HAL
open science

Robustesse des classements bibliométriques, à travers la convergence des thèmes obtenus par citations et lexiques : une méthode hybride pour une représentation mixte

Alain Lelu, Michel Zitt, Elise Bassecoulard

To cite this version:

Alain Lelu, Michel Zitt, Elise Bassecoulard. Robustesse des classements bibliométriques, à travers la convergence des thèmes obtenus par citations et lexiques : une méthode hybride pour une représentation mixte. VSST 2013, Oct 2013, Nancy, France. pp.1-17. hal-00926631

HAL Id: hal-00926631

https://hal.science/hal-00926631v1

Submitted on 9 Jan 2014

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

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

Robustesse des classements bibliométriques, à travers la convergence des thèmes obtenus par citations et lexiques : une méthode hybride pour une représentation mixte

Alain Lelu* *, Michel Zitt**, Elise Bassecoulard****
alain.lelu@univ-fcomte.fr, zitt@nantes.inra.fr, bassecou@numericable.fr

(*) Université de Franche-Comté, 1 rue Claude Goudimel, 25030 Besançon (France),
(**) INRA Angers-Nantes, rue de la Géraudière 44316 Nantes (France),
(***) LORIA, Campus scientifique, BP 239, 54506 Vandœuvre-lès-Nancy (France).

Mots clefs :
Multi-clustering, scientométrie, ingénierie des connaissances, distance de Hellinger, analyse factorielle sphérique, K-moyennes axiales
Keywords:
Multi-clustering, scientometrics, knowledge engineering, Hellinger distance, spherical factor analysis, Axial K-Means
Palabras clave :
Multi-clustering, scientometría, ingeniería del conocimiento, distancia de Hellinger, análisis factorial esférica, K-medias axiales

Résumé

Alors que les méthodes d'analyse multidimensionnelles habituelles considèrent les objets étudiés sous un seul point de vue, nous présentons ici une exploration méthodologique autour de la prise en compte de points de vue multiples, dans la lignée d'une étude précédente de 4 années de nanosciences sur le WoS. Nous avons utilisé une méthode hybride entre clustering et décomposition matricielle pour comparer deux partitions des articles, l'une, lexicale, à partir de l'extraction de termes simples et composés, l'autre à partir des citations émises. Méthode qui nous a permis de placer les clusters des deux types dans un espace commun, et de conclure en présentant son apport par rapport à nos résultats antérieurs.

1 Introduction, problématique, état de l'art

L'analyse multidimensionnelle des bases de données scientifiques et techniques est possible à partir de plusieurs points de vue : les descripteurs d'un ensemble d'articles ou brevets peuvent être les citations émises (couplage bibliographique) ou reçues, les termes utilisés dans les titres, le résumé ou le corps du texte (couplage lexical), les co-auteurs, les institutions de ces co-auteurs, etc. Une riche information peut être tirée à la croisée de deux ou plusieurs de ces approches, tant du point de vue de l'analyse du contenu que du point de vue méthodologique. Il devient urgent de répondre à la nécessité pour les sciences humaines en général de ne pas se contenter d'un regard univoque sur les réalités complexes observées, regard assisté jusqu'ici par l'observation informatique, *séparément*, d'un ou plusieurs tableau(x) de données individus × descripteurs.
Pour résumer l'état de l'art en cette matière de complémentarité des points de vue, on peut dire qu'une première approche, directe, consiste à concaténer les vecteurs descriptifs de chaque point de vue : c'est ce qui a été proposé dans notre domaine d'application par [1]. Si cette approche rend bien compte des convergences, elle tend à gommer les spécificités, et le problème de pondération de chaque point de vue reste entier.
Une deuxième approche directe n'a pas ces inconvénients, mais commence tout juste à apparaître dans notre domaine d'application : il s'agit d'analyser le tableau multimode, ou tenseur (individus × descripteurs1× descripteurs2× …) par des techniques dites n-modes généralisant la décomposition aux valeurs singulières (SVD) des tableaux bimodes habituels – INDSCAL, Parafac, Tucker2-Multilinear SVD... Les problèmes de cette approche sont tout d'abord techniques, la dimension des tenseurs traitables restant limitée malgré les progrès informatiques. Les auteurs de [2] ont abordé l'analyse simultanée et la comparaison approche lexicale vs. citations en « réduisant » leurs données à un tableau à 3 entrées, c'est à dire une pile des deux tableaux (revues × revues) croisant plus de 8000 revues du WoS sous les deux aspects de similarité lexicale et d'intercitations. L'autre difficulté est d'interpréter les résultats, fournis sous forme 1) de tableaux (items × facteurs) pour chacune des n entrées, 2) d'un « tenseur-cœur » résumant dans un petit tableau à n entrées l'influence et l'interaction des facteurs extraits – équivalent n-modes de la matrice diagonale des valeurs propres du cas 2-modes habituel. Les auteurs cités ont utilisé l'espace défini par le tableau (revues × facteurs généralisés) pour partitionner les revues en 22 clusters, comparés quantitativement aux 22 catégories du WoS et illustrés par les lemmes des 5 titres les plus centraux de chaque cluster, mais n'ont pas présenté et commenté le tenseur-cœur, d'interprétation certainement délicate – la « courbe d'expérience » des chercheurs en sciences humaines pour ce type de résultats est à peine amorcée.

Comparer deux partitions d'un même ensemble.

Pour poursuivre vers notre objectif, notre problème « 3-way » (articles × termes × citations) peut être simplifié en le segmentant en deux étapes : réaliser d'abord deux classifications non supervisées séparément sur les données (articles × termes, et articles × citations) de chaque point de vue, puis les exploiter de façon « mixte » dans un deuxième temps, de façon à mettre en valeur les convergences aussi bien que les différences.

Plusieurs indices existent pour mesurer l'accord global de deux partitions [23]. Ils sont adéquats quand on s'intéresse à plus de deux partitions, ou à l'évolution de deux partitions : une valeur isolée, sans références à laquelle la comparer, comme dans notre cas, serait de peu d'utilité. Pour affiner l'analyse, on pourrait penser que traiter par Analyse Factorielle des Correspondances [3] le tableau croisé des effectifs des deux classifications permettrait d'obtenir un espace de comparaison des deux ensembles de clusters. Mais cette représentation simultanée des lignes et des colonnes d'un tableau d'effectifs croisés est illusoire et conventionnelle : si les positions relatives à l'intérieur de chaque nuage de points sont bien interprétables, rien ne peut être inféré entre points des deux types. En effet chaque point-ligne

constituant le centre de gravité des points-colonnes correspondants, et réciproquement, il faut « normaliser » chaque nuage pour obtenir une représentation simultanée lisible [4].
Pour notre part nous avons exploité dans [5] ce tableau croisé pour comparer les classifications automatiques non-supervisées (*clustering*) effectuées sur un même corpus de descriptions d'articles à partir du couplage bibliographique d'une part, du couplage lexical d'autre part, et ce en demandant le même nombre « raisonnable » de clusters (50), compromis estimé entre finesse d'analyse et possibilités cognitives d'interprétation. Notre traitement, semi-automatique, a consisté à réordonner en lignes et en colonnes le tableau des effectifs croisés des deux classifications de façon à obtenir une zone diagonale la plus chargée possible, traduite visuellement sous forme d'« archipel » ou « serpent » dans une carte plus qualitative que quantitative, comparable à un « portulan » (cf. Fig. 3). Le résultat est donc la création d'un ordre unique pour les clusters de chacun des deux types, c'est à dire une façon simple de les positionner dans un espace commun unidimensionnel.
Pour atteindre notre objectif de positionnement dans un espace commun *multidimensionnel*, et rendre davantage justice à des entités aussi complexes que des domaines de recherche, nous exploitons ici la richesse méthodologique permise par notre méthode de clustering, les K-Moyennes Axiales (KMA) [6], qui délivre un type de représentation hybride entre clustering strict (un document appartient en tout ou rien à un seul cluster), utilisé dans notre étude antérieure, et décomposition factorielle (un document est caractérisé par ses valeurs de projection – ou « centralités » – sur l'ensemble des axes factoriels extraits). Ces axes pointant vers les zones denses du nuage sphérique des données n'ont pas de raisons d'être mutuellement orthogonaux, alors que la majorité des décompositions factorielles contraignent les axes extraits à l'être. Dans le premier cas on a affaire à des axes locaux, propres au travail d'interprétation, alors que dans le deuxième il s'agit d'axes globaux, le plus souvent impossibles à interpréter individuellement au delà du 6e ou 7e. C'est cette dernière caractéristique de notre méthode – définir des clusters nuancés[1] et recouvrants – qui nous a permis de positionner les clusters des deux types dans un espace commun, celui des documents, et en particulier de calculer l'ensemble des cosinus au sein des 100 axes de clusters, cosinus intra-types comme inter-types. Nous présenterons et discuterons plus loin les caractéristiques des KMA par rapport à celles d'autres méthodes de clustering et décomposition factorielle, ainsi que les avantages et inconvénients de leur utilisation.

