Approches statistiques et sémantiques pour la recherche des signaux faibles

Bernard Dousset

Approches statistiques et sémantiques pour la recherche des signaux faibles

Bernard DOUSSET Professeur émérite IRIT/ SIG, Université Paul Sabatier
118 route de Narbonne, 31062 Toulouse cedex 09, bernard.dousset@irit.fr

***Résumé* — Nous présentons dans cet article une méthode d'extraction de signaux faibles basée sur une double approche. Dans un premier temps un traitement sémantique permet de détecter tous les multi-termes utilisés dans l'ensemble des documents analysés qu'ils soient issus d'une base homogène ou de la fusion de plusieurs bases hétérogènes et ce par traitement du texte libre enrichi par le vocabulaire contrôlé (mots clés, thésaurus, …). Dans un second temps, ne sont retenus que les termes récents à forte densité dans certains documents. Ce vocabulaire retenu est ensuite croisé avec lui-même dans une matrice de cooccurrences qui est ensuite triée par blocs afin d'en extraire des clusters sémantiques cohérents et nouveaux. Ces clusters correspondent à des signaux faibles qu'il est ensuite facile de valider en les croisant avec les autres champs : auteurs, laboratoires, pays, journaux, reste du vocabulaire.**

***Termes d'indexation* — signaux faibles, clusters, mots-clés, tris de matrices, multi-termes, innovation.**

I. 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 sujets ? »
- « Dans quel contexte se situent ces innovations ? »

Il fallait donc trouver des méthodes simples et fiables pour, très rapidement, répondre à ces questions qui conditionnent le reste de l'étude.

Notre approche est 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 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.

II. TRIS DE GRANDES MATRICES

A. 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).

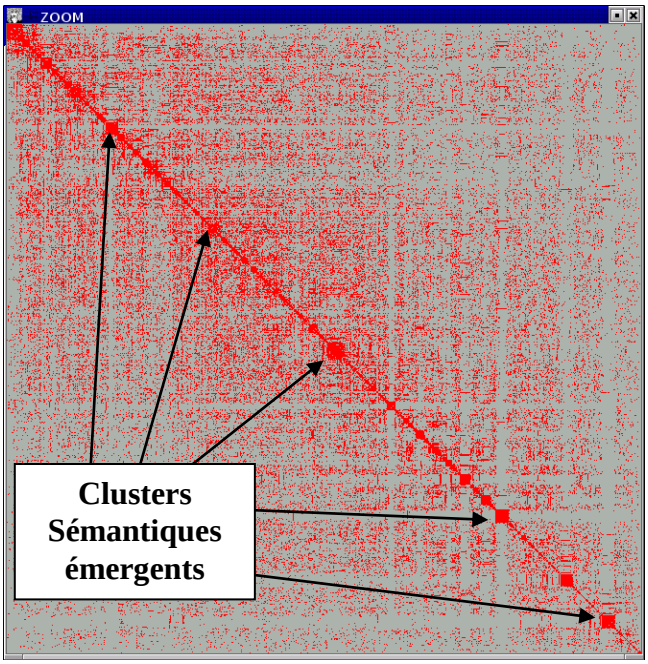

Figure 1. : Tri par blocs diagonaux sur une matrice de cooccurrence sémantique.

La matrice ci-dessus regroupe déjà près de 25 millions de cellules (5000 x 5000) et nous avons déjà travaillé sur des matrices de 10 000 lignes et colonnes.

B. 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 éléments 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.

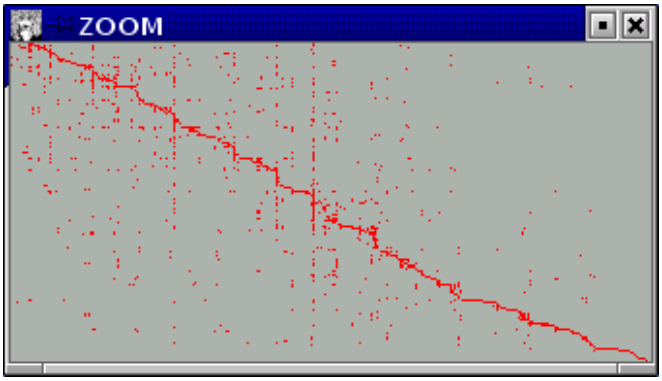

Figure 2. : Tri par blocs diagonaux d'une matrice asymétrique Auteurs - Journaux.

