Méthode d'extraction des signaux faibles

Cristelle ROUX
GFI Bénélux, Luxembourg
cristelle.roux@gfi.be

1. Introduction

Au début d'une analyse stratégique, la première question posée est très souvent la suivante :

- « Quels sont les sujets émergents du domaine étudié ? »

Elle est invariablement suivie par :

- « Quels sont les acteurs qui travaillent sur ces nouveaux sujets ? »
- « Dans quel contexte se situent ces innovations ? »

Il fallait donc trouver une méthode simple et fiable pour, très rapidement, répondre à ces questions qui conditionnent le reste de l'étude.
Notre approche a été très pragmatique et elle se décompose de la façon suivante :
Rechercher la terminologie émergente si possible dans le texte libre (multi-termes) plutôt qu'au niveau des mots-clés,
L'extraire au dessus d'un seuil d'appartenance à la période la plus récente (minimum deux fois la valeur attendue),
Etablir la matrice de cooccurrence de cette terminologie émergente,
Trier convenablement cette matrice pour faire ressortir des classes sémantiques,
Extraire le contenu de chaque classe homogène ainsi trouvée en précisant s'il est extrait d'un ou plusieurs documents,
Rechercher les documents pointés par chaque classe,
Les soumettre aux experts du domaine et les aider dans leur interprétation car, comme il s'agit d'un nouveau concept, ils sont souvent totalement incompétents sur ce point spécifique.
Nous avons très souvent utilisé cette méthode dans des études rétrospectives et nous avons pu montrer que les signaux faibles sont détectables bien avant le réel décollage d'un nouveau concept qu'il soit scientifique, technologique ou économique.

2. Algorithmes de tri de grandes matrices

Tri par blocs sur les liens absolus

Cette technique a de nombreuses applications :

- Comme précédemment recherche de classes connexes,
- Pour chacune des classes, un tri interne par blocs permet de regrouper directement les éléments les plus liés,
- Réorganisation d'une matrice connexe en blocs diagonaux.

Son utilisation en analyse de textes permet, comme ci-dessous, de détecter les classes sémantiques émergentes les plus marquées. Nous partons pour cela de la matrice de croisement des nouveaux termes (extraits suivant la procédure évoquée plus bas). Cette terminologie émergente peut éventuellement former des groupes correspondants à des concepts émergents. Un seul terme ne suffit pas, car il peut s'agir d'une évolution terminologique qui consacre un concept déjà ancien qui, maintenant, bénéficie d'un vocabulaire spécifique (souvent un mot simple remplace ainsi une expression ou un mot composé).

