC, dont le principe correspond par analogie à celui de la combinaison entre l'inertie intra-cluster et l'inertie inter-cluster dans les modèles habituels, est basé sur la maximisation du compromis moyen pondéré entre le contraste des éléments actifs et le contraste inversé des éléments passifs pour une partition optimale :

$$EC = arg\max_{k}\left[\frac{1}{k}\sum_{i=1}^{k}\left(\frac{|s_i|\sum_{f\in S_i} G_i(f)+|\overline{s_i}|\sum_{h\in\overline{s_i}}\frac{1}{G_i(h)}}{|s_i|+|\overline{s_i}|}\right)\right] \qquad (9)$$

où n_i représente le nombre de données associées au cluster i, $|s_i|$ représente le nombre de traits actifs dans i, and $|\overline{s_i}|$, le nombre de traits passifs dans i.

[10] Tels que l'indice de Dunn (Dunn, 1974), l'indice de Davies-Bouldin (Davies et Bouldin, 1979), l'indice Silhouette (Rousseeuw, 1987), l'indice Calinski-Harabasz (Caliński et Harabasz, 1974) ou l'indice Xie-Beni (Xie et Beni, 1991).

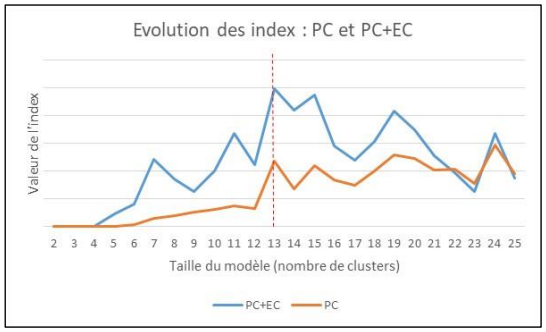

Figure 4. *Evaluation de la qualité du clustering (indices PC et EC) ; modèle optimal (13 clusters). Les valeurs de l'index ont été rééchelonnées pour une meilleure visualisation.*

Dans le cadre de notre expérience, nous faisons varier le nombre de clusters dans une fourchette allant jusqu'à 1/50 du nombre de données. Nous rejetons les modèles de taille 1 pour deux raisons principales : les indices ne sont pas prévus pour produire des résultats dans ce cas, et ces modèles correspondent naturellement à une opération de clustering triviale. Nous opérons un clustering strict. Dans ce cas chaque donnée est ré-associée à un seul cluster et la forme habituelle de la fonction *af* d'affectation d'une donnée à un cluster suit :

$$af(d) = \underset{k}{argmin}\left(Dist(\vec{k}, \vec{d})\right) \qquad (10)$$

où *Dist* représente un fonction distance (générallement la distance euclidienne), $\vec{k}$ représente le vecteur de profil du cluster *k* et $\vec{d}$ représente le vecteur descripteur de la donnée *d*.

En exploitant les données associées aux clusters, nous sélectionnons le modèle qui optimise la combinaison PC+EC des indices mentionnés précédemment. Cette technique permet d'obtenir le nombre pertinent de clusters mettant en évidence les principaux sujets, ou thèmes, de recherche en Science de la Science au cours de la période considérée. Une analyse experte des résultats obtenus confirme que le modèle optimal obtenu est conforme pour représenter avec précision tous les principaux thèmes de recherche du domaine analysé. La figure 4 présente les tendances de l'évolution des indices PC et EC ainsi que le point optimal (c'est-à-dire le nombre optimal de clusters ou le modèle optimal), la figure 5 présente la description d'un cluster basée sur ses caractéristiques les plus contrastées et le tableau 2 présente la liste des titres des clusters que l'expert a caractérisé en exploitant les éléments les plus contrastés.

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 5. *Exemple de description d'un cluster par la liste de ses mots les plus contrastés. Le thème concerné par le cluster est la cartographie des connaissances.*

	Label (expert)	Résumé du contenu (traits principaux)
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

Tableau 2. *Liste et description synthétiques des clusters obtenus. Pour une meilleure clarté, les labels, ou titres, des clusters sont rajoutés par l'expert du domaine.*

4.2. *Construction des graphes de contraste*

Dans le domaine mathématique de la théorie des graphes, un graphe bipartite (ou bigraphe) est un graphe dont les sommets peuvent être divisés en deux ensembles disjoints et indépendants U et V *de* sorte que chaque arête relie un sommet de U à un sommet de V. Les graphes de contraste sont des graphes bipartites basés sur les relations entre un ensemble de traits S et un ensemble d'étiquettes L. Théoriquement, l'ensemble d'étiquettes L pourrait représenter n'importe quel type d'information à

laquelle les caractéristiques peuvent être reliées et l'ensemble de caractéristiques S est un sous-ensemble d'un ensemble global F (c.-à-d. l'espace original des caractéristiques sur lequel reposent les données) qui a été obtenu par un processus de sélection, comme la maximisation des traits présentée ci-dessus. Dans le cas de l'utilisation de la maximisation des traits, le poids $c_{(u,v)}$ d'une arête $(u, v), u \in S, v \in L$ représente alors le contraste du trait u pour une étiquette v telle qu'il est défini par l'équation 7 [11].