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 :

C. *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.

III. EXTRACTION D'INFORMATIONS STRATEGIQUES

A. *Extraction interactive des 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.

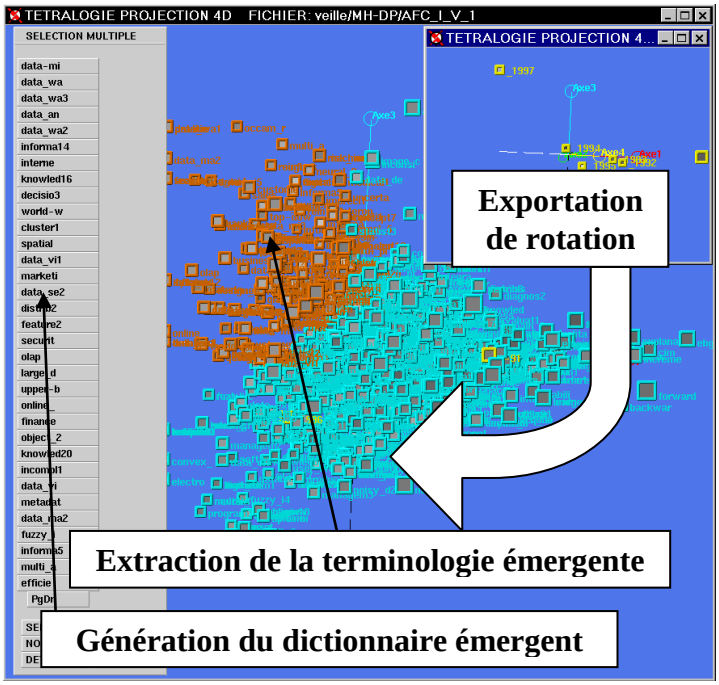

Figure 3. : Extraction d'éléments émergents basée sur une AFC Thématique – Temps.

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.