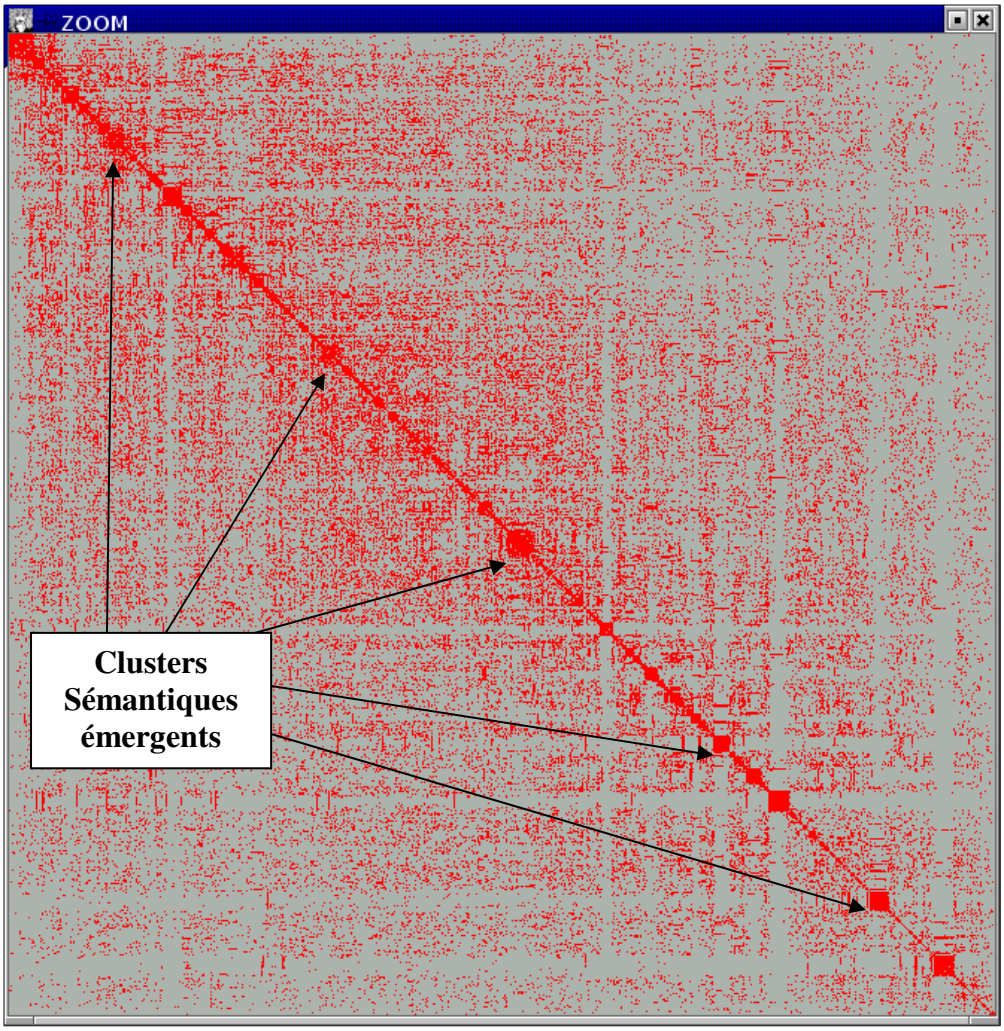

Figure 1. : Tri par blocs diagonaux sur une matrice de cooccurrence sémantique.

La matrices ci-dessus regroupe déjà près de 25 millions de cellules (5000 x 5000) et nous avons déjà travaillé sur des matrices de10 000 lignes et colonnes.

Tri par blocs sur les liens relatifs

Cette technique est utilisée lorsque les termes croisés ont des fréquences très différentes. En effet, dans les textes sont mêlés des termes courants ou très utilisés dans le domaine à d'autres beaucoup plus précis qui ciblent des spécificités. Si nous voulons découvrir les groupes sémantiques qui correspondent à ces sujets émergents ou rares, nous devons préalablement passer en mode relatif avant de faire le tri. Remarquons que, pour les matrices de cooccurrences symétriques croisant des modalités exclusives (par exemple : auteurs ou mots-clés), les éléments diagonaux représentent en fait les fréquences dans le corpus. Nous devons procéder de même pour des croisements asymétriques entre deux variables distinctes posant les mêmes problèmes de dispersion de fréquences. Nous proposons plusieurs méthodes pour passer en mode relatif :

- Division de chaque élément de la matrice par la racine carrée des éléments diagonaux qui lui correspondent, nous obtenons alors une matrice à diagonale unitaire (cas symétrique uniquement). Ce principe fonctionne bien sur les matrices sémantiques et tient compte des liens faibles.

$$S_{ij}=\frac{a_{ij}}{\sqrt{a_{ii}\,a_{jj}}}$$

- Division du carré de chaque élément par les éléments diagonaux, nous obtenons alors une matrice d'équivalence elle aussi à diagonale unitaire (cas symétrique uniquement). Cette méthode est très utilisée pour analyser les réseaux sémantiques, mais elle a tendance à pénaliser les liens faibles : c'est en fait le carré de la similarité précédente et donc une valeur de ½ ne représente plus ici que ¼ .
- La similarité de Kulzinsky est de même ordre que l'équivalence, mais la moyenne des fréquences vient remplacer un des facteurs du numérateur. Elle est utilisée dans la détection des réseaux sémantiques associés aux signaux forts.

$$S_{ij}=\frac{a_{ij}(a_{ii}+a_{jj})}{2\,a_{ii}\,a_{jj}}$$

- Nous pouvons estomper l'effet réducteur des deux propositions précédentes en utilisant l'indice dit de proximité, qui est obtenu en divisant chaque terme de la matrice par les éléments diagonaux associés (cas symétrique uniquement).

