L'ANALYSE DES MOTS ASSOCIES POUR L'INFORMATION ECONOMIQUE ET COMMERCIALE

EXEMPLE SUR DES DEPECHES "REUTERS BUSINESS BRIEFING"

B. Delecroix
France Telecom
CESD/ISIS - Université De Marne la Vallée

R.Eppstein
CESD/ISIS - Université De Marne la Vallée

Résumé : L'analyse des mots associés se fonde sur une théorie sociologique principalement développée par le CSI et le SERPIA (Callon, Courtial, Turner) au milieu des années 80. Cette analyse mesure la force d'association entre les termes d'un corpus documentaire pour dégager et mettre en évidence l'évolution d'un domaine scientifique à l'aide de la construction d'agrégats de termes (clusters) et d'un diagramme stratégique. Depuis ces premiers travaux, la méthode des mots associés a été employée avec succès pour aider à comprendre la structure et l'évolution de nombreux champs scientifiques. Aujourd'hui, elle est implantée dans plusieurs logiciels qui sont utilisés par les entreprises dans des systèmes d'aide à la décision leur permettant d'améliorer leur compétitivité et de définir leur stratégie. Cependant, il n'existe pas de travaux permettant de valider cette méthode dans ce type de contexte.
Au travers d'un exemple d'analyse d'un corpus d'information à caractère économique et marketing sur les technologies DSL tiré de "Reuters Business Briefing", cette présentation donne une interprétation aux résultats obtenus par l'analyse des mots associés. Après un rapide aperçu du logiciel utilisé (Sampler), et après avoir clairement défini le protocole expérimental utilisé, nous examinons chaque étape du processus en œuvre lors de l'analyse des mots associés : extraction terminologique, calcul des clusters et du diagramme stratégique. Nous analysons le rôle de chaque paramètre de cette méthode avant de donner une interprétation plus générale de la méthode des mots associés dans un contexte d'information économique tirée de dépêches d'agences. D'autres travaux viendront compléter ceux-ci afin de généraliser les résultats obtenus.

Abstract : Co-word analysis is based on a sociological theory developed by the CSI and the SERPIA (Callon, Courtial, Turner) in the middle of the eighties. It measures association strength between terms in documents to reveal and visualise evolution of science through the construction of clusters and strategic diagram. Since, this method has been successfully applied to investigate the structure of many scientific fields. Nowadays it occurs in many software systems which are used by companies to improve their business and define their strategy but the relevance in this kind of application has not been proved yet.

Through the example of economic and marketing information on DSL technologies from Reuters Business Briefing, this presentation gives an interpretation of co-word analysis for this kind of information. After an outlook of the software we used (Sampler) and after a survey of the experimental protocol, we investigate and explain each step of the co-word analysis process : terminological extraction, computation of clusters and strategic diagram. In particular, we explain the meaning of every parameter of the method. Finally we try to give global interpretation of the method in an economic context. Further studies will be added to this work in order to allow a generalisation of these results.

Mots Clés : classification, mots associés, intelligence économique

Keywords : clustering, co-word analysis, competitive intelligence

l'Analyse des mots associes pour l'Information economique et commerciale

Exemple sur des dépèches "Reuters Business Briefing"

INTRODUCTION

Beaucoup d'entreprises utilisent la veille et l'intelligence économique pour développer leur activité et mieux appréhender leur environnement. Parmi les logiciels présents sur ce marché, nombre d'entre eux mettent en œuvre la méthode des mots associés. Il existe de nombreuses références sur cette méthode d'analyse (Callon, courtial, Turner, Bauin, 1983; Courtial, Callon & Laville, 1991; Courtial, 1994; Law & Whittaker, 1992). La méthode des mots associés révèle la structure d'un champ scientifique et en dessine l'évolution en mesurant la force d'association entre les expressions représentatives du corpus considéré. Cette méthode n'a donc pas été conçue pour l'analyse de l'information économique ou financière et en particulier des informations provenant de dépêches d'agences de presse.

Le corpus de 700 dépèches extraites de Reuters Business Briefing traite des technologies DSL, secteur économique hautement compétitif durant ces derniers mois. Que devient "l'analyse de l'évolution d'un champ scientifique" sur un tel type de document? Dans cette étude, nous tentons de valider l'utilisation de cette méthode et de fournir une interprétation aux résultats obtenus dans ce contexte.