B. *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).

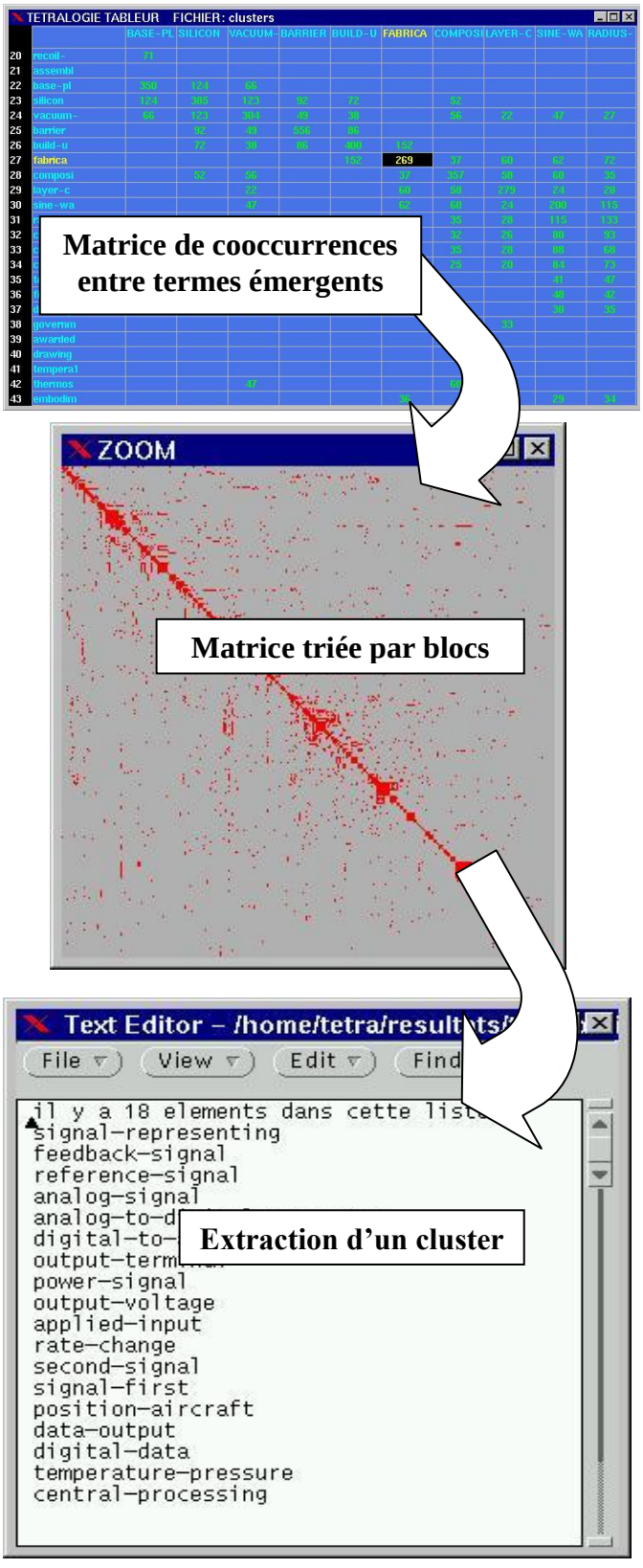

Figure 4. : Illustration de la méthode d'extraction de signaux faibles.

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 réinterroger 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.

C. 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.

D. Bilan

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.

E. Amélioration possible

Il reste toutes fois une difficulté due au traitement sémantique initial des textes que nous devons analyser. Sont-ils issus de champs contrôlés : mots clés, thésaurus, indexations de tous types ou bien du texte libre : titre, résumé original ou retravaillé, texte intégral, commentaires. Les sources traitées sont-elles hétérogènes (plusieurs bases en même temps) car il est alors illusoire de vouloir fusionner les différents champs d'indexation. En effet, non seulement les mots-clés ne sont pas à jour, mais ils typent les documents en fonction de leur source (vocabulaires d'indexation quasiment disjoints). La solution que nous préconisons est de réaliser une indexation automatique et homogène des champs en texte libre (titre, résumé, texte intégral) s'appuyant sur un thésaurus à jour du domaine étudié et qui est généré par l'analyse syntaxique, sémantique et statistique de la totalité des champs traitant du sujet sur l'ensemble des sources. Ainsi, un mot-clé (venant d'une base ou d'un auteur) pourra alors servir d'index à tous les documents du corpus.

IV. NOTION DE MULTITERMES

La terminologie rencontrée dans les champs sémantiques peut se décomposer en trois entités :

- Les mots simples ou unitermes : le dictionnaire de la langue au sens propre du terme.

- Les radicaux auxquels il est possible de ramener certains unitermes par des algorithmes de radicalisation (stimming, algorithme de Porter, …).
- Les mots composés, syntagmes ou expressions : suite ordonnée d'unitermes comme : « analyse de données », « état de l'art ». Ils peuvent éventuellement être reliés par des tirets et/ou suivis de leur acronyme comme « bovine-spongiform-encephalopathy (BSE) » ou « Creutsfeld Jacob desease CJD ».

Ces derniers sont bien entendu beaucoup plus précis et leur valeur sémantique leur permet de faire office de mots-clés. Aussi, doit-on les rechercher systématiquement dans les textes et, si possible, en générer un dictionnaire à jour qui va servir de base à l'indexation automatique et à l'analyse sémantique.