$$S_{ij}=\frac{a_{ij}}{a_{ii}\,a_{jj}}$$

- Toujours dans le cadre des matrices symétriques, nous pouvons utiliser la similarité d'inclusion qui rend compte, si elle s'approche de 1, du fait qu'un terme est toujours relié à un autre ou qu'un auteur appartient exclusivement à un équipe dont le directeur signe toutes les publications. Cette métrique est très utile pour faire la différence entre les éléments spécifiques à un groupe et ceux qui interfèrent avec les autres groupes.

$$S_{ij}=\frac{a_{ij}}{\min(a_{ii},a_{jj})}$$

- Division par la racine carrée des marginales : ce procédé est applicable aux matrices asymétriques. Comme les marginales sont toujours supérieures aux élément diagonaux, cette méthode a tendance à pénaliser les termes très fréquents (mots outils, termes généraux, termes de l'équation de recherche), elle privilégie donc les termes rares qui sont fréquemment en cooccurrence. Il est donc possible de détecter certains signaux faibles (groupes cohérents de termes peu répandus).

$$S_{ij}=\frac{a_{ij}}{\sqrt{a_{i\bullet}\,a_{\bullet j}}}\quad avec:\ a_{i\bullet}=\sum_j a_{ij}\ \ et\ \ a_{\bullet j}=\sum_i a_{ij}$$

- Division par la norme des lignes (ou des colonnes). Cette méthode de réduction permet d'uniformiser une des deux variables, les modalités sont alors de même taille et l'effet fréquentiel est estompé.

$$S_{ij}=\frac{a_{ij}}{N_n(L_i)}\quad avec:\ N_1(L_i)=\sqrt{\sum_j a_{ij}^2}\ \ ou\ \ N_2(L_i)=\sum_j |a_{ij}|\ \ ou\ \ N_3(L_i)=\max_j(|a_{ij}|)$$

- Division par le maximum de la ligne (ou de la colonne) comme dans le cas de la norme n°3 ci-dessus. A remarquer que pour une matrice symétrique, la diagonale devient unitaire car, initialement, elle est dominante dans nos matrices.

Nous avons conservé deux techniques dans Tétralogie.

- La première consiste à normaliser la matrice, donc la modifier, puis à la trier. Elle a l'avantage du choix de la normalisation, mais détruit les valeurs initiales de la matrice.
- La seconde est basée sur une normalisation compatible avec les matrices non symétriques, elle trie la matrice en fonction des nouvelles valeurs, mais conserve les anciennes. Donc seule la structure de la matrice change mais plus les valeurs.

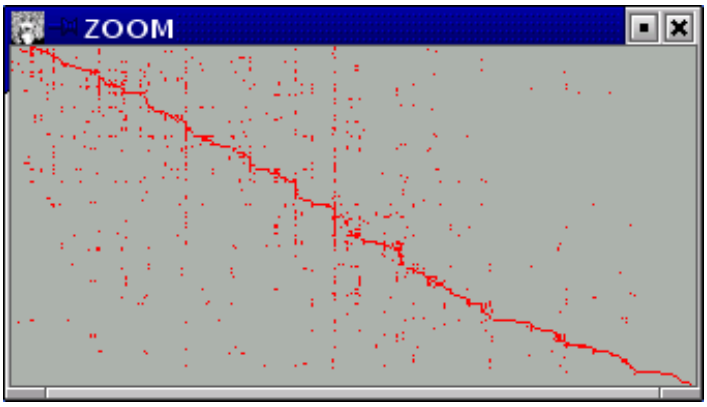

Figure 2. : Tri par blocs diagonaux d'une matrice asymétrique Auteurs - Journaux.

Dans l'exemple ci-dessus, nous détectons les chapelles d'un domaine de recherche à partir d'une matrice Auteurs – Journaux triée par blocs en mode relatif :

Extraction automatique des classes

Etant donnée la très grande taille de certaines des matrices analysées et le nombre important de classes (clusters) mis en évidence, il nous a semblé opportun de rechercher une technique automatique permettant d'isoler chacune d'elles. Comme ici, les éléments à agréger arrivent séquentiellement pour former la diagonale ou la pseudo-diagonale dominante de la matrice, il suffit de détecter les sauts de ressemblance pour isoler chaque classe de la suivante. Une baisse de cette mesure traduit, en effet, l'absence dans le reste des items non classés d'éléments susceptibles de venir compléter la classe en cours d'élaboration. Un seuil convenablement choisi permet alors de réaliser un découpage efficace, seuls les classes ayant suffisamment d'éléments seront ensuite analysées.