Nous rappellerons tout d'abord les principes des K-Moyennes, des K-Moyennes Axiales et des décompositions factorielles, puis nous établirons comment les sorties des KMA peuvent être employées dans une optique de clustering strict, ou nuancé, ou de décomposition factorielle pour le problème qui nous occupe. Nous présenterons enfin les résultats obtenus sur le corpus volumineux déjà utilisé dans [5], et en tirerons comparaisons et conclusions.

2 Les K-Moyennes

Cette méthode ancienne de clustering date des années 1950 [7] mais reste toujours d'actualité, grâce à son efficacité temporelle (en O(nombre d'individus, nombre d'attributs, nombre de clusters) × nombre d'itérations) et spatiale (en O(nombre d'attributs, nombre de clusters)), et aux différences raisonnables de taille entre clusters, qui mènent le plus souvent à des résultats favorables à l'interprétation. Le but est de maximiser la somme des variances des K clusters (K est fixé par l'utilisateur), ce qui revient à minimiser la norme de Frobenius (la racine carrée de la somme des carrés des éléments d'une matrice) de la différence entre la matrice des données $\mathbf{X}$ et son approximation d'ordre K $\overline{\mathbf{X}}_{\mathbf{K}} = \mathbf{U}_{\mathbf{K}}\ \mathbf{V}_{\mathbf{K}}'$, où $\mathbf{V}_{\mathbf{K}}$ est une matrice dont les K colonnes sont les moyennes de chaque cluster, et $\mathbf{U}_{\mathbf{K}}$ est une

[1] Nuancés, et non *flous* : la notion de flou implique une interprétation probabiliste, une incertitude sur l'appartenance d'un individu à un cluster – un document est donc caractérisé par ses probabilités d'appartenir (strictement !) à chaque cluster, probabilités dont la somme est 1. Ici « nuance » signifie que certains documents (ou termes) sont typiques d'un seul cluster, mais que d'autres peuvent être également importants dans d'autres contextes, i.e. plusieurs clusters, ou typiques d'aucun cluster.

matrice indicatrice d'appartenance (u_{ik} = 1 si l'individu i appartient au cluster k, 0 sinon)[2]. Comme les K-Moyennes réalisent un clustering strict, toutes les colonnes de $\mathbf{U_K}$ ont pour somme 1.
L'inconvénient principal est la sensibilité de l'algorithme aux conditions initiales (il converge vers des optima *locaux* de la fonction objectif), contrairement aux méthodes menant à un optimum global unique, comme la Décomposition aux Valeurs Singulières (SVD), ou menant à l'énumération d'un ensemble fini d'optima locaux, comme le font les méthodes de densité, à paramètre de finesse d'analyse constant, comme le fait aussi la SVD, respectant le principe " une représentation et une seule pour un tableau de nombres donné ".

2.1 Les K-Moyennes Axiales

Cette variante des K-Moyennes [6] utilise comme fonction objectif la somme des inerties des clusters autour des *axes principaux* de chaque cluster (somme des carrés des projections des points du cluster), et non autour de leurs *centres de gravité* comme précédemment. De plus cet axe principal résulte de son Analyse Factorielle Sphérique (AFS) [8, 9], utilisée ici dans sa variante « différences au tableau nul ».

2.1.1 L'Analyse Factorielle Sphérique

L'AFS du tableau (documents × attributs) $\mathbf{X} = \{x_{ti}\}$, de sommes en lignes et colonnes respectivement $\mathbf{x}_{t.}$ et $\mathbf{x}_{.i}$ consiste :

1 – à effectuer la décomposition aux valeurs singulières (SVD) du tableau$\{\sqrt{x_{ti}}\}$, noté $\mathbf{X}^{[1/2]}$:

$\mathbf{X}^{[1/2]} = \mathbf{U\,L\,V'}$, où $\mathbf{L}$ est la matrice diagonale des valeurs propres, $\mathbf{U}$ et $\mathbf{V}$ les matrices orthonormales des vecteurs singuliers lignes et colonnes ($\mathbf{U'\,U} = \mathbf{V'\,V} = \mathbf{I}$)
Et pour l'ordre 1 : $\mathbf{X}^{[1/2]} \approx \mathbf{u_1}\,\boldsymbol{\lambda_1}\,\mathbf{v_1'}$

2 – à calculer les facteurs AFS :

Les nuages sphériques définis par les vecteurs unitaires de coordonnées $\{ \mathrm{sqrt}(x_{ti} / x_{.i})\}$ et $\{ \mathrm{sqrt}(x_{ti} / x_{t.}\}$, de respectivement T et I points, se projettent sur les axes singuliers avec les valeurs $\mathbf{F}$ et $\mathbf{G}$, interprétables comme des cosinus, en tant que produits scalaires de vecteurs normalisés :

$$\mathbf{F} = \mathbf{Dr}^{-1/2}\,\mathbf{X}^{[1/2]}\,\mathbf{V}$$
$$\mathbf{G} = \mathbf{Dc}^{-1/2}\,\mathbf{X}^{[1/2]\prime}\,\mathbf{U}$$

où $\mathbf{Dr}^{-1/2}$ est la matrice diagonale des $\{ x_{t.}^{\ -1/2}\}$ (respectivement $\mathbf{Dc}^{-1/2}$ avec les $\{ x_{.i}^{\ -1/2}\}$)
Le calcul des facteurs F et G en découle :

$$\mathbf{F} = \mathbf{Dr}^{-1/2}\,\mathbf{U\,L}$$
$$\mathbf{G} = \mathbf{Dc}^{-1/2}\,\mathbf{V\,L}$$

Et pour l'ordre 1 :

$$\mathbf{f_1} = \mathbf{Dr}^{-1/2}\,\mathbf{u_1}\,\lambda_1$$
$$\mathbf{g_1} = \mathbf{Dc}^{-1/2}\,\mathbf{v_1}\,\lambda_1$$

A noter que dans le cas, le plus courant, d'une matrice de données non-négative, son premier vecteur propre est non-négatif, ainsi que les projections de l'ensemble des points du nuage ; ces projections sont les plus élevées pour les points les plus centraux, les plus proches du vecteur singulier qui résume au premier chef l'information contenue dans les données. Ce premier facteur peut donc être interprété comme un *indice de centralité* dans le nuage, propriété qui nous sera utile par la suite.

---

[2] Notations : les matrices sont en capitales grasses, les vecteurs en minuscules grasses, $\mathbf{X'}$ signifie la transposée de la matrice $\mathbf{X} = \{x_{ti}\}$, où t et i sont respectivement les indices lignes et colonnes.