Habituellement nous procédons comme suit :

- Générer les dictionnaires de tous les champs sémantiques (thesaurus, mots-clés, index, classifications, termes d'indexation des auteurs, titres, résumés, texte intégral, …).
- Fusionner ces dictionnaires et dé-doublonner.
- Ne garder que les mots composés sans leurs acronymes,
- Générer un dictionnaire de multi-termes de la spécialité (suite d'unitermes séparés par des espaces).
- Eventuellement générer un dictionnaire de synonymes notamment pour prendre en compte les acronymes avec ou sans les variations morphologiques (inversion, terminaisons, pluriels, multilinguisme, …).

Cette première phase permet d'extraire, d'un volumineux corpus, tout l'aspect conscient de l'information dite explicite qui comprend :

- Le conscient collectif : terminologie sur laquelle tout le monde est d'accord et essentiellement représentée par la notion de mots-clés.
- Le conscient individuel : mots composés signalés par certains auteurs et détectables par la présence de tirets et/ou d'acronymes.

Mais il s'agit d'un vocabulaire convenu qui n'a pas une grande utilité pour détecter l'innovation et les sujets tout juste émergents. Il sert toutefois à parfaitement cibler les grands axes du domaine ainsi que leurs interconnexions.

V. DETECTION D'UNE NOUVELLE TERMINOLOGIE

Dans une seconde phase, nous allons rechercher l'ensemble des multi-termes qui ne sont signalés par personne, car totalement nouveaux dans le domaine, et qui représentent, à nos yeux, le front de recherche (ou d'innovation) encore inconscient, mais qui se trouve à l'état de traces dans les textes que nous analysons.

Pour cela, nous recherchons de façon statistique quelles sont les expressions qui reviennent suffisamment souvent (au moins deux fois) et qui sont absentes du dictionnaire « conscient » précédent. Cette détection nous permet d'accéder à une information qui n'est ni pointée par les mots-clés traditionnels, ni proposée par les auteurs, car trop récente ou non encore officiellement reconnue.

Elle présente elle aussi deux niveaux bien distincts :

- L'inconscient collectif : une même expression se retrouve chez plusieurs auteurs, mais personne ne la signale comme mot-clé éventuel (ces auteurs ont donc dû lire ou écouter un message qui les a séduits, ils sont d'accord mais ne le savent pas encore. Il y a donc un consensus inconscient).
- L'inconscient individuel : un même auteur utilise plusieurs fois dans ses écrits une nouvelle expression, il s'agit d'un « segment répété » qui représente l'idée importante de son discours.

Bien entendu, ces deux niveaux d'information sont beaucoup plus subtils que les précédents, ils nous permettent d'accéder à la nouveauté ou à tout ce qui touche au marginal. Leurs croisements avec tous les autres éléments d'information vont nous permettre de savoir quels sont les acteurs concernés, dans quels axes apparaissent ces nouveaux concepts et éventuellement s'ils s'inscrivent dans des stratégies visibles. Les propositions faites par cette méthode automatique de détection doivent être validées en deux temps :

- Tout d'abord, par les documentalistes, afin d'éliminer certaines expressions usuelles de la langue et des éléments trivialement inutiles.
- Ensuite, par les experts du domaine, afin de ne garder parmi les multi-termes détectés (expressions, mots composés, molécules, sigles, …) que ceux ayant une réelle valeur sémantique dans le cadre de la problématique étudiée.

Mais bien souvent, nous détectons également des évidences qui ont échappé à tout le monde, notamment aux indexations traditionnelles, et qui ont une importance stratégique indéniable.