3. Extraction d'informations stratégiques

Extraction interactive d'information : les émergences

Outre la visualisation en 4D, un de nos apports les plus appréciés au niveau des méthodes d'analyse multidimensionnelles est l'introduction de la variable temps à de nombreux niveaux de l'exploration. Voici une méthode d'extraction des émergences utilisant les manipulations interactives sur une AFC réalisée en fonction de la variable temps :

- Croiser la variable à analyser avec le temps exprimé en périodes aux effectifs suffisamment homogènes (rapport au plus de 1 à 2) ,
- Faire une AFC de la matrice obtenue,
- Visualiser la carte des modalités temporelles (colonnes seules),
- Par des rotations, manipuler le nuage jusqu'à isoler la dernière composante temporelle dans un coin de la fenêtre (dans la figure suivante : 1997 en haut à gauche),
- Visualiser la carte globale (variable à analyser plus le temps),
- Exporter, vers cette carte, l'azimut trouvé dans la première,
- Extraire les items qui se trouvent au delà ou à proximité de l'icône associé à la dernière période (en orange sur la carte 4D),
- Générer le filtre contenant toutes les modalités émergentes de la variable analysée.

Ce filtre peut ensuite être réutilisé pour croiser les émergences entre elles et trouver ainsi les concepts émergents.