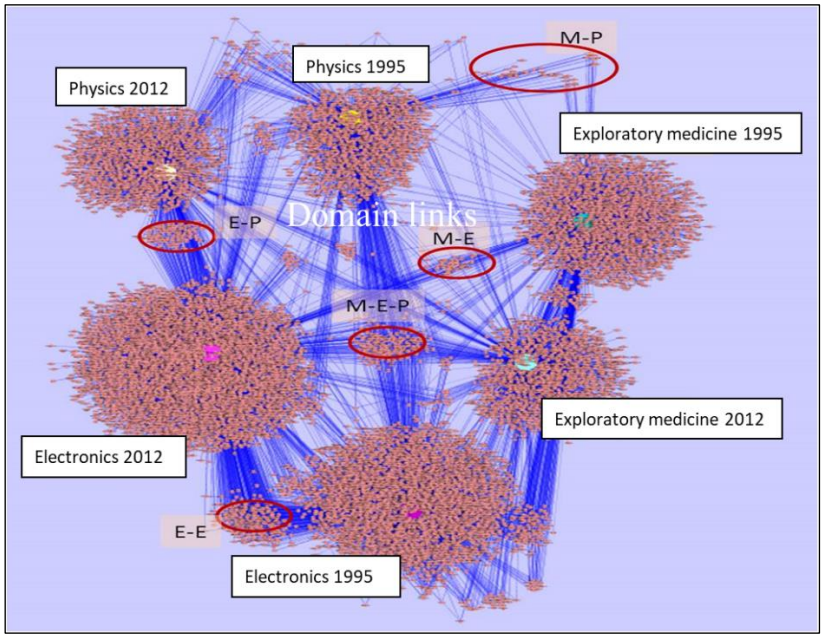

Figure 6. *Un exemple de graphe de contraste matérialisant la relation entre les auteurs et les domaines scientifiques associés aux périodes de temps.*
Les cercles rouges mettent en évidence les auteurs qui représentent les transmetteurs de connaissances entre domaines et périodes.

De tels graphes ont de nombreuses propriétés intéressantes. Tout d'abord, ils réduisent la surcharge cognitive produite par la représentation graphique classique en raison du processus de sélection de traits qui réduit le nombre de connexions potentielles. Deuxièmement, ils peuvent être utilisés pour mettre en évidence indirectement les relations entre les étiquettes, lorsque les caractéristiques ont une interaction contrastée avec plusieurs étiquettes. Troisièmement, la combinaison de cette approche avec le modèle pondéré dirigé par la force (Kobourov, 2012) permet de mettre en évidence les étiquettes centrales ou les plus influentes de la série L et d'identifier facilement les étiquettes qui sont les plus étroitement liées par des caractéristiques associées, ces dernières apparaissant en position proche dans le graphe.

Nous avons proposé une première utilisation originale des graphes de contraste dans le cas de l'analyse de la transdisciplinarité entre différents domaines de recherche et périodes de temps (Cuxac et Lamirel, 2013). La figure 6 montre un graphe de contraste résultant de cette approche où les caractéristiques représentent les auteurs des articles de recherche et les étiquettes représentent une combinaison de périodes et de domaines de recherche. Des auteurs faisant la connexion entre les domaines et les périodes de temps apparaissent clairement sur le graphe ; ils peuvent être considérés comme les transmetteurs de connaissances et, par conséquent, mettent en évidence leur rôle social et scientifique majeur.

[11] Dans l'équation 7, les étiquettes matérialisent les catégories ou les clusters auxquels les données sont associées.

4.3. *Informations complémentaires par l'exploitation d'étiquettes externes*

Comme elles sont définies dans (Attik et al., 2006), les étiquettes externes sont des informations qui sont associées aux données mais qui ne jouent aucun rôle dans le processus initial d'analyse de celles-ci. Toutefois, ces informations peuvent a posteriori fournir des indices importants pour améliorer la précision de l'analyse. Dans le cas du processus de clustering présenté précédemment, les étiquettes externes peuvent ainsi être exploitées dans second temps (i.e. après formation des clusters) en évaluant leur distribution dans les clusters obtenue par l'intermédiaire des données associées à ces derniers. Elles peuvent ainsi fournir des informations complémentaires sur les thèmes représentés par les clusters.

Dans le cas de notre ensemble de données sur la Science de la Science, nous nous concentrons sur deux types d'étiquettes externes, les dates de publication des articles et les auteurs des articles. Les dates de publication des articles sont exploitées pour effectuer une analyse diachronique de l'activité des thèmes, en soulignant l'importance de chaque thème dans chaque période de temps, soit cette activité est considérée individuellement, soit par rapport aux autres thèmes. Comme le montre la section suivante relative à l'analyse des résultats, cette approche permet de comprendre précisément la chronologie globale de l'activité de recherche d'un domaine de recherche, dans note cas celui de la Science de la Science. Les auteurs des articles peuvent être exploités pour mettre en évidence les contributeurs les plus importants qui dirigent ou influencent un thème de recherche et qui peuvent même être considérés comme des contributeurs centraux s'ils dirigent/influencent/coordonnent plusieurs domaines de recherche en même temps[12].

Dans le contexte de notre expérience, notre analyse des étiquettes externes est basée sur deux mesures différentes qui sont la fréquence et la prévalence des étiquettes. La fréquence d'étiquetage F_c^l d'une étiquette *l* de type *t* dans un cluster *c* peut être définie comme :

$$F_c^l = Card\{d \in D \mid af(d) = c \wedge l \in Extlab_t(d)\} \qquad (11)$$

où *Card* est la fonction cardinal, *D* est l'ensemble des données exploitées, *af* la fonction définie à l'équation 10 qui fournit le cluster associé à chaque donnée et $Extlab_t(d)$ une fonction qui fournit la liste des étiquettes externes de type *t* associées à la donnée d.

La prévalence d'une l'étiquette est une mesure fondée sur les clusters. Une étiquette *l* est prévalente dans un cluster *c* si :

$$\nexists\, c' \in C, c' \neq c, F_{c'}^l > F_c^l \wedge \nexists\, l' \in L_c\,,\ l' \neq l\,,\ F_c^{l'} > F_c^l \qquad (12)$$

où L_c est l'ensemble des étiquettes présentes dans le cluster *c à* travers ses données associées.

La prévalence est utilisée pour mettre en évidence l'influence préalable d'une étiquette. Par conséquent, selon cette définition, une étiquette peut être prévalente uniquement dans un seul cluster et certain clusters peuvent ne pas avoir d'étiquettes prévalentes.