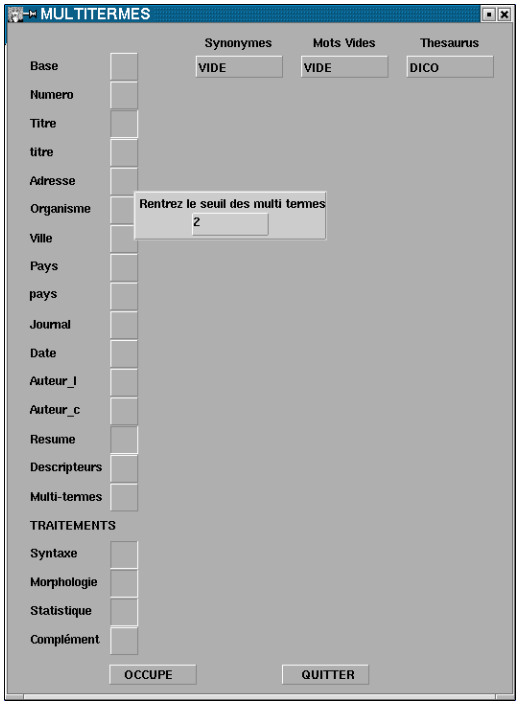

Figure 5. : Génération des multi-termes

VI. CREATION D'UN CHAMP D'INDEXATION COMPLEMENTAIRE

Nous allons générer, dans chaque document, un nouveau champ d'indexation reprenant la présence de la terminologie simple ou composée au niveau des différents champs

sémantiques. Cette génération s'appuie sur le dictionnaire validé des multi-termes, sur un dictionnaire de mots vides à éliminer ainsi que sur un dictionnaire de synonymes d'unitermes. Elle produit un nouveau champ d'indexation (cette fois-ci homogène) qui reprend la liste des éléments détectés.

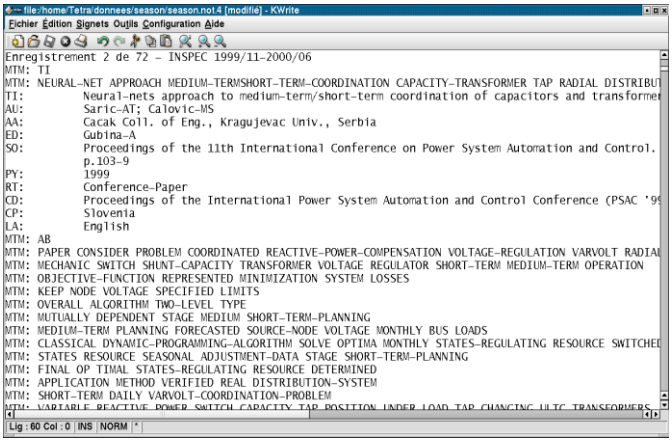

Figure 6. : Ajout d'un nouveau champ d'indexation MTM

Il est possible d'y adjoindre certaines options :

- Détection statistique pour générer le dictionnaire « inconscient ».
- Traitement syntaxique pour tenir compte de la ponctuation qu'un multi-terme ne peut franchir.
- Traitement morphologique (radicalisation pour éviter d'utiliser de trop gros dictionnaires de correspondance).

Mais outre les quatre niveaux d'information cités plus haut, le principal intérêt de cette indexation est d'être homogène pour l'ensemble des sources utilisées. Il est donc possible, au sein d'une analyse multi-bases, d'utiliser les qualités sémantiques offertes par ces nouveaux mots-clés. De plus, les sources collaborent et donc améliorent mutuellement l'indexation commune, car une expression détectée dans l'une d'entre elles permet d'indexer correctement toutes les autres, de même pour un multi-terme découvert grâce à sa présence sur au moins deux supports différents.

VII. CHOIX DE LA TERMINOLOGIE OPTIMALE

Le principal problème posé par le champ d'indexation complémentaire des multi- termes est la très grande taille du dictionnaire qui permet de le générer. Souvent plusieurs dizaines de milliers d'entrées. Il n'est donc pas toujours possible de charger en mémoire l'ensemble des croisements sémantiques qu'il propose. Nous avons donc pensé à simplifier ce dictionnaire en ne conservant que les termes les plus représentatifs, c'est à dire ceux qui typent le mieux les classes sémantiques du domaine et qui créent le moins possible de connexions non significatives. Pour cela nous éliminons :

- les mots vides de la langue,
- les termes qui ne sont présents qu'une fois dans le corpus (apax),
- ceux qui sont distribués uniformément sur l'ensemble des documents (termes de l'équation de recherche, mots usuels ou trop généraux, …),