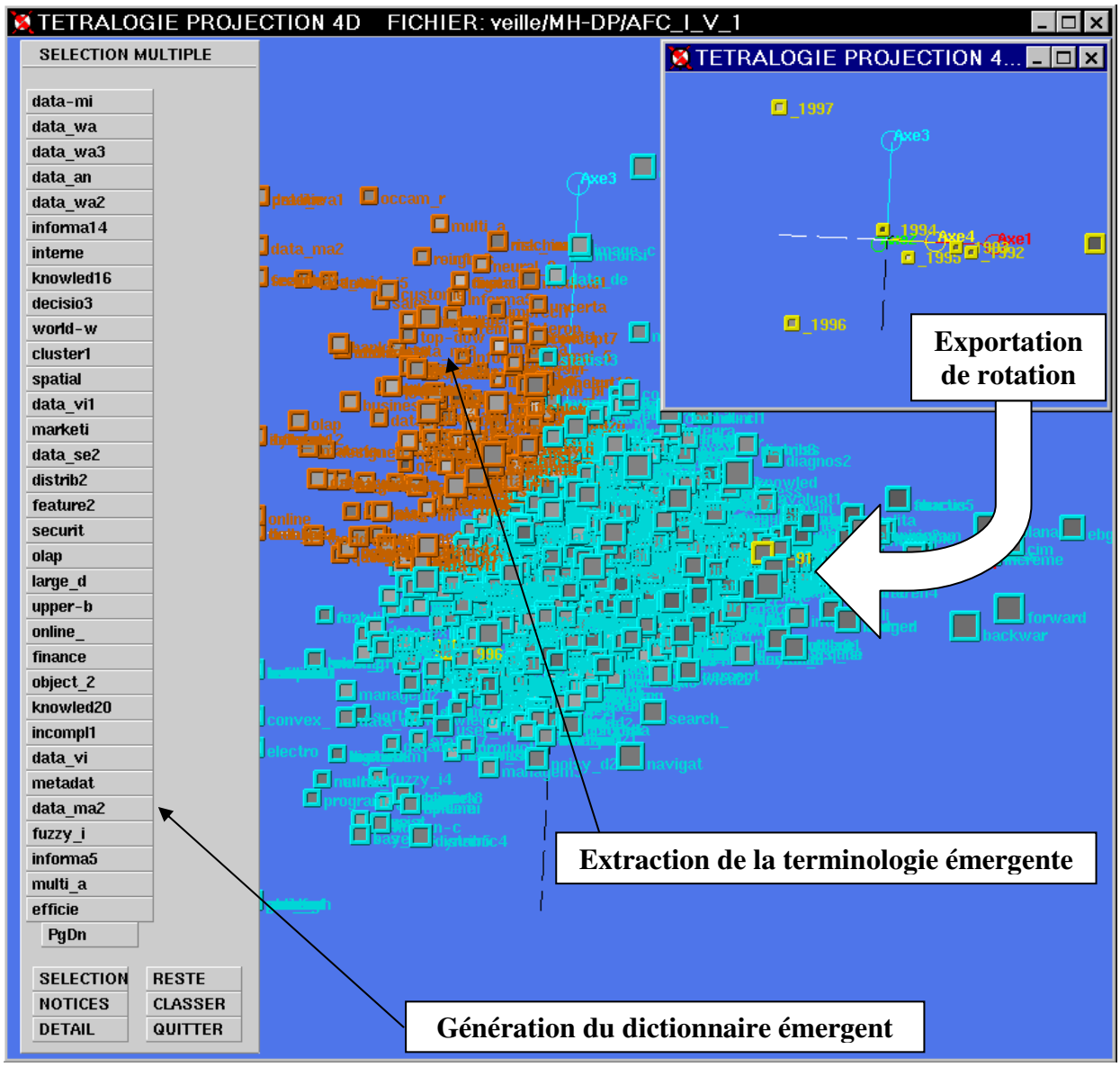

Figure 3. : Extraction d'éléments émergents basée sur une AFC Thématique – Temps.

Nous allons, par la suite, étendre ce type de démarche à d'autres stratégies de découverte de connaissances essentiellement basées sur l'interactivité. L'outil de visualisation servant à décupler les facultés sensorielles de l'utilisateur, qui, par ses capacités de déduction et sa maîtrise du sujet, est le seul à pouvoir amener l'analyse à son terme.

Détection des signaux faibles

Cette méthode, très appréciée des décideurs, consiste à extraire des classes sémantiques émergentes qui représentent ce qui ce fait de nouveau dans un domaine donné. Pour cela, nous devons :

- Partir d'une matrice Mots-clés – Dates ou mieux Multi-termes – Dates,
- Extraire la terminologie émergente comme ci-dessus
- La croiser avec elle même (matrice carrée de cooccurrences),
- Trier cette matrice par blocs diagonaux,
- Extraire les classes les plus visibles,
- Demander le détail (liste des termes connectés entre eux),
- Recroiser le tout avec les autres champs (contexte).