Cette analyse possède des propriétés intéressantes :

- l'extension aux données réelles quelconques, positives et négatives, est assurée en posant que $x_{i.}$ (resp. $x_{.j}$) se définit comme la somme des valeurs absolues de la ligne i (resp. de la colonne j), x.. devient la somme totale des valeurs absolues des x_{ij} , et x_{ij} devient sign(x_{ij}) sqrt [$|x_{ij}|$ / ($x_{i.}$ $x_{.j}$ /x..)] ; nous utiliserons plus loin cette propriété pour utiliser cette méthode dans un espace des données transformé et réduit, à valeurs positives et négatives.
- étant de simples cosinus, les valeurs factorielles résident dans l'intervalle [-1:+1], [0:+1] pour des données non-négatives.
- le premier facteur peut s'interpréter comme un indice de centralité spectrale de chaque point dans son nuage [10],
- la propriété de dualité entre l'analyse des lignes et des colonnes : à l'axe principal du nuage de points des documents, exprimé par un vecteur avec une coordonnée par attribut, correspond l'axe principal du nuage des attributs, formellement symétrique, exprimé par un vecteur avec une coordonnée par document (« formule de transition») :

$$\mathbf{F} = \mathbf{Dr}^{-1/2}\, \mathbf{X}^{[1/2]}\, \mathbf{Dc}^{+1/2}\, \mathbf{G}\, \mathbf{L}^{-1}$$
$$\mathbf{G} = \mathbf{Dc}^{-1/2}\, \mathbf{X}^{[1/2]'}\, \mathbf{Dr}^{+1/2}\, \mathbf{F}\, \mathbf{L}^{-1}$$

- de cette propriété découle un « avantage compétitif » en faveur de l'intégration d'AFS dans les algorithme incrémentaux au fil de l'ajout de nouveaux documents : les facteurs **G** des attributs dérivent de façon directe des vecteurs singuliers **U** de la matrice $\mathbf{X}^{[1/2]}$, laquelle matrice se modifie par simple ajout de lignes, contrairement, par exemple à une matrice pondérée TF-IDF, dont les valeurs peuvent changer à chaque introduction de données.
- il découle de la formule $\mathbf{G} = \mathbf{Dc}^{-1/2}\, \mathbf{X}^{[1/2]'}\, \mathbf{U}$ que les facteurs **G** appliquent une correction de type TF-IDF aux vecteurs propres **U** issus des données « brutes » non pondérées : cette correction a pour effet de normaliser les distributions des mots, donc de rendre leur représentation insensible aux distributions très dissymétriques, comme les répartitions Zipfiennes. Il n'est donc pas étonnant que dans l'analyse par AFS de sous-tableaux documents × mots sémantiquement homogènes dont nous parlerons plus loin, on trouve des termes génériques ou rhétoriques fréquents en position peu centrale, en « pied d'axe », et à l'opposé en tant que termes centraux et spécifiques un mélange de termes rares et/ou plus ou moins fréquents, sans influence évidente de la fréquence des mots sur la centralité de leur position.
- la propriété d'équivalence distributionnelle que possède aussi l'Analyse Factorielle des Correspondances [3] : l'analyse des lignes est inchangée quand on fusionne deux colonnes de même profil relatif, ce qui confère de la stabilité au regard des fusions ou éclatement de termes, par exemple, qui apparaissent de façon semblable dans les mêmes contextes sémantiques [9].
- dans la publication [11] nous avons pu comparer empiriquement les quatre distances TF-IDF, de Hellinger, euclidienne sur la sphère, de compression d'information, dans le cadre du défi d'apprentissage de données textuelles DEFT 2011, dont nous avons remporté la première place ex-aequo : la distance de Hellinger l'emporte sur toutes les autres, y compris, de peu, sur la distance TF-IDF.

A noter que ce processus est formellement relié à l'analyse des correspondances (AFC), où la SVD est appliquée à la matrice transformée $\mathbf{Q} = \mathbf{Dr}^{-1/2}\, \mathbf{X}\, \mathbf{Dc}^{-1/2}$ conduisant à la décomposition : $\mathbf{Q} = \mathbf{U_{ca}}\, \mathbf{D_{ca}}\, \mathbf{V_{ca}}'$

Les facteurs CA s'écrivent : $\mathbf{F_{ca}} = x..^{1/2}\, \mathbf{Dr}^{-1/2}\, \mathbf{U_{ca}}\, \mathbf{D_{ca}}$

$$\mathbf{G_{ca}} = x..^{1/2}\, \mathbf{Dc}^{-1/2}\, \mathbf{V_{ca}}\, \mathbf{D_{ca}}$$

Du point de vue géométrique, l'AFC projette le nuage de points d'origine sur la surface d'un « simplexe étiré » tangent à la sphère unité [4], nuage dont le barycentre est pointé par le premier facteur, un vecteur trivial de nombres un (la première valeur propre correspondante est égale à un).

Ce qui contraste avec l'AFS, où les facteurs-documents, i.e. les projections des documents sur le premier axe, définissent les indices de centralité de ces documents, et le carré de la première valeur propre $\lambda(1)^2$ définit la fraction de la somme du tableau de données x.. dont rend compte la reconstitution de premier

ordre de **X** : { $x_{ti} = F_1(t)^2\, G_1(i)^2\, x_{t.}\, x_{.i} / \lambda(1)$ }, où $F_1(t)$ indique le t-ième composant du premier facteur-ligne, et symétriquement pour $G_1(i)$. Matriciellement, la décomposition exacte s'écrit : $\mathbf{X}^{[1/2]} = \mathbf{Dr}^{1/2}\, \mathbf{F}\ \mathbf{L}^{-1}\, \mathbf{G'}\, \mathbf{Dc}^{1/2}$ où **L** a sur sa diagonale l'ensemble des K valeurs propres, K étant le rang de $\mathbf{X}^{[1/2]}$.
En AFC les points d'effectifs faibles peuvent se trouver très éloignés du barycentre, d'où une surreprésentation de ces effectifs et des précautions à prendre pour l'interprétation des graphiques factoriels, problème inexistant en AFS. Si les effectifs du tableau à analyser sont forts ou moyens, les représentations obtenues par AFC et AFS sont très voisines – en effet au voisinage du barycentre les points de la sphère sont proches de ceux du plan simplexe qui lui est tangent.

- L'intérêt d'une représentation sphérique pour les données creuses comme les données textuelles ou les grands graphes a été argumentée et implantée par [12] sous le nom de Spherical K-Means, où la normalisation classique $\mathbf{x} \leftarrow \mathbf{x}/|\mathbf{x}|$ est utilisée, et non la nôtre
$\mathbf{x_t} \leftarrow \mathrm{sqrt}(\mathbf{x_t}/\Sigma_i|x_{it}|)$, où le vecteur normé est non colinéaire au vecteur d'origine, sauf exception.

- Dernier argument : sous-jacente à l'AFS est la distance de Hellinger, discutée et utilisée par quelques auteurs, comme [13] et [14]. En effet :

$d_{Hell}(i_1,i_2)^2 = 2\,(1 - \cos(i_1,i_2))$, où $\cos(i_1,i_2) = (\mathbf{x_{i1}}/x_{i1.})^{[1/2]\,\prime}\,(\mathbf{x_{i2}}/x_{i2.})^{[1/2]}$

En termes géométriques, on reconnaît la relation, pour 2 points du cercle unitaire, entre la distance de la corde et le cosinus de leur angle au centre, distance variant entre 0 pour des points confondus et 2 pour des points opposés.