5. Résultat de l'analyse et de la visualisation des données

5.1. *Structure générale du domaine de la Science de la Science*

Dans le cas spécifique de notre expérience sur les données de la Science de la Science, nous proposons de construire un graphe de contraste entre un ensemble de clusters (ensemble *L*)

[12] Dans de nombreux cas, chaque donnée peut avoir plusieurs étiquettes externes du même type. Par exemple, un document de recherche peut avoir plusieurs auteurs différents.

représentant les principaux thèmes de recherche du domaine qui ont été extraits par le processus de clustering et les éléments les plus contrastés (ensemble *S*) issus des descriptions des clusters (voir figure 5). Cette approche qui combine clustering et graphe de contraste d'une manière originale s'avèrera particulièrement utile aux experts en Science de la Science [13] pour comprendre la construction de leur domaine, en mettant en évidence les thèmes les plus centraux du domaine (générateurs de domaine) ainsi que les thèmes périphériques qui sont les plus connectés. Dans le graphe résultant, seuls les traits dont le contraste est supérieur à 1,4 sont conservés pour participer à la représentation (1074 traits parmi les 1576 utilisées pour le clustering sont ainsi conservés). La figure 7 représente l'ensemble du graphe et la figure 8 représente une sous-partie mettant en évidence des domaines étroitement liés qui partagent beaucoup de caractéristiques.

La distribution spatiale des 13 thèmes est présentée sur la figure 7 (dont un extrait est également présenté en 7-b). Selon les experts, ce graphe met en évidence une structure très clairement interprétable du domaine de la Science de la Science en Chine. Dans un tel modèle, les thèmes fortement interconnectés auront tendance à apparaître au centre de la représentation (voir section 4.2). Dans notre cas, cette information sur les domaines de bases est représentée par deux thèmes complémentaires : "8# Attributs du domaine de la Science de la Science" et "7# Politique de recherche et impacts sur la société".

L'explication donnée par les experts est la suivante :

1) La Science de la Science est un domaine réflexif, elle doit donc faire face à son propre développement, étudier les modèles de la science eux-mêmes (thème 8#).

2) La Science de la Science est une science appliquée qui guide la pratique. Par conséquent, elle doit être guidée par la pratique, étudier la fonction sociale de la science et servir la politique de recherche (thème 7#).

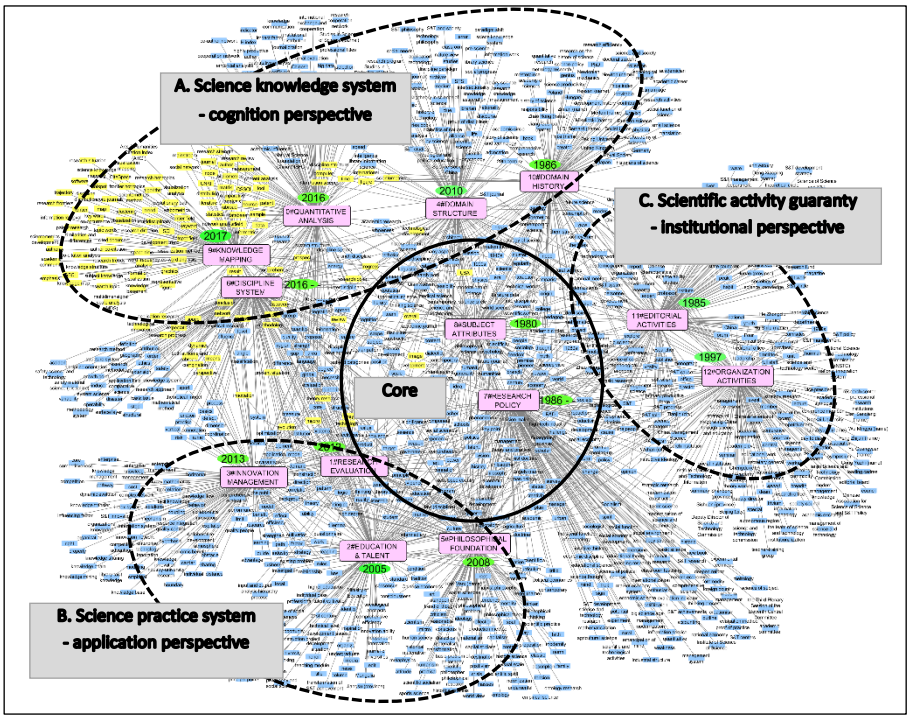

Figure 7. *Graphe de contraste global représentant les principaux thèmes en Science de la Science en Chine.*

(Le cluster 9# est mis en évidence et les informations détaillées sont présentées dans le tableau 2)

[13] Selon les dires des experts eux-mêmes.

Autour de l'information de base, apparaissent trois grands champs de recherche qui correspondent aux activités scientifiques habituelles, à savoir "A. Système de connaissances scientifiques", "B. Système de pratique de la Science" et "C. Garantie de l'activité scientifique", qui forment la structure logique complète de la Science de la Science du point de vue respectivement de la cognition, de l'application et de l'institutionnel.

Le champ "A. Système de connaissances scientifiques" est associé aux cinq thèmes connexes "10# Histoire de la Science de la Science", "4# Structure du domaine et disciplines périphériques sur la Science de la Science", "6# Système disciplinaire", "0# Analyse quantitative de la Science" et "9# Cartographie des connaissances de la Science". Les preuves historiques (10#) ont servi de point de départ aux premières études de la structure de la Science de Science et des disciplines périphériques et à la poursuite de la construction du domaine (4#) et finalement, au développement des approches quantitatives de la recherche scientifique (0#). L'apparition récente de nouvelles techniques de gestion des données et de technologies logicielles a conduit au développement d'approches de visualisation élaborées (9#). Bien qu'il soit évidemment lié à la structure du domaine (4#), le système disciplinaire (6#) est un domaine de recherche plus indépendant lié à l'étude sur la méthodologie de recherche, les approches comparatives et les sciences de la sécurité.