1 - L'ANALYSE DES MOTS ASSOCIES

La méthode des mots associés réduit un vaste espace d'expressions à de multiples espaces plus petits et donc plus facilement compréhensibles et interprétables qui sont également représentatif des relations entre thèmes abordés dans l'ensemble des documents considérés. Cette analyse nécessite la définition d'une mesure d'association, mais également la mise en œuvre d'un algorithme permettant la mise en évidence de ces thèmes.

Différentes mesures d'association ont été étudiées (Callon, Law, & Rip, 1986; Grivel et François, 1995), mais la plus couramment employée reste l'indice d'équivalence. Deux termes i et j cooccurrent s'ils sont employés simultanément dans un document. Considérons un corpus de N documents. Chaque document est indexé par un ensemble de termes ou d'expressions qui peuvent apparaître dans de nombreux documents. Soit C_k le nombre d'occurrences du terme k, Soit C_{ij} le nombre de cooccurrences des termes i et j, alors l'indice d'équivalence E_{ij} est défini par :

$$E_{ij}=\frac{C_{ij}^2}{C_i * C_j} \text{ , où } 0 \le E_{ij} \le 1$$

Cette mesure permet la mise en évidence d'associations fortes, en tenant compte de la fréquence d'apparition de chaque terme dans le corpus considéré. Ainsi deux termes qui apparaissent souvent dans le corpus mais simplement quelques fois ensemble seront représentés par un lien plus faible que deux expressions apparaissant moins souvent mais systématiquement de façon simultanée. Le réseau de termes est constitué de nœuds (les termes) interconnectés par des liens représentant leur force d'association respective. L'algorithme utilisé produit deux types de liens. Les liens internes représentent les relations les plus fortes entre expressions ; les liens externes qui représentent les associations plus faibles et servent à mettre en relation les clusters. Chaque cluster est donc formé par l'ensemble des expressions liées par un lien interne. Les clusters sont liés entre eux par les liens externes.

2 - LE LOGICIEL SAMPLER

Sampler est un logiciel d'analyse lexico-statistique développé par la Cisi (Jouve 1996), aujourd'hui filiale du Groupe CS Communication & Systèmes. Ce logiciel se fonde sur les recherches du *Service d'Etude et de Réalisation de Produits d'Information Avancés (SERPIA)* et du *Centre de Sociologie de l'Innovation (CSI) de l'Ecole des Mines* (Callon&Courtial&Turner,1983) et sur le développement par ces équipes du logiciel Leximappe (Michelet 1988). Dans Sampler, l'analyse des mots associés est complétée par une analyse morpho-syntaxique capable d'extraire des unitermes mais également des multitermes, amenant ainsi une réduction de la polysémie.

Les termes et expressions sont agrégés par la méthode des mots associés pour former des réseaux d'associations au sein desquels il est possible de naviguer graphiquement. Chaque réseau, également appelé "cluster", ne correspond aucunement à une structure sémantique mais plutôt à des associations contextuelles entre expressions. Les clusters ne sont pas orientés et contiennent les deux types de liens décrits dans la méthode des mots associés. Les liens internes représentent les associations où l'occurrence de chaque terme est peu différente de la cooccurrence. Les liens externes représentent les associations de termes qui apparaissent dans un contexte différent. La construction des clusters est effectuée par l'application d'un algorithme de Classification Ascendante Hiérarchique utilisant l'indice d'équivalence comme distance entre expressions.

3 - PROTOCOLE EXPERIMENTAL

Au total, 800 dépêches en langue anglaise concernant les technologies DSL (Digital Subscriber Lines) sur une période de six mois (d'Octobre 2001 à Avril 2002) ont été extraites du service "Reuters Business Briefing". L'équation exacte de recherche utilisée est "dsl OR adsl OR xdsl OR digital subscriber lines".

Pour rendre notre analyse pertinente, un nettoyage complet des données récupérées a été effectué. Nous avons tout d'abord effacé les doublons (le service Reuters Business Briefing collecte les dépêches depuis de nombreux fils de presse sans effectuer ce travail) pour obtenir 700 dépêches uniques. Nous avons ensuite nettoyé ces documents afin d'éliminer les mots, balises ou expressions interférant avec l'analyse tels que le nom de l'auteur, la ville depuis laquelle a été émise la dépêche ou bien encore le nom de l'agence de presse émettrice.