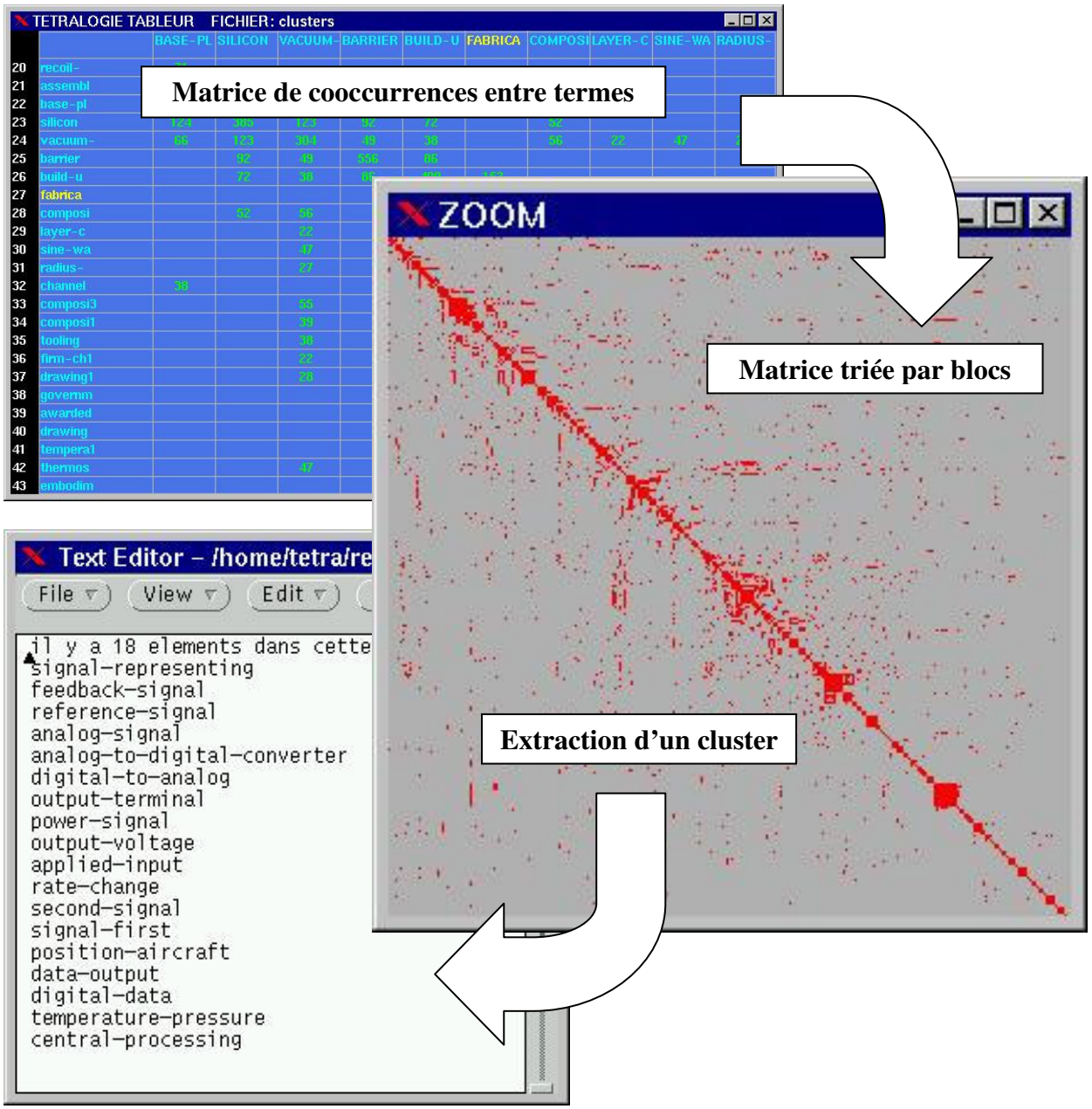

Figure 4. : Illustration de la méthode d'extraction de signaux faibles.

Le résultat dépasse souvent toute prévision, car les concepts sous-jacents sont complètement nouveaux, ce qui déstabilise les experts, qui s'avouent bien souvent incompétents en la matière [**ROUX98**]. Les nouveaux sujets, ainsi détectés, doivent bien entendu faire l'objet d'un zoom détaillé, qui peut être obtenu en croisant leur terminologie spécifique avec les acteurs du domaine et les autres concepts. Il est aussi souhaitable de re-interroger les bases d'information sur ce nouveau thème (dont l'équation de recherche nous est donnée), afin de compléter sa carte d'identité et de mieux en cerner la potentialité.

Phénomènes de rupture

La disparition brutale d'un sous domaine, d'une équipe, d'un acteur majeur peut être une information stratégique. La consultation d'une matrice ayant le temps comme seconde variable est souvent suffisante (histogramme d'évolution, classification en fonction du temps, tri d'une colonne par consistance). Par contre, lorsqu'il s'agit de mettre à jour une réorientation thématique, un changement d'alliance ou tout simplement l'arrêt d'une collaboration, il est nécessaire de faire intervenir deux

variables et le temps. On se tourne alors vers l'analyse des matrices 3D et l'ensemble des méthodes que nous que nous avons développées pour cela.

4. Conclusion

Cette méthode nous a permis, dans chacune de nos analyses, de détecter les concepts émergents des domaines que nous avons étudiés. La validation de la méthode a été réalisée grâce à des études rétrospectives qui ont mis en évidence des émergences (constatées à posteriori) avec souvent plusieurs années d'avance. Nous avons ainsi pu démontrer que les facteurs de l'innovation sont présents dans les informations ouvertes bien avant de pouvoir les détecter et les identifier par les techniques plus classiques. Le problème qui reste à résoudre est celui de l'interprétation de ces concepts émergents (groupes de termes simples ou multiples émergents de façon cohérente dans quelques documents) car les experts ne connaissent bien évidemment pas ces nouveaux domaines à moins d'y être personnellement impliqués. A chaque fois, il faut donc rechercher le contexte d'apparition de ces signaux faibles, c'est à dire les domaines connexes et les acteurs impliqués.