Mais ce nettoyage n'est pas toujours suffisant pour ramener le nombre de multi-termes à un volume exploitable en mémoire. Nous avons alors recours à une technique mise au point par un de mes étudiants en DEA [**KANOb98**] et qui consiste à ne conserver que les termes qui ont une forte densité dans certains documents. Pour cela nous calculons le rapport entre densité locale (dans chaque document) et densité globale (sur l'ensemble du corpus) et nous qualifions prioritairement les éléments dont le rapport est important dans au moins deux documents. En abaissant progressivement le seuil de qualification nous pouvons ainsi générer un dictionnaire de la taille désirée. Cette procédure permet de détecter le cœur des classes sémantiques par leur terminologie la plus typique. Les liens éventuels entre classes sont alors dus à des termes précis et non plus provoqués par des expressions courantes sans grande signification.

Dans la figure suivante nous illustrons le principe de qualification des termes :

- Nous calculons la densité relative d'un terme dans chaque document,
- Le seuil est initialement fixé très haut,
- On abaisse progressivement ce seuil,
- Dès qu'un terme dépasse ce seuil sur au moins deux documents, il est qualifié,
- On arrête le processus de qualification dès qu'un nombre fixé de termes qualifiés est atteint.

Quelque soit le volume imposé aux dictionnaires et aux matrices, on est sûr d'avoir choisi les termes qui génèrent les classes sémantiques les plus précises. Une fois trouvé le noyau de chaque classe, il est toujours possible de requalifier certains des termes écartés afin, notamment, de retrouver le contexte et d'enrichir le réseau sémantique. Il suffit pour cela de réaliser un simple croisement entre la classe formée de termes qualifiés et l'ensemble des autres multi-termes.

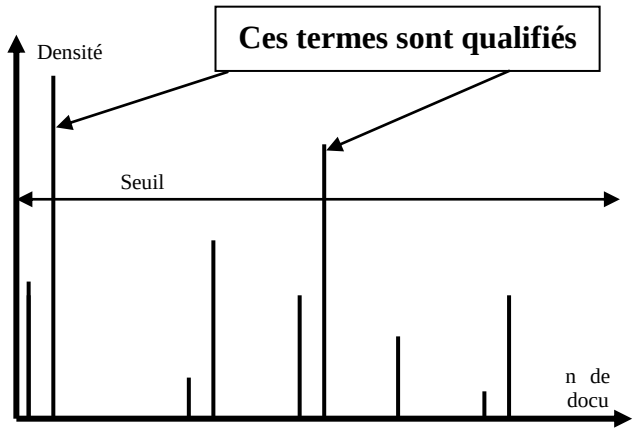

Figure 7. : Qualification d'un multi-terme dans le dictionnaire à conserver

VIII. UTILISATION DU CHAMP MULTI-TERMES

Un nouveau champ est alors ajouté au corpus, il est composé de la liste des multi-termes retenus et présents dans chaque document indexé. Les multi-termes sont reliés par des tirets et séparés par un espace. Il suffit de décrire ce nouveau champ de la base dans les métadonnées pour qu'il soit reconnu par les fonctions d'extraction et de croisement d'information.