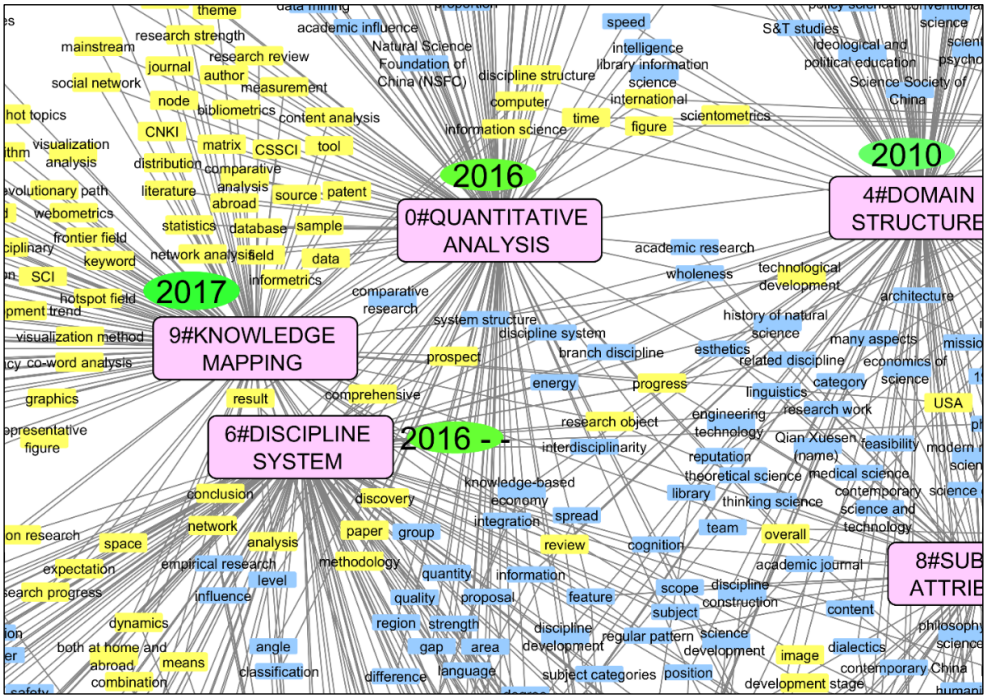

Figure 7-b. *Zoom présentant un extrait ciblé du graphe de contraste global représentant les principaux thèmes en Science de la Science en Chine.*

Le champ "B. Système de pratique scientifique" est composé de quatre grands thèmes connexes : "1# Evaluation de la recherche", "3# Gestion de l'innovation", "2# Sciences de l'éducation et culture des talents" et "5# Fondements philosophiques de la Science de la Science". Aujourd'hui, la stratégie d'innovation représente l'une des préoccupations majeures du gouvernement chinois car elle vise à lier la science et la technologie à l'économie et aux applications à valeur ajoutée pour le développement de la société chinoise. Le rôle de la science et de la technologie dans la pratique sociale est donc logiquement matérialisé par la gestion de l'innovation (3#) et ce domaine s'est considérablement développé ces dernières années (voir aussi figure 11). L'évaluation de la recherche (1#) (analyse des intrants et extrants scientifiques, prise de décision scientifique), ainsi que l'éducation et la fertilisation des talents (2#), ont des rôles complémentaires dans la stratégie d'innovation. Par conséquent, ces domaines sont également devenus des questions indispensables dans le système de pratique scientifique en Chine. Enfin, l'ensemble du système de pratique

scientifique est clairement guidé par les bases philosophiques (5#) héritées de la philosophie de Marx et de la dialectique de la nature d'Engels.

Le champ "C. Garantie du système d'activité scientifique" est composé de deux thèmes principaux : "#11 Publications sur la Science de la Science et "#12 : Organisation pour la Science de la Science". Ce champ est clairement lié à la gestion de la production de la recherche scientifique (#11 : publications, périodiques de recherche), ainsi qu'à l'organisation des activités du domaine (#12 : sociétés savantes, conférences et colloques). Ces tâches soutiennent le développement réussi du domaine et garantissent sa pérennité.

5.2. *Evolution de la Science de la Science*

Au cours des 40 dernières années, 13 thèmes de recherche ont été observés dans la Science de la Science chinoise, et leur évolution, matérialisée à l'aide des dates de publications des articles analysés (voir section 4.3), est très clairement mise en évidence par notre méthode, comme présenté à la Figure 8. Elle est également justifiée par l'expertise, comme cela est décrit ci-après.

Dans les années 1980, l'activité en Science de la Science ne faisait que commencer en Chine. Le thème le plus discuté dans le monde académique était la question des attributs propres au domaine (8#). A cette époque, les chercheurs ont tenté d'identifier la nature et les modèles généraux du domaine combinant la pensée de Bernal et la réalité chinoise (7#). Lorsque la Conférence Nationale des Sciences s'est tenue en 1978, le système scientifique et technologique chinois a commencé à entrer dans une période de réforme et trois grandes revues du domaine Science de la Science (11#), ont été créées successivement : *Science Research Management* (1978), *Science of Science and S&T Management (*1980) et *Studies in Science of Science (*1983). Ces revues, poussées par le gouvernement à leurs débuts, ont rapidement attiré un grand nombre d'articles, ce qui a fait de la publication et de la gestion des résultats de la recherche un thème important.

Dans le même temps, les milieux académiques se sont intéressés à l'étude de l'histoire de la Science de la Science (10#) pour trouver des preuves de la base théorique et de la construction de la discipline en Chine en utilisant les travaux des principaux intervenants étrangers du domaine. C'est ainsi qu'ils ont jeté les bases théoriques de la politique de recherche en Chine.