2.1.2 L'algorithme des K-Moyennes Axiales :

Pour cette sous-section, nous utiliserons des notations non matricielles : les réels sont symbolisés par des minuscules grecques, les vecteurs par des minuscules latines, i est l'indice courant des colonnes, de 1 à I, t celui des lignes, de 1 à T, k celui des clusters, de 1 à K. Ainsi ξ_{ti} est la valeur située à la t-ième ligne, i-ième colonne de la matrice **X**, $\xi_{t.}$ est la somme en ligne de la ligne t, $\xi_{.i}$ la somme en colonne de la colonne i, $\xi_{..}$ la somme totale.

Initialisation

Choix au hasard de *K* axes $v(0)^{(k)} = \left\lfloor v_1(0)^{(k)},\ldots,v_I(0)^{(k)} \right\rfloor$ normalisés $\left\| v(0)^{(k)} \right\| = 1$

et de *K* scalaires $\tau_0(k) = 0$. *k* est l'indice des clusters (*k*=1,...,*K*)

Passage de *t*-1 à *t*

Pour chaque ligne $a_t = \left\lfloor \sqrt{\xi_{t1}},\ldots,\sqrt{\xi_{tI}} \right\rfloor$ du tableau de données transformé

- Calcul des *K* projections sur les axes $v^{(k)}$ soit $\eta_t^{(k)} = < v(t)^{(k)}, a_t >$
- a_t est intégré au cluster *k* pour laquelle sa projection $\eta_t^{(k)}$ est maximale
- $\tau_t(k) = \tau_{t-1}(k) + \eta_t^{(k)^2}$

- le vecteur « accumulateur d'apprentissage » $v^{(k)}$ représentatif du cluster est mis à jour $v(t)^{(k)} = v(t-1)^{(k)} + \left(\eta_t^{(k)} / \tau_t(k)\right) a_t$ et normalisé.

Test d'arrêt

La totalité des vecteurs lignes est parcourue. Les axes des clusters sont remplacés par $v^{(k)}$. Le calcul du critère $\tau = \sum_k \tau(k)$ définit le test d'arrêt. En effet, si ce critère augmente par rapport au passage précédent d'une quantité supérieure à un seuil paramétrable ou s'il n'y plus de changements d'affectation l'algorithme passe au calcul des valeurs factorielles :

- Facteurs colonne $\gamma_i^{(k)} = v_i^{(k)} \sqrt{\lambda^{(k)} / \xi_{.i}^{(k)}}$
- Facteurs ligne $\varphi_i^{(k)} = \eta_t^{(k)} / \sqrt{\xi_{t.}}$

Erreur relative de reconstitution et indicateur de qualité

Dans ces conditions, l'erreur relative de reconstitution des données est :

$$\varepsilon \cong \left(\xi_{..} - \sum_k \tau(k) \right) \Big/ \xi_{..}$$

La valeur de l'indicateur de qualité de l'analyse est par conséquent :

$$1 - \varepsilon \cong \sum_k \tau(k) \Big/ \xi_{..}$$

Avantage de l'algorithme : Il est rapide (O(nombre d'individus, nombre d'attributs, nombre de clusters) × nombre d'itérations) et consomme peu de mémoire (O(nombre d'attributs, nombre de clusters)).

Inconvénient : Le résultat varie avec l'initialisation des *K* demi axes, car la fonction objectif $\tau = \sum_k \tau(k)$ converge vers des maxima locaux, selon l'initialisation.

3 Décomposition matricielle

Un principe commun qui sous-tend la plupart des décompositions matricielles est d'exprimer une approximation d'une matrice réelle **X** à r lignes and c colonnes comme le produit de trois matrices réelles, i.e. une matrice **U de** dimensions (r, k), une matrice **L** de dimension (k, k) (sauf précision contraire), une matrice **V** de dimensions (c, k) :

X ≈ U L V'

où la dimension k résulte d'un compromis entre la précision de reconstitution et la compacité, ou l'interprétabilité de la représentation.

Le cas spécifique et important de la bifactorisation **X ≈ U V'** se ramène au schéma de la trifactorisation, en imposant **L** comme la matrice identité **I** de dimensions (k, k).

Plus généralement, les concepteurs des méthodes de factorisation disposent d'une palette étendue de possibilités pour spécifier et contraindre **U**, **V**, **L**, tout comme de pré-traiter **X**.
Pour ce qui concerne les K-Moyennes Axiales, la fonction objectif à maximiser est une somme d'inerties des clusters obtenues (somme des carrés des projections des points du cluster sur son axe), n'impliquant que les projections des points d'un cluster sur son propre axe de cluster (clustering nuancé). Mais il est possible de prendre en considération l'*ensemble* des projections des points sur l'ensemble des axes de cluster, ce qui mène d'une part à une interprétation des KMA en tant que méthode de clustering nuancée et recouvrante, par exemple en définissant l'appartenance à un cluster par un seuil sur les projections, d'autre part à un schéma de reconstitution approchée des données de type $\mathbf{X} \approx \mathbf{U}\,\mathbf{L}\,\mathbf{V'}$
En effet nous avons vu que l'algorithme calculait le vecteur - qu'on notera $\mathbf{y}_k = \left\{\eta_t^{(k)}\right\}$ - projection des points appartenant à chaque cluster sur les axes dominants de chaque cluster, normalisés, rassemblés dans **V**, autrement dit : $\mathbf{y}_k = \mathbf{X}^{[1/2]}\,\mathbf{v}_k$, où $\mathbf{v}_k = \{\, v_i^{(k)} \}$
Pour se ramener au schéma précédent, on peut considérer qu'on a créé implicitement une matrice indicatrice $\mathbf{Y}_k$ creuse, dont les valeurs non-nulles sont les projections des vecteurs-données sur l'axe 1 des clusters auquel chacun appartient strictement, d'où $\mathbf{X}^{[1/2]} \approx \mathbf{Y}_k\,\mathbf{V'}$
Mais les vecteurs-données peuvent aussi se projeter sur *tous* les autres axes, ce qui peut se traduire par la matrice pleine $\mathbf{Y} = \mathbf{X}^{[1/2]}\,\mathbf{V'}$
Ce qui suggère la reconstitution approchée : $\mathbf{X}^{[1/2]} \approx \mathbf{Y}\,\mathbf{V'}^{+}$ où $\mathbf{V'}^{+}$ est la pseudo-inverse de **V'**.
Nos essais sur diverses données ont montré que cette approximation se situait à moins de 10% d'erreur supplémentaire par rapport à celle obtenue après convergence de l'algorithme (rappelons que celui-ci optimise, au sens des moindres carrés, la somme des reconstitutions de premier ordre de chaque cluster).
A noter que sur des données réelles et volumineuses, « directionnelles » comme les données textuelles, ou binaires comme les graphes de citations, il y a peu de chances pour qu'un axe oblique soit une combinaison linéaire de plusieurs autres : l'espace des K axes obliques se confond alors avec celui des K premiers vecteurs singuliers. Et la reconstitution approchée $\mathbf{X}^{[1/2]} \approx \mathbf{Y}\,\mathbf{V'}^{+}$ donne le même résultat que celle découlant de la SVD $\mathbf{X}^{[1/2]} \approx \mathbf{W}\,\mathbf{L}\,\mathbf{Z'}$, si on note **W** la matrice des K vecteurs propres lignes, **Z** celle des colonnes, **L** la matrice diagonale des K valeurs propres correspondantes.