Les paramètres du logiciel Sampler ont été fixés comme suit :

- Nombre minimum de cooccurrences : 3
- Nombre minimum d'occurrences : 3
- Nombre maximum de liens internes : 20
- Nombre maximum de liens externes : 20
- Nombre maximum de mots par cluster : 10

Ces choix ont été effectués en tenant compte des paramètres par défaut ainsi que de notre expérience d'analyse utilisant la méthode des mots associés sur des documents scientifiques mais également des précédentes expériences sur ce sujet (Ding et al., 2000 ; Grivel et al., 1995)

Après une première phase d'extraction terminologique, les descripteurs obtenus ont été standardisés manuellement afin d'éliminer les variantes les plus remarquables. Cette opération a été supervisée par un expert des technologies DSL de France Telecom. On remarquera ici l'importance de cette étape dans le processus d'analyse. Cette expérience a confirmé que l'index final n'était obtenu qu'après plusieurs itérations et dans un délai de plusieurs jours. Cette remarque doit nous conduire à une réflexion sur l'utilisation de ce type d'outil en fonction du problème considéré et des moyens humains à investir pour obtenir un résultat significatif.

4 - ANALYSE

4.1 - examen de l'index

L'index produit par Sampler contient 1394 termes d'occurrence variant entre 3 et 1590. Les termes ayant une occurrence de 1 ou 2 ont été volontairement occultés.

Un examen rapide de l'index permet de situer le thème traité par les documents. Les dix premiers mots sont *Dsl, adsl, broadband, customers, Internet, business, data, users, dsl services, alcatel*. Un peu plus loin, d'autres termes tels que *communication, adsl service, high speed, Internet service, high speed Internet, bandwith*…permettent de situer le contenu des données. D'autre part, un examen rapide permet également d'identifier les principaux acteurs du domaine : *Alcatel* (10^{eme}), *sbc communications* (15^{eme}), *verizon communications* (25^{eme}, *bellsouth* (34^{eme}), *lucent technologies* (36^{eme}), *chunghwa telecom* (46^{eme}), *France telecom* (49^{eme}), *deutsche telekom* (59^{eme}) …

On peut ici remarquer que le nombre d'occurrence (nombre d'apparition du terme dans le corpus) des noms de sociétés est relativement élevé alors que leur fréquence (nombre de documents dans lesquels ces termes apparaissent) est relativement faible. Ce phénomène est facilement explicable par la nature des données à traiter. En effet, les dépêches d'agences qui ont comme sujet le *DSL* sont souvent en relation avec une ou plusieurs sociétés. De plus chaque société est citée plus fréquemment dans chaque dépêche (en moyenne 3 fois par dépêche contre 2 pour des termes plus génériques).

La principale conséquence est la surreprésentation des sociétés dans les clusters produits par l'analyse.

4.2 - Analyse des clusters

Le logiciel Sampler produit 72 clusters à partir de l'index et des paramètres initialement fixés. On peut distinguer deux grandes catégories de clusters :

Les clusters constitués de termes génériques

Les termes génériques sont par définition centraux mais peu denses (par exemple *dsl*). Les clusters constitués de termes génériques permettent de distinguer les grands sous-thèmes du domaine analysé (*dsl equipment*, *telecom*, *dslam*…). Ces clusters apportent peu de valeur ajoutée à l'expert mais sont utiles pour les personnes désirant se familiariser avec le domaine étudié.