Ensuite, la Science de la Science en Chine a connu une période de ralentissement, jusqu'à ce que le troisième conseil de l'Association chinoise pour la Science de la Science et la politique scientifique et technologique soient établis en 1997 (12#). Avec l'amélioration du système institutionnel de la discipline, initié par le programme de doctorat lancé la même année à l'Université de Technologie de Dalian, les méthodes d'enseignement des Sciences et la fertilisation des talents sont devenus des thèmes d'actualité au cours de l'année 2005 (2#).

Les idées philosophiques jouant un rôle clé dans le système éducatif pour guider la pratique, ce point est apparu plus important dans la fertilisation des talents (5#). Ainsi, en 2008, les milieux académiques chinois se sont intéressés davantage à l'origine philosophique de la Science de la Science, et plus particulièrement aux fondements de la philosophie de Marx.

Le développement précoce des méthodes de cartographie des connaissances scientifiques en Chine (Chen et Zeyuan, 2005) a ouvert la voie à un nouveau domaine de recherche dont le but était d'obtenir des informations sur la structure des domaines et des disciplines périphériques de la Science de la Science (4#), comme par exemple, les études scientifiques et technologiques, la théorie technologique, la philosophie technologique, la bibliothéconomie ou l'économie du savoir.

En 2012, les milieux universitaires chinois ont souligné que les activités scientifiques sont un système en soit. Par conséquent, pour accroître l'efficacité, ces activités doivent être évaluées et planifiées à l'aide d'approches systémiques scientifiques et techniques (1#). Ce changement de point

de vue contextuel s'explique par 2 raisons principales. D'une part, l'accélération du processus de prise de décision en science et technologie a nécessité une évaluation rapide et objective des intrants et des extrants de la recherche, et d'autre part, un grand nombre de nouvelles approches ont été mises au point en scientométrie, celles-ci étant susceptibles de compléter fructueusement les méthodes traditionnelles d'évaluation.

Yang Xiaolin rapporte dans son livre les propos de Wu Mingyu, un des pionniers du domaine de la Science de la Science en Chine, qui pensait que : "Les personnes qui s'engagent dans la Science de la Science devraient d'abord mettre l'accent sur le concept d'innovation"[14]. Dans cette même veine, peu après sa création en 1992, la NSFC (National Natural Science Foundation of China) a commencé à soutenir la recherche sur l'innovation. La prise de conscience du rôle central de l'innovation en Chine s'intensifiant, les thèmes de recherche liés à l'innovation s'enrichissent rapidement ("innovation technologique", "innovation globale", "innovation indépendante", "innovation collaborative", "innovation disruptive",....), ce qui fait que le management de l'innovation (3#) devient un thème très populaire en 2013.

En 2016, l'analyse scientifique quantitative (0#) et le système disciplinaire sont devenus plus populaires, et en 2017, la recherche pertinente sur la cartographie des connaissances sur la science (9#) est devenue le point culminant de la recherche scientifique de la Science de la Science en Chine[15].

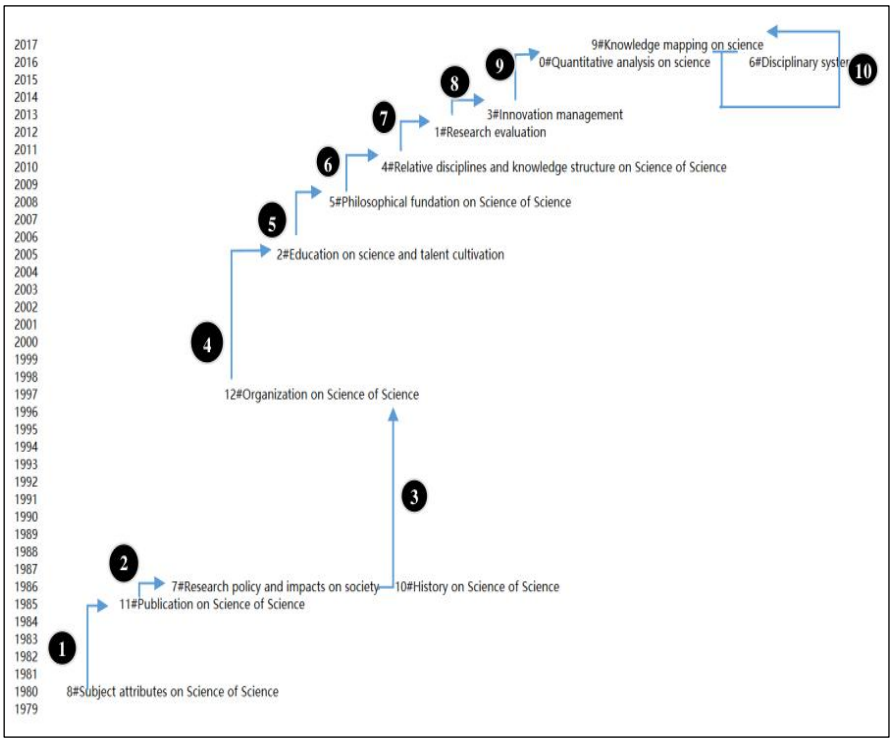

Figure 8. *Carte de cheminement des changements de thèmes de recherche dans la Science de la Science chinoise.*

[14] Yang Xiaolin. Trente ans de recherche sur les politiques scientifiques et technologiques - L'autobiographie orale de Wu Mingyu. Hunan Education Press, 2015.)

[15] A partir de 2009, 5 séminaires de formation sur la cartographie des connaissances ont été organisés au WISELAB (Université de Technologie de Dalian) diffusant largement les méthodes et réflexions en Chine. Cette approche a également conduit à la présentation de cet article en utilisant un outil de cartographie spécifique pour mettre en évidence la structure et l'évolution du domaine de la Science de la Science en Chine.