4 Utilisation des KMA pour l'analyse mixte citations-mots d'un corpus documentaire

4.1 En tant que clustering strict :

Dans la référence [5] nous nous sommes limités à utiliser les sorties des KMA en tant qu'appartenance en tout ou rien d'un document à un cluster. D'où un tableau carré des recouvrements entre les 50 clusters issus des citations (« c-clusters ») et les 50 clusters issus des mots (« w-clusters »). Plusieurs traitements impliquant divers indicateurs de liens entre lignes et colonnes de ce tableau (le meilleur s'étant avéré l'indice d'Ochiai), automatiques ou semi-automatiques, ont été testés pour sérier et faire coïncider au mieux ces deux types de clusters, aboutissant à des représentations graphiques en « serpent », ou en « archipel » 3D des tableaux croisés c-clusters × w-clusters pour divers indices, portés en 3e dimension z ; en effet à un c-cluster peuvent correspondre approximativement plusieurs w-clusters, et réciproquement.
On peut interpréter la représentation « archipel » (cf. figure 3) comme une tentative de dépliage unidimensionnel des clusters de chaque type, c'est à dire une synthèse globale, un compromis, entre de multiples dépliages locaux correspondant au voisinage de chaque cluster : nous allons ci-dessous présenter une synthèse moins locale, mais en contrepartie, de représentation plus complexe.

4.2 En tant que décomposition matricielle et support d'un espace commun de données :

Les tableaux factoriels pleins **Fc** et **Fw** issus respectivement des passages sur les matrices (documents × références citées) et (documents × termes) ont été concaténés en une matrice **F** (documents × (c-clusters + w-clusters)). Ainsi l'espace des documents peut-t-il servir de dimension commune pour les traitements ultérieurs mêlant c-clusters et w-clusters. Nous avons de cette façon effectué :

1) le calcul du tableau complet (100 × 100) des cosinus intra- et extra- axes c et w à partir des colonnes de **F**.

2) une réduction de cet espace par SVD de la matrice des cosinus : la figure 1 montre l'« éboulis » des premières dizaines de valeurs propres obtenues sur les données décrites ci-dessous. L'heuristique du « décrochement » (*gap*) dans cet éboulis [15] suggère que seules les 15 premières sont significatives. La transformation des données consistant à remplacer chaque vecteur-individu d'origine par ses coordonnées dans les 15 dimensions pertinentes constitue un « blanchiment » (*whitening*), selon la terminologie utilisée en analyse du signal, c'est à dire crée un nouvel espace dans lequel la variance est identique (et égale à 1) le long de toutes les coordonnées. Un tel espace est utilisé également par la méthode Analyse en Composantes Indépendantes (ICA) [16] qui y cherche les axes maximisant un indice de non-Gaussianité de répartition pour les projections des vecteurs-données centrés, en vue de résoudre des problèmes d'Analyse du Signal comme la séparation de sources (e.g. problème de la *cocktail-party* : démêler les conversations à partir de l'enregistrement du brouhaha en plusieurs points de la salle).

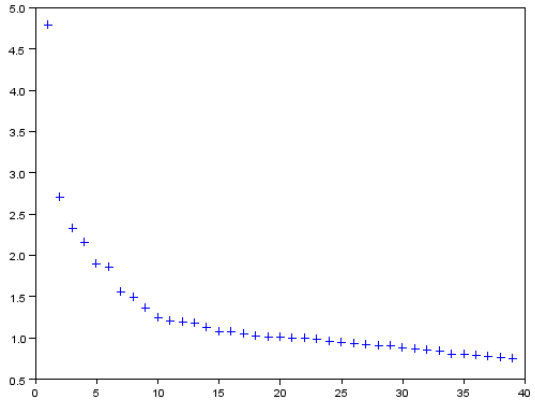

Figure 1 L'« éboulis » des valeurs singulières de la matrice 100 × 100 des cosinus inter- et intra-types de clusters.

3) un « clustering de clusters » dans cet espace réduit aux 14 dimensions U2 à U15 (la première, U1, qui sépare strictement les c-clusters des w-clusters, donc triviale pour nous, a été éliminée), via un nouveau passage KMA, sans garantie que ces nouveaux clusters incluent à la fois des c-clusters et des w-clusters (mais c'est tout l'intérêt de cette procédure que de mêler des proximités angulaires intra-c (ou –w) et extra-c (ou –w), procédure impossible via la seule analyse du tableau des recouvrements par AFC, par exemple). De cette façon nous créons autant de logiques partielles de dépliage unidimensionnel que de « thèmes », c'est à dire de clusters de clusters (« super-clusters »).

Nous avons également obtenu une cartographie commune des c-clusters et w-clusters par SVD (cf. fig. 2) et par AFC mais les deux dimensions U2 et U3 semblent insuffisantes pour tirer des conclusions de façon visuelle, les deux c- et w- sous-nuages étant relativement imbriqués.

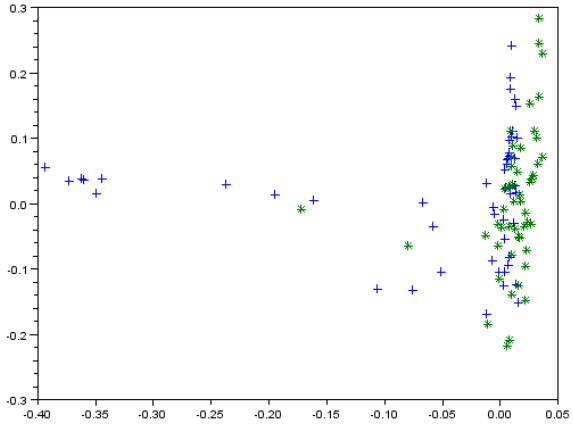

Figure 2 : nuage de points des c-clusters (*) et w-clusters (+) dans l'espace réduit commun (U2, U3)

Pour revenir à notre procédure de clustering au 2e degré, nous avons fixé à 14 le nombre demandé de « super-clusters ». En effet notre but, dans le sous espace réduit à 14 dimensions orthogonales U2-U15, est d'utiliser les K-Moyennes Axiales comme moyen rapide et approché d'extraire les 14 dimensions *obliques*, sans contrainte d'orthogonalité, pointant vers les zones de forte densité relative du nuage de points, et ce sans perte d'information puisqu'on aboutit à la même reconstitution des données (cf. section 3 plus haut), ni surestimation du nombre de clusters puisqu'il s'agit d'un minimum pour des sous-nuages de points à la surface de la sphère unité dans l'espace d'origine, donc appuyés sur la seule information angulaire. D'autres méthodes, comme la NMF (Non-negative Matrix Factorization, [17]) et ses variantes, tentent de résoudre directement le problème dans le cadre de la factorisation de matrice, mais ne le font qu'avec des algorithmes aboutissant à des optima locaux d'une fonction objectif globale, comme c'est déjà le cas pour les KMA. Elles ne sont donc pas supérieures en cela aux KMA. Nous avons vérifié sur des données-jouet que les solutions KMA et NMF étaient très proches angulairement - après transformation des données pour que la NMF se réalise dans l'espace de Hellinger décrit plus haut, ce qui n'est pas le cas de la NMF d'origine.

5 Le corpus étudié

La délimitation de ce corpus de 167 340 références sur le sujet des nanosciences publiées de 1999 à 2003 et des 375 062 références citées à au moins 3 reprises, extrait du Web of Science (Thomson-Reuters ed.), ainsi que son découpage en 50 clusters par les K-Moyennes Axiales, a été décrite dans [18]. Le vocabulaire extrait dans l'environnement NeuroNav, mots simples et termes composés, a fait l'objet de traitements multiples, en particulier d'unification des variantes de mêmes termes aussi bien automatique, que semi-automatique et manuel, aboutissant à un ensemble de 188 472 termes différents, simples et composés, apparaissant au moins 3 fois [5]. A noter que sélectionner un vocabulaire aussi étendu a permis à la fois de limiter à quantité négligeable le nombre de documents « orphelins » caractérisés par aucun terme, ainsi que celui des documents ne se trouvant indexés que par des termes triviaux ou génériques les décrivant très mal.