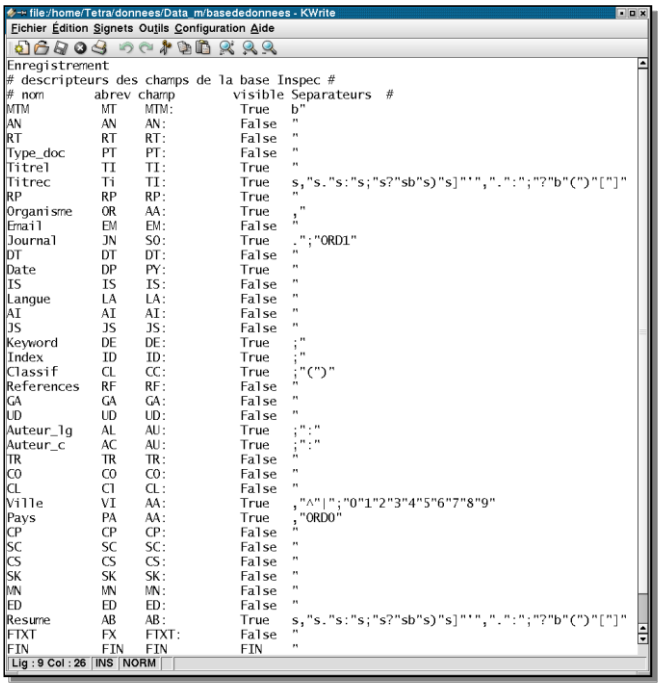

Figure 8. : Prise en compte du champ multi-termes dans les métadonnées.

Il sera alors utilisé comme un champ natif. Son utilité est multiple :

- Il représente un champ d'indexation à jour :
  - conscient collectif (CC)
  - conscient individuel (CI)
  - inconscient collectif (IC)
  - inconscient individuel (II)
- Il est homogène dans un environnement multi-base
- Il peut contenir des termes très spécifiques :
  - formules chimiques
  - expressions
  - sigles complexes

Il se substitue donc aux champs d'indexation hétérogènes.

IX. EXEMPLES DE DICTIONNAIRES MTM

A. Dictionnaire initial (conscients collectif et individuel)

Ce dictionnaire est issu de la compilation des tous les champs d'indexation et de ceux en texte libre qui indiquent des mots composés.

ACID COMMON INORGANIC ANION
ACOUSTIC EMISSION AE
ADJACENT HIGH MELTING POINT
ADSORBED CARBON BLACK
ADSORBED LAYER PVP
ADVANCED ANALYSIS METHOD
AGREEMENT EXPERIMENTAL DATA
ALL ALUMINUM CYLINDER
ALLOY PARTICLE MELT
ALLOY PARTICLE PAD
ALUMINIUM ALLOY FOAM
ALUMINIUM ALLOY SHEET
ALUMINUM ALLOY SHEET
AMINO ACID COMMON INORGANIC ANION
AMPHIPHILIC BLOCK COPOLYMER
ANALYSE CIRCUMFERENTIAL THROUGH WALL

B. Dictionnaire complémentaire (inconscients collectif et individuel)

Ce dictionnaire est généré par étude statistique sur la redondance dans le corpus d'expressions qui ne sont pas présentes dans le dictionnaire précédent.

hydrogen-trapped
plastic-non-symmetric
load-displacement
molecular-building-block
organic-inorganic
polyurethane-foam
poly-paraphenylene
iron-fibre
oxygen-facilitated
molecular-crystal
ethyl-cellulose
aluminium-alloy-foam
accurate-determination
knitted-fabric-composite
autofrettaged-pressure-vessel
solvent-mediator
temperature-reduction
shape-recovery
sol-gel-process
porphyrinosilica-template
iron-porphyrinosilica
sandwich-shell
foamed-metal
perfect-shell
sheet-yielding
optimally-designed
elastic-deformation
functionally-graded
integrated-process
sixth-conference-hole-burning-related-
platinum-tetra-pentafluorophenyl-porphine-pttfpp
temperature-dependence-atm
pressure-temperature
aqueous-solution-poly

X. EXTRACTION DE CLUSTERS SEMANTIQUES

Voici quelques clusters sémantiques obtenus à partir du croisement des multi-termes les plus typés et tri de la matrice de cooccurrence obtenue par blocs diagonaux :

CLUSTER 1
composite-beam
thermal-expansion
embedded-shape-memory-alloy
thermal-buckling
shape-recovery
temperature-reduction
laminated-composite

CLUSTER 2
solvent-polymeric
liquid-chromatographic
metalloporphyrin-based
porphyrin-based

anion-selectivity
potentiometric-selectivity
lipophilic-anionic
membrane-electrode
membrane-electrode-based
anion-selective
ion-selective-electrode
paper-presented
mnppix-psp4
quaternary-ammonium
excellent-selectivity
cationic-site