La figure 9 présente la répartition du nombre d'articles par an dans tous les groupes thématiques. Une telle approche peut être utilisée pour mettre en évidence des périodes d'activité thématique spécifiques, qu'il s'agisse d'une croissance d'activité indiquant des thèmes émergents (0#, 9#, 1#, 3#, 6#), de thèmes initiateurs du domaine caractérisés par une forte croissance initiale suivie d'une forte diminution d'activité, (#8), de thèmes qui ont eu une période significative de maturation/d'activité dans la période d'analyse (#10), voire de thèmes présentant des pics d'activité locaux pouvant être expliqués par les événements particuliers dans le développement historique du domaine (la période de reprise de la Science de la Science initiée par la création du troisième conseil du China Association for Science and Science & Technology en 1997 correspond à un tel pic local dans le thème 12#).

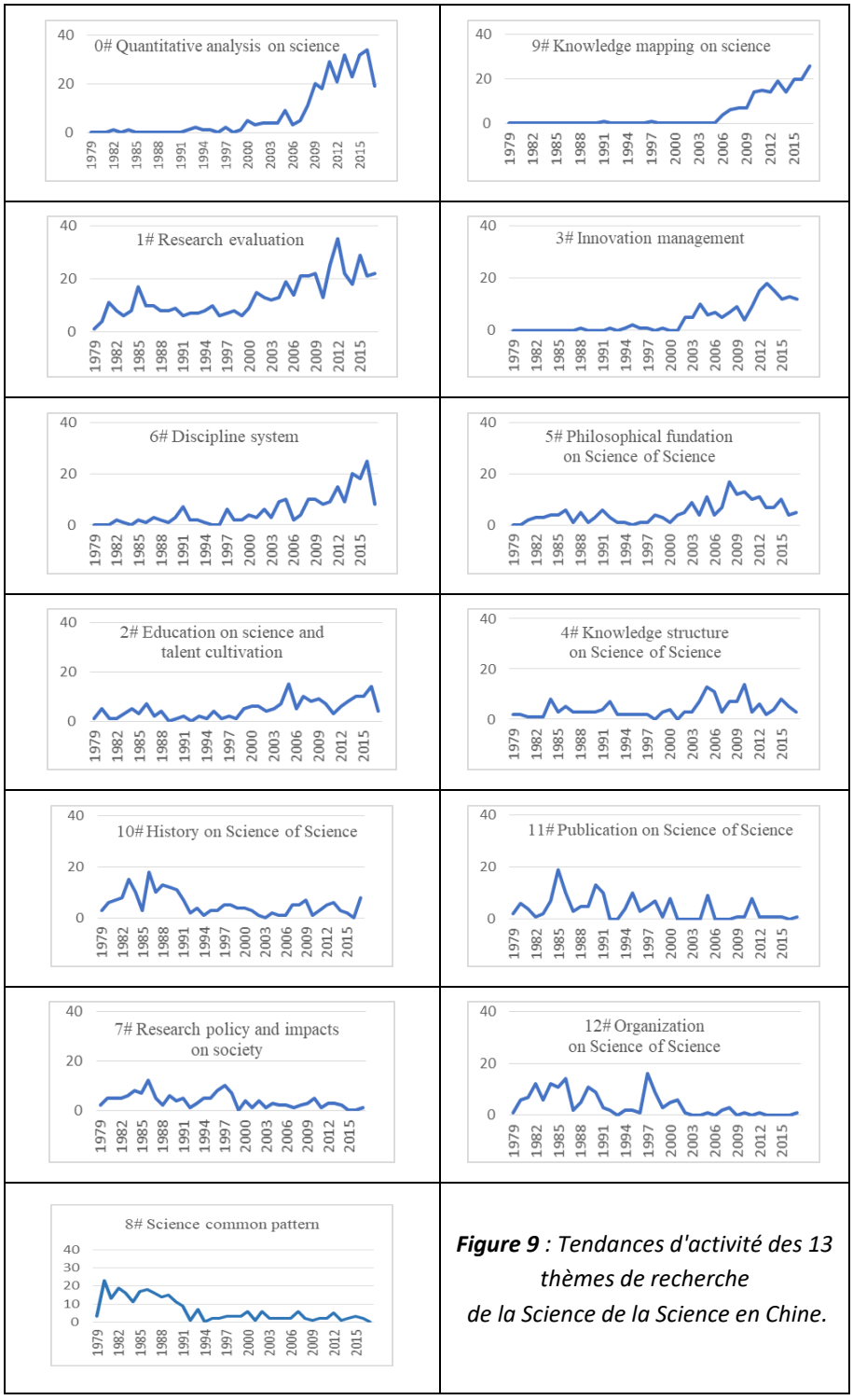

***Figure 9** : Tendances d'activité des 13 thèmes de recherche de la Science de la Science en Chine.*

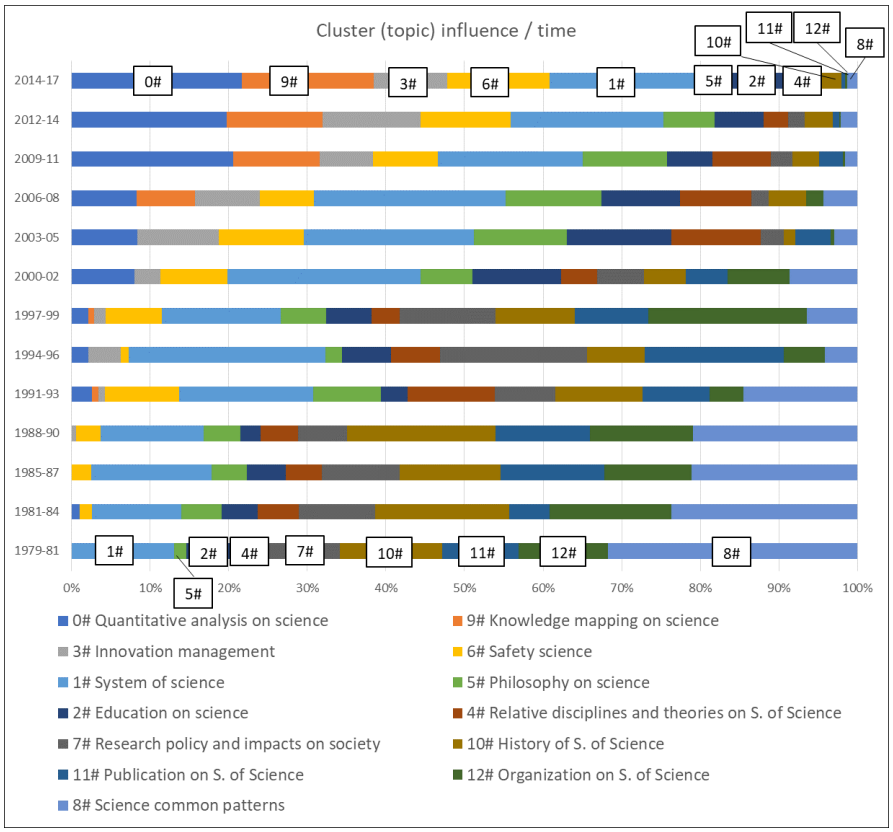

Figure 10. *Influence coordonnée des thèmes de recherche en Science de la Science en Chine.*

Une représentation plus globale de l'influence de chaque cluster (c.-à-d. thème) dans les différentes périodes (en utilisant des blocs de 3 ans) peut être dérivée des distributions précédentes. Cette représentation présentée à la figure 10 peut ensuite être utilisée pour mieux comprendre les lois de développement de la Science de la Science en Chine. Ce point de vue peut surtout aider à distinguer entre les thèmes importants mais accidentels qui ont une certaine chance de se développer à court terme et les thèmes importants rationnels qui jouent un rôle principal à long terme dans la construction du domaine.

Les thèmes "0# Analyse quantitative de la Science", "9# Cartographie des connaissances sur la science" et "3# Gestion de l'innovation" ne sont pas apparus au début de la recherche scientifique dans le domaine de la Science de la Science en Chine, mais seulement ces dernières années, le statut de ces thèmes devenant de plus en plus important. L'établissement de la position dominante du thème "0# Analyse quantitative de la Science" montre que la Science de la Science a atteint sa maturité en tant que matière. L'importance du thème "9# Cartographie des connaissances sur la science" indique que la Science de la Science s'est transformée en un sujet ouvert, intégrant les approches computationnelles et les technologies de visualisation de l'information. La prospérité croissante du thème "3# Gestion de l'innovation" montre de son côté que la Science de la Science est un thème de plus en plus étroitement lié à la pratique qui met l'accent sur la valeur économique de la science et de la technologie, et montre sa position stratégique en Chine aujourd'hui. En comparaison, les thèmes de recherche sur les attributs du domaine (8#), la construction de l'organisation scientifique et des processus de publication (#11) et la gestion des résultats de la recherche scientifique (#12) se sont progressivement affaiblis, ce qui indique également que la recherche scientifique en Science de la Science devient progressivement mature et normalisée en Chine.