6 Les résultats

6.1 Résultats sur la base d'un clustering strict

Nous avons exploré dans notre publication [5] la possibilité de représenter les liens entre deux classifications non supervisées du même corpus (NanoSciences 1999-2003, les résultats sont naturellement obsolètes aujourd'hui) sous la forme d'une quasi-carte, un « portulan » croisant le découpage en clusters obtenus sur les citations avec celui obtenu sur les termes. Une sériation simultanée des lignes et des colonnes de cette matrice de recouvrements, utilisant des procédures automatiques de réarrangement, a conduit à des matrices réordonnées dont un exemple (recouvrement semi-normalisé par indice d'Ochiai) est repris ici en visualisation 2D+ (figure 3). Elle montre globalement l'accord des deux partitions, dans la mesure où la majorité des îlots sont disposés sur une quasi-diagonale : la forme de « serpent » provient de distributions différentes des tailles de clusters - à un cluster d'un type peuvent correspondre plusieurs clusters de l'autre type. Les classifications diffèrent toutefois dans le détail, il existe des îlots hors de cette séquence principale. Ceci peut-être en partie dû à une non-optimalité du réarrangement, mais en partie à de réelles divergences entre approches par les mots et les citations, appelant à l'examen approfondi et individualisé de ces îlots. A partir de l'existence de groupes partageant les mêmes références mais pas le même vocabulaire, on peut faire plusieurs hypothèses d'interprétation : cela peut être le signe de la scission d'une discipline en plusieurs branches, ou bien de domaines d'application distincts qui se reconnaissent une racine commune. Dans le cas extrême où les citations ne sortent pas de ces groupes, l'hypothèse d'un réseau de collusion pourrait être évoquée… Inversement des groupes partageant le même vocabulaire mais pas les mêmes références, peuvent être interprétés par la sociologie des citations ou comme concurrence de « chapelles » rivales travaillant sur les mêmes sujets à partir de présupposés différents… Seul un retour sur les données d'origine permettant de cerner les sous-ensembles de vocabulaire ou citations communes et spécifiques, ainsi que des éléments externes aux données recueillies, pourraient permettre d'affiner à ce point l'interprétation.

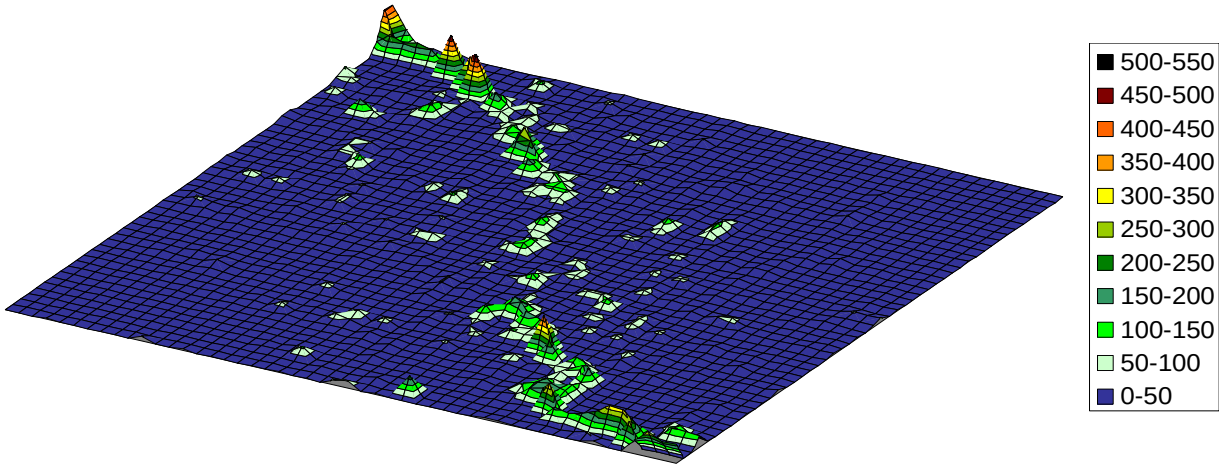

Figure 3 – Le « portulan » des liens entre les 50 w-clusters et les 50 c-clusters. Verticalement, un multiple de l'indice d'Ochiaï des liens entre paires de ces clusters

6.2 Résultats sur la base d'un espace commun de données et d'une décomposition matricielle

La matrice 100 × 100 des cosinus inter- et intra-types de clusters une fois calculée dans les seules 14 dimensions U2-U15, a été analysée par la méthode des K-Moyennes Axiales paramétrée avec K=14 super-clusters à extraire. La répartition des effectifs des clusters en résultant s'est trouvée assez homogène, entre deux super-clusters d'effectif minimum 3 et deux autres d'effectif maximum 12. L'analyse rend compte de 93,26% de la somme du tableau analysé.
Voici le contenu des super-clusters en c-clusters et w-clusters, ordonné pour chaque super-cluster par centralité décroissante (pour la signification de chaque numéro de cluster, voir l'annexe):

```
Super-cluster 1 :   C44    C47    C29    C37    C32    C4    W4    C39    C24

Super-cluster 2 :   C21    C41    C25    C38    W23    C36

Super-cluster 3 :   C17    W30    W25

Super-cluster 4 :   C2    C20    C11

Super-cluster 5 :   W45    C30    C31    W3    W41    W32    W12    W13    W42    W48

Super-cluster 6 :   W24    W17    W29    W37    C42    C8    W20    W40    W43    W44    W22    W1

Super-cluster 7 :   C48    C33    C35    C18    C50    W21    W14    W6    C26    C9    W9    C28

Super-cluster 8 :   W31    W18    W11    W2    C10    C13    W49    W5    W50    W39

Super-cluster 9 :   W35    W33    W28    C3    W27    W16

Super-cluster 10 :   C22    C46    C27    C5    C40    C6    C34    C45

Super-cluster 11 :   W38    W15    W36    W10

Super-cluster 12 :   W47    C16    C12    C49    C7    C19

Super-cluster 13 :   W8    W46    W34    C15    W19    W7

Super-cluster 14 :   C23    C1    C14    C43    W26
```

On constate que si les super-clusters 10 et 11 sont strictement intra-w ou intra–c (4 clusters sur 8 du super-cluster 10 ont été classés dans le « superthème » *Nanomaterials* de notre publication [18]), les 12 autres mélangent c-clusters et w-clusters. Cette dominance des mélanges renforce le constat établi plus haut d'une correspondance approchée *1 vers n* ou *m vers 1* entre les c-clusters et les w-clusters. Le super-cluster 12 est particulièrement typique de ce point de vue, car le cluster W47 (*mesoporous silica / carbon*) y est à la fois le seul w-cluster et le plus central de tous, à savoir des c-clusters : *Titin/protein mechanics ; mesoporous*

I/organosilica… ; mesoporous II / silica… ; DNA / mech. properties… ; biomolecular interactions…. Le super-cluster 3, plus petit, est aussi dans ce cas : le cluster C17 (*porous silicon properties*) « domine » W30 (*thin films…porous materials*) et W25 (*ion implantation/ ion beam*). Mais les autres cas de mélange d'un seul w-(ou c-)cluster avec plusieurs autres clusters de type opposé ne sont pas aussi clairs (super-clusters 1, 2, 9, 13, 14).
Parmi les 12 super-clusters mixtes, nous allons développer les exemples des super-clusters 1 et 7, à défaut de développer l'ensemble des autres, faute de place.
Le super-cluster 1 associe au seul w-cluster W4 *nanotubes* huit c-clusters. En réordonnant le tableau des cosinus selon l'ordre de centralité décroissante dans ce super-cluster, on fait apparaître visuellement la « logique locale » qu'il traduit (figure 4). A noter que sur ces huit c-clusters, sept sont proches dans l'« archipel» précédent. Le 8ème (C24 : *Fullerenes*…) a une intersection plus faible avec W4 qu'avec plusieurs autres w-clusters : ceci explique que la sériation des intersections de clusters à partir des effectifs et des indices qui en dérivent directement le place assez loin des sept autres dans l'archipel. Ce qui peut paraître comme une anomalie tient à la logique récursive sous-jacente à l'extraction de vecteurs propres : deux clusters d'intersection négligeable mais de voisinages proches entre eux - voire de voisinages de voisinages proches, etc. - pourront être considérés comme proches par notre méthode de réduction d'espace des données. C'est le même principe qui est à l'œuvre dans la méthode d'analyse sémantique latente ou LSA [19], et permet de déterminer comme proches deux documents sans aucun mot en commun, mais appartenant au même contexte sémantique. Mais cette place est en définitive justifiée puisque le Fullerène fut le premier nano-objet historique.
Le super-cluster 7 rassemble des clusters consacrés aux travaux sur la croissance épitaxique et la photo-luminescence, dans un mélange relativement équilibré de types W et C. Ce qui est en accord avec nos résultats précédents, à la différence près que notre contrainte d'un seul ordre répartissait ces clusters sur deux sous-paquets, de façon à préserver leur proximité, également importante, avec les clusters du thème « Quantum Dots » - une bonne illustration.de distorsion évitée par notre nouvelle méthode.