CLUSTER 3
manganese-tetraphenylporphyrin
serum-sample
show-high-selectivity-histamine-amino-acid-common-carrier-poly-vinyl-chloride-membrane-potentiometri
histamine-synthetic
near-nernstian-response-concentration-range-detect
surface-graphite-electrode
applied-determination

CLUSTER 4
chain-transfer-agent
controlled-molecular-weight
immortal-polymerization
lewis-acid
narrow-molecular-weight-distribution
turnover-number
polymerization-methyl-methacrylate
aluminum-porphyrin-initiator
living-anionic-polymerization
growing-specy
aluminum-porphyrin
acid-assisted
anionic-polymerization
proceeded-rapidly
living-polymerization
ring-opening-polymerization
molecular-weight-distribution
ester-group
micellar-aggregate
two-stage

CLUSTER 5
catalytic-oxidation
development-biomimetic-oxidation-catalysis
structural-feature
thiolate-ligand
n-oxide
drug-metabolism-study
polypeptide-bound
ruthenium-porphyrin
iron-prophyrin
polymer-complex
recent-study
high-yield
extremely-high
one-step
porphyrin-complex
high-selectivity

XI. CONCLUSION

Cette méthode d'indexation s'impose dans le cas d'analyses multi-bases afin d'homogénéiser les termes du vocabulaire d'indexation et ainsi d'harmoniser les contributions de chaque document. En effet, les formats peuvent être très différents ainsi que la qualité et la disparité de l'indexation initiale quand elle existe. Dans le cas d'une seule base, son intérêt reste grand car les indexations proposées ne sont pas à jour et il est alors très difficile de détecter des signaux faibles. Enfin, certains termes techniques ne sont en général pas indexés comme, notamment, les formules de chimie, les gènes ou des expressions très longues sans compter la possibilité de détecter le copier-coller, opération qui peut être pleine d'enseignements. Le recours à la détection automatique et à la génération des multi-termes améliore d'une façon sensible la détection des signaux faibles. En effet, l'innovation s'exprime dans un premier temps par l'utilisation de mots assez standards du vocabulaire existant via de nouvelles expressions composées: les multi-termes. Par la suite, si l'innovation devient un sujet de premier plan au niveau recherche ou développement, certaines de ces expressions sont rebaptisées et figées dans le vocabulaire, mais il est beaucoup trop tard pour s'apercevoir qu'il s'agit d'une innovation majeure. Citons pour exemple la « mort cellulaire programmée » qui est devenue des années plus tard l'apoptose.

XII. BIBLIOGRAPHIE

[1] DOUSSET B., Integration of interactive konwledge diecovry for environmental scanning. Phd. Rapport, University Paul Sabatier, Toulouse, (2003).

[2] EL HADDADI A., DOUSSET B., BERRADA I., Securing a competitive intelligence platfor

m, conference communication, INFORSID 2010.

[3] EL HADDADI A., DOUSSET B., BERRADA I., LOUBIER I., The multi-sources in the context of competitive intelligence , EGC 2010, P A1-125 A1-136, Tunisie, (2010).

[4] GHALAMALLAH I., GRIMEH A., DOUSSET B., Processing data stream by relational analysis, EUROPEAN WORKSHOP ON DATA STREAM ANALYSIS, March, 14-16, CASERTA, ITALY, N° 36 , pp 67-70 (2007).

[5] GHALAMALLAH I., A proposed model of exploratory multivariate analysis in competitive intelligence, thesis rapport, Toulouse university, Dec. (2009).

[6] HATIM H., EL HADDADI A., EL BAKKALI H., DOUSSET B., BERRADA I., Generic approach to control access and tratement in a competitive intelligence platform, conference communication, VSST 2010.

[7] SOSSON D., M. VASSARD, DOUSSET B., Portal for navigation in the strategic analysis , VSST'01, Vol 1, pp 347-358, Barcelone – Espagne, (2001).

[8] ROUX C., DOUSSET B., Une méthode de détection des signaux faibles: application à l'émergence des Dendrimères. Veille stratégique, scientifique et technologique : VSST'98, pp 349-357, (Toulouse 3, France), (1998).