5.3. *Contribution des auteurs*

C0	Freq.	Freq.T		Quantitative analysis on science
6	18	27	Jiang Chunlin	
7	8	23	Qiu Junping	
0	8	45	Liu Zeyuan	
10	5	14	Hou Haiyan	
C1				System of science
2	6	32	Wang Xukun	
7	5	23	Qiu Junping	
49	4	6	Pan Yujun	
33	4	7	Qian Xuemin	
C2				Education on science
16	4	10	Jin Lei	
0	3	45	Liu Zeyuan	
C3				Innovation management
128	3	3	Zhang Zigang	
161	3	3	Sun Rui	

Figure 11. *Identification des contributeurs à quelques un des principaux thèmes utilisant la distribution fréquentielle des auteurs dans les clusters. Les principaux contributeurs sont représentés par une surbrillance grise*

(Freq. est la fréquence de l'auteur dans le cluster et Freq. T est la fréquence globale de l'auteur dans le corpus).

La figure 11 présente un aperçu des résultats qui est possible d'obtenir concernant les auteurs les plus influents qui peuvent être caractérisés dans chaque thème en utilisant la distribution de fréquence des auteurs dans les données des clusters et en suivant la même démarche que celle adoptée pour les dates de publications. Ces auteurs sont susceptibles d'être projetés sur le graphe de contraste à une position qui dépend du nombre et de la position des thèmes dans lesquels ils sont jugés influents. Leur degré de centralité sur l'ensemble du domaine de recherche étudié ou sur une partie sélective de celui-ci peut ainsi être "visuellement évalué".

6. Conclusions et discussion

En tant que recherche théorique fondamentale orientée vers la pratique, la Science de la Science en Chine est née avec la réforme et l'ouverture de la nation. Nous utilisons dans ce travail des méthodes élaborées et originales d'analyse de données et de cartographie des connaissances pour révéler objectivement les changements historiques des thèmes de la Science de la Science en Chine et pour refléter le rôle central de ce domaine dans le processus de développement national. Notre approche a également montré que le développement rapide de l'économie chinoise et sa pratique de plus en plus active de l'innovation ont corrélativement mis en avant de nouveaux thèmes de recherche dans le domaine de la Science de la Science.

Les découvertes les plus spécifiques que les experts ont réalisée avec notre approche est que la recherche est passée d'une période de pré-maturation du sujet aux disciplines connexes et à l'analyse de la structure du savoir, de l'analyse qualitative à l'analyse quantitative et à l'analyse visuelle, de la recherche générale sur la fonction sociale à la recherche plus spécifique sur la fonction économique et la fonction stratégique.