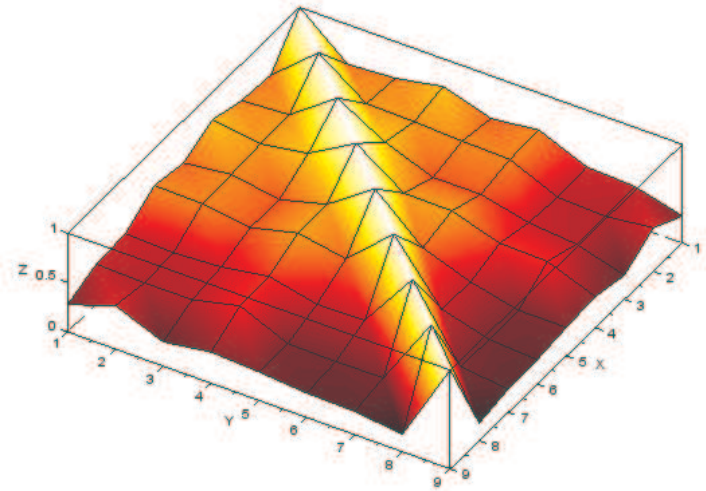

Figure 4 – Cosinus, dans l'espace réduit, entre les axes des 9 clusters du super-cluster 1, ordonnés par centralité décroissante dans ce super-cluster : `C44 C47 C29 C37 C32 C4 W4 C39 C24.`

7 Conclusions, perspectives

Alors qu'en principe les clusters recouvrants et/ou nuancés, flous, sont intéressants dans beaucoup de domaines d'application, la pratique montre qu'on fait rarement appel à eux. Les KMA laissent le choix à leur utilisateur : clustering strict, ou nuancé, ou recouvrant (via un seuil faible sur les projections des documents), ou encore détection des seuls « pôles marquants » (via un seuil fort), ou enfin décomposition factorielle oblique – mais l'expérience a montré que les options sophistiquées étaient peu utilisées. La méthode présentée ici de comparaison entre découpages de clusters issus de deux sources différentes utilise le caractère nuancé des sorties des KMA, et est à coup sûr un complément utile pour aller au delà de la représentation « en archipel » présentée précédemment. L'intérêt de plus en plus marqué pour le *multiclustering* des traces multiples d'activité issues de sources diverses, sur le Net ou ailleurs, ne devrait pas manquer de lui ouvrir des opportunités d'application.

Une généralisation importante serait de pouvoir croiser un nombre de points de vue sur les données supérieur à deux : il faudrait alors passer de la décomposition matricielle de l'espace commun à sa décomposition tensorielle, ce qui est possible au moyen de méthodes comme INDSCAL, Parafac, Tucker2, ...[20].

A plus court terme, notre expérience nous a montré que notre méthode est très sensible à la dimension de l'espace réduit dans lequel opère le clustering : une piste de perfectionnement est d'en finir avec l'heuristique de la « marche » dans l'éboulis des valeurs propres, peu évidente à mettre en pratique, en utilisant une procédure statistiquement rigoureuse pour établir la « dimension intrinsèque » de la matrice des données [21] ; une autre piste, pour échapper au problème de l'instabilité des frontières de clusters et du nombre optimal de ceux-ci, propre aux méthodes itératives de type « centres mobiles » ou « *Expectation Maximization* » (EM), serait d'utiliser une méthode densitaire, par exemple l'Analyse en Composantes Locales [6], cette fois dans l'espace réduit, telle qu'utilisée et validée pour les graphes dans [22]. C'est ce que nous nous efforcerons de transposer pour les grosses matrices de données, incontournables dans le domaine scientométrique.

Remerciements :

A Thomas Largillier pour avoir attiré notre attention sur la détection de possibles réseaux de collusion académiques.

Bibliographie

[1] VAN DEN BESSELAAR P. et HEIMERIKS G, Mapping *research topics using word-reference co-occurrences: A method and an exploratory case study*, Scientometrics 68 (3), pp. 377-393, 2006

[2] Liu X, Glänzel W, De Moor B, *Optimal and hierarchical clustering of large-scale hybrid networks for scientific mapping*, Scientometrics 91(2): 473-493, 2012

[3] BENZECRI J.-P., & coll., *L'Analyse des données 1 - L'analyse des correspondances* (Vol. 1). Paris: Dunod. 619 p.,1973

[4] GREENACRE M, HASTIE T - *The geometric interpretation of correspondence analysis*, Journal of the American Statistical Association, vol. 82, no398, pp. 437-447, 1987

[5] ZITT M., LELU A., BASSECOULARD E., *Hybrid citation-word representations in science mapping: portolan charts of research fields?*, JASIST 62, 1, pp. 19–39, 2011

[6] LELU A., *Clusters* and *factors: neural algorithms for a novel representation of huge and highly multidimensional data sets*, In E. Diday, Y. Lechevallier & al. editors. pp 241-248, Springer-Verlag, Berlin, 1994

[7] MC QUEEN J., *Some Methods for classification and Analysis of Multivariate Observations*. In: LeCam L.M., and Neyman, J. (eds). Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability., Berkeley: University of California Press, 1: 281–297, 1957

[8] DOMENGES D., VOLLE M., *Analyse factorielle sphérique : une exploration*, Annales de l'INSEE, 35-1979 :3-84, Paris, 1979

[9] ESCOFIER B., *Analyses factorielles et distances répondant au principe d'équivalence distributionnelle*, Revue de Stat. Appliquée, 26(4):29-37, Paris, 1978

[10] BRANDES U. et CORNELSEN S., *Visual Ranking of Link Structures*, Journal of Graph Algorithms and Applications 7(2):181-201, 2003.