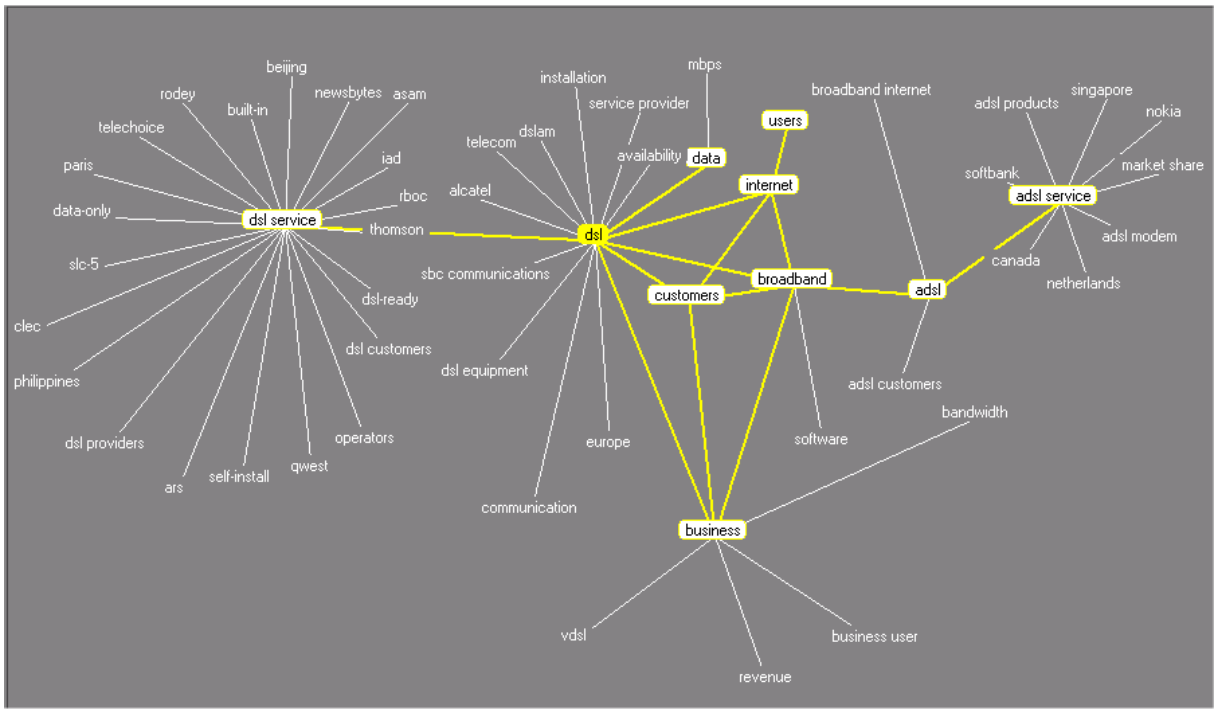

Fig. 1 : Cluster DSL

Dans le deuxième exemple (figure 2), le cluster Alcatel permet d'identifier les cinq acteurs majeurs du marché des équipements DSL. Ces relations ne doivent pas être interprétées comme des accords de partenariats ou tout autre lien qui pourrait exister entre ces sociétés, mais reflètent l'association effectuée dans les communiqués à titre de comparaison ou de référence dans ce secteur.

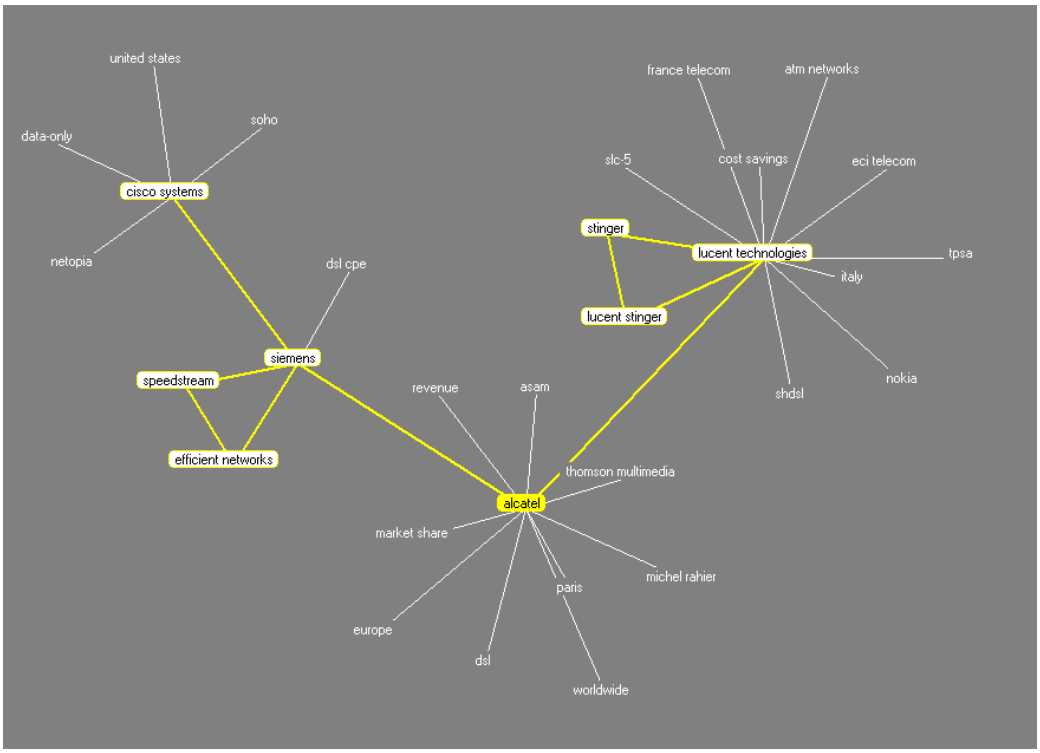

Fig 2. Cluster Alcatel

Détection de signaux faibles

Il existe, en très petite quantité, des clusters révélant des signaux faibles. Ces clusters ont été identifiés par des experts. Par exemple, le cluster Nokia présente son implantation géographique en Chine au cours de la période considérée.