La combinaison de la mesure de maximisation des traits et de l'apprentissage non supervisé et l'exploitation conjointe des graphes de contraste pour la visualisation est une approche originale que nous avons proposée dans ce travail. Les expériences que nous avons menées en vraie grandeur et qui ont été validées par des experts du domaine ont montré que cette méthode pouvait, sans supervision, sans paramètres et sans l'appui d'aucune source de connaissance extérieure, révéler très efficacement les thèmes de recherche, leurs interactions et leurs changements dans un domaine de

recherche très complexe comme celui de la Science de la Science en Chine. Dans cet article, nous proposons notamment une méthode de visualisation des résultats d'analyse à l'aide de la maximisation des traits. Cette méthode s'avère très adaptée à l'analyse de données à grande échelle dans des dimensions élevées. Elle tolère de plus l'intégration de nombreuses informations complémentaires qui peuvent enrichir les résultats d'analyse et permit d'obtenir une clarté et une précision de résultat que les méthodes concurrentes actuelles ne peuvent pas fournir. A titre d'exemple, des méthodes comme LDA qui seraient susceptibles de se substituer à l'approche proposée, pour la partie qui concerne l'extraction de sujets, souffrent trop fortement de la dépendance de paramètres très difficiles à contrôler et d'hypothèses de travail difficiles à vérifier sur la distribution des mots, ces problèmes endiguant sévèrement la qualité des résultats (niveau de généralité, précision).

Finalement, comme nous l'avons montré, les résultats que nous avons obtenus se sont déjà révélés suffisamment parlant pour les 2 experts que nous avons mobilisé mais mener une analyse qualitative plus poussée basée sur des questionnaires semi-directifs fournis à plusieurs experts pourrait certainement permettre une validation plus poussée de ces résultats. Par manque de temps et de moyens pour la mettre en place, et, nous réservons cependant cette étape pour un travail ultérieur.

Remerciements

Nous tenons à vivement remercier les deux experts en Science de la Science chinois que nous avons mobilisés pour ce travail et sans l'aide desquels qui se sont révélés d'une aide précieuse pour mener à bien notre analyse des résultats :

Le professeur Liu ZEYUAN est un des pionniers de la Science de la Science en Chine et un des contributeurs le plus important du domaine. Il est l'un des fondateurs des sociétés savantes en Science de la Science en Chine et par ailleurs le fondateur du WISELAB à l'université de Dalian, laboratoire lui-même pionnier dans l'étude quantitative de la Science de la Science en Chine.

Le Professeur Yue CHEN est l'actuel directeur du WISELAB. Sous son influence, ce laboratoire a développé l'exploitation des méthodes modernes d'analyse de la Science, notamment celle des méthodes cartographiques. Ce laboratoire est aujourd'hui considéré aujourd'hui comme l'un des trois principaux laboratoires chinois dans le domaine de la Science de la Science.

Bibliographie non numérotée

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.

Bernal, J. (1939). *The Social Function of Science*. London: George Routledge & Sons Ltd.

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

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1), 1–27.

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*.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), 224–227.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1), 95–104.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., et al. (2018). Science of science. Science, 359(6379). doi:10.1126/science.aao0185

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

Genxiang Pu,Renkun Di. (1998). The cognitive turn of Sociology of Science. *Journal of Dialetics of Nature*, (5), 29–34. (In Chinese)

Hongzhou Zhao,Guohua Jiang. (1983). Great facts, great subjects. Science of Science and S&T Management, Volume. (In Chinese)

Hongzhou Zhao,Guohua Jiang. (1988). Hessen Episode and the origin of Science of Science. *Studies in Science of Science*, 6(1), 14–23. (In Chinese)

Hsueshen Tsien. (1979). Science of science, Studies in Science and Technology System, *Marx's Philosophy. Philosophical Researches*, (1), 20–27. (In Chinese)

Huang J, Cheng X Q, Shen H W, et al. (2012). Exploring social influence via posterior effect of word-of-mouth recommendations, *ACM International Conference on Web Search and Data Mining. ACM*, 2012:573-582.

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.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). Oakland, CA, USA.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, 53–65.

Shen H W, Barabási A L. (2014). Collective credit allocation in science. *Proceedings of the National Academy of Sciences of the United States of America*, 2014, 111(34):12325.

Sinatra R, Wang D, Deville P, et al. (2017). Quantifying the evolution of individual scientific impact. *Science*, 2017, 354(6312): aaf5239-aaf5239.

Wang D, Barabási A. L. (2013). Quantifying long-term scientific impact. *Science*, 2013, 342(6154):127-32.

Wei Qian,Xinxin Li. (2012). J. D. Bernal and China. *Science & Culture Review*, 16–32. (In Chinese)

Xie, X. L., & Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on pattern analysis and machine intelligence*, 13(8), 841–847.

Yue Chen, Zeyuan Liu. (2005). The rise of mapping knowledge domain. *Studies in Science of Science*. 2005, 23(2):149-154. (In Chinese)

Yue Chen, Liwei Zhang, Zeyuan Liu. (2017). The prelude of the science of science in the world——The third copernican revolution initiated in Poland. *Studies in Science of Science*, 35(1), 4–10. (In Chinese)

Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., & Stanley, H. E. (2017). The science of science: From the perspective of complex systems. *Physics Reports*, 714–715, 1–73. doi:10.1016/j.physrep.2017.10.001

Zeyuan Liu, Yue Chen, Xiaoyu Zhu. (2013). D.J.Price's contribution to theory of the science of science. *Studies in Science of Science*, 31(12), 1762–1772. (In Chinese)

Zeyuan Liu. (2017). Feng Zhijun's Puzzle: What is the core theory of the science of science? *Studies in Science of Science*, 35(5), 655–660. (In Chinese)