[11] CADOT M., AUBIN S., LELU A., *Indexer, comparer, apparier des textes et leurs résumés : une exploration*, Actes de l'atelier DEFT, TALN 2011, Montpellier , 2011

[12] BANERJEE A., DHILLON. I., GHOSH J., SRA S., Clustering *on the Unit Hypersphere using Von Mises-Fisher Distributions*, Journal of Machine Learning Research, JMLR, 2005

[13] LEGENDRE P. et GALLAGHER E.D., *Ecologically meaningful transformations for ordination of species data*, Oecologia, vol 129, n°2, pp. 271-280, 2001

[14] RAO C., *A Review of Canonical Coordinates and an Alternative to Correspondence Analysis using Hellinger Distance*, Questiio (Quaderns d'Estadistica i Investigacio Operativa) 19:23–63, 1995

[15] CATTELL R. B., *The scree test for the number of factors*, Multivariate Behavioral Research, 1(2), 245-276, 1966

[16] JUTTEN C., HERAULT J., *Une solution neuromimétique au problème de séparation de sources*, Traitement du Signal, Vol. 5, N° 6-NS, p. 389-403, 1988

[17] LEE D. D. et SEUNG H. S., *Learning the parts of objects by nonnegative matrix factorization*, Nature, 401, 788-791, 1999

[18] BASSECOULARD E., LELU A. et ZITT M., *Mapping nanosciences by citation flows: a preliminary analysis*, Scientometrics, vol 70, n°3, pp. 859-880, 2007

[19] DEERWESTER S., DUMAIS S. T., FURNAS G. W., LANDAUER T. K., et HARSHMAN R., *Indexing By Latent Semantic Analysis*, Journal of the American Society For Information Science, 41, 391-407, 1990

[20] KROONENBERG P.M., *Applied Multiway Data Analysis*, John Wiley & Sons, 2008

[21] CADOT M., LELU A., *Optimization of the representation space for qualitative data: A preliminary validation on classification problems*, International Journal On Advances in Software, (2011) 4 1&2

[22] LELU A., CADOT M., Espace *intrinsèque d'un graphe et recherche de communautés*, Revue i3 (INFORMATION-INTERACTION-INTELLIGENCE) 2011 n°2, CEPADUES Editions, Toulouse, 2011.

[23] GUENOCHE A., *Partitions optimisées selon différents critères : évaluation et comparaison*, Mathématiques et sciences humaines [En ligne], 161, Printemps 2003

Annexe : liste des clusters

C-CLUSTERS (par ordre de la Fig. 5)	#	W-CLUSTERS (par ordre de la Fig. 5)	
Gene_carrier/Delivey/Therapy/Transfection/Transfer	C15	Gene_therapy/Gene_delivery/Gene_transfer	W 8
Protein/DNA_Microarrays	C 5	Drug_delivery/Pharmacokinetics/Drug_carrier/Encapsulation/Drug_evaluation	W 34
Titin/Protein_mechanics	C16	Integrin/Extracellular_matrix/Protein_expression/Cell_adhesion	W 46
Kinesin/Activity/Motors	C42	Mesoporous_silica/Mesoporous_carbons	W 47
Biomolecular_interaction/Binding/Adhesion	C19	Self-assembled_monolayers	W 45
DNA_molecules/Mechanical_properties	C 7	Metal_complexes/Ligands/Hydrogen_bonds	W 27
Mesoporous_II/Silica/Synthesis	C49	Copolymers/Block_copolymers/Polymer_architecture	W 28
Mesoporous_I/Organosilica/Ordered/Materials	C12	Nanofiltration/Membranes/Aqueous_solutions/Polyelectrolytes	W 16
Self_assembled_monolayer/thiolmonolayers	C30	Langmuir-Blodgett_films/Langmuir_monolayers/Air/water_interface	W 32
Self_assembly/Rotaxanes	C 6	Electrochemical_studies/Electrochemical_processes/Electron_transfer	W 3
Molecular_magnet	C45	Thin_films/Photoelectrochemistry/Porous_materials	W 30
Dendrimers	C22	AFM_Studies/AFM_imaging	W 24
Microchips/Biomedical_analysis /Proteomics	C40	Microscopy_techniques/AFM/STM/SPM	W 17
Porphyrin_fullerene_dyads/Fullerene_hybrids	C46	Nanoparticles/Gold/Silver_nanoparticles/TEM	W 11
Polyelectrolytes_fims/Multilayers	C31	Gold/Silver_nanoparticles/Metallic_nanoparticles	W 48
Photo_electro_chemistry/Solar_cells/photovoltaic	C36	Mechanical_alloying/nanosize_powders/Fe_compounds	W 38
AFM/STM_studies	C 8	Solvothermal_synthesis/Room_temperature_synthesis/Nanocrystalline_struct	W 15
Single_molecule_level/Spectro/Imaging	C 3	Nanopowders_synthesis/Nanoparticles_synthesis	W 44
Monolayer_protected_cluster/Gold_nanoparticles	C43	XPS_studies/Surface_studies/X-ray/Films	W 22
Photonic/Colloidal_crystals	C13	Adsorption/Interfaces/Solid-liquid	W 13
Nanocrystals	C23	AFM_studies/Thin_films/Surface_studies	W 37
Nanocomposites/Clay_polymer_hybrids	C25	Glasses/Glass_transition/Phase_transition/themosensitivity	W 26
Nanobelts/Nanoribbons/Nanowires_fabrication	C14	X-ray_studies/Grazing_incidence_X-ray	W 36
Glass_transition/Glass_forming	C27	Modelling/Simulation/Model_systems/Transport	W 39
Magnetoresistance	C26	Room_temperature/Magnetoresistance/Temperature_dependence	W 42
Molecular_electronics/First_principles	C 1	Thin_films/Ultrathin_fims/Thickness	W 29

Fullerene/C-60/Production/Formation/behavior	C24	Raman_scattering/Raman_spectroscopy/SERS	W 23
Theoretical_studies/First_principles/Ab_initio_studies	C 2	Ion_implantation/Ion_beam/	W 25
Carbo_nitride_films/MechanicalTribological_properties	C21	Molecular_dynamics_simulations	W 50
Carbon_films/Diamond_like/Amorphous	C41	Ab_initio_studies/Models/Calculations/Density_functional_studies	W 5
TiO2_surface/Particles/Films/Catalysis	C38	Amorphous_alloys/Mechanical_alloying/Heat_treatment/Thin_films/X-Ray	W 10
Density_functional_theory_studies/Self_diffusion/Modelling	C11	Surface_morphology/Thin_films/Morphology_formation/roughness	W 31
Electronic_structures/Calculations	C20	TEM_studies/Structure_studies/Nanostructures/Dislocations	W 18
Quasicrystals/Amorphous alloys	C34	Nitride_films/Sputtering/Annealing/X-Ray	W 2
Porous_silicon_properties	C17	Diamond_films/Silicon films/Nanocrystalline/Vapor_deposition	W 1
Semiconductors/Ferromagnetism	C35	Adsorption_analysis/Adsorption_desorption/Catalysis/Adsorption_methods	W 49
Diodes/Optical_properties/Applications	C28	STM_studies/	W 40
Band_Structure/Parameters/Gap	C 9	Surface_structures/Surface_grown_films/Surface_adsorption/LEED	W 20
2D-3D_ordering/Self_assembledqQuantum_dots	C50	Alloys/Grain_boundary/Microstructures/Metallic_alloys/Magnetic_properties	W 41
Carbon_nanotubes_functionalization	C37	Photoluminescence/SiO2_substrates/Light_emission/Films	W 14
Carbon_nanotubes/Hydrogen_storage	C39	Band_gap/Electronic_structures/Conduction_band/Photoemission	W 21
Carbon_nanotubes/Mechanical_properties	C32	Growth_behaviour/Epitaxial_growth/Films	W 12
Carbon_nanotubes/Production/Properties	C47	Epitaxial_growth/Thin_films/RHEED_studies	W 35
Single_walled_carbon_nanotubes/Electronic_properties	C 4	Molecular_beam_epitaxy_growth/Qdots_formation	W 33
Single_walled_carbon_nanotubes/Synthesis/Raman/Arc_method	C44	Molecular_beam_epitaxy/Epitaxial_growth/GaAs	W 9
Carbon_nanotubes_growth	C29	Photoluminescence/Optical_properties/Quantum_wells/Quantum_wires	W 6
Spin_injection/Filter/Polarization/Transport	C18	Quantum_dots/Structural_properties/Optical_properties	W 7
Quantum_dost/Magnetic_fields/Coulom/Interactions	C48	Carbon_nanotubes/Single_wall_nanotubes /Synthesis/Growth/Characterization	W 4
Quantum_dots/Bits/Josephson	C33	Carbon_nanotube_field_emission	W 43
Kondo_effect	C10	Quantum_dots/Kondo_effect/Electron_transport/Quantum_wires	W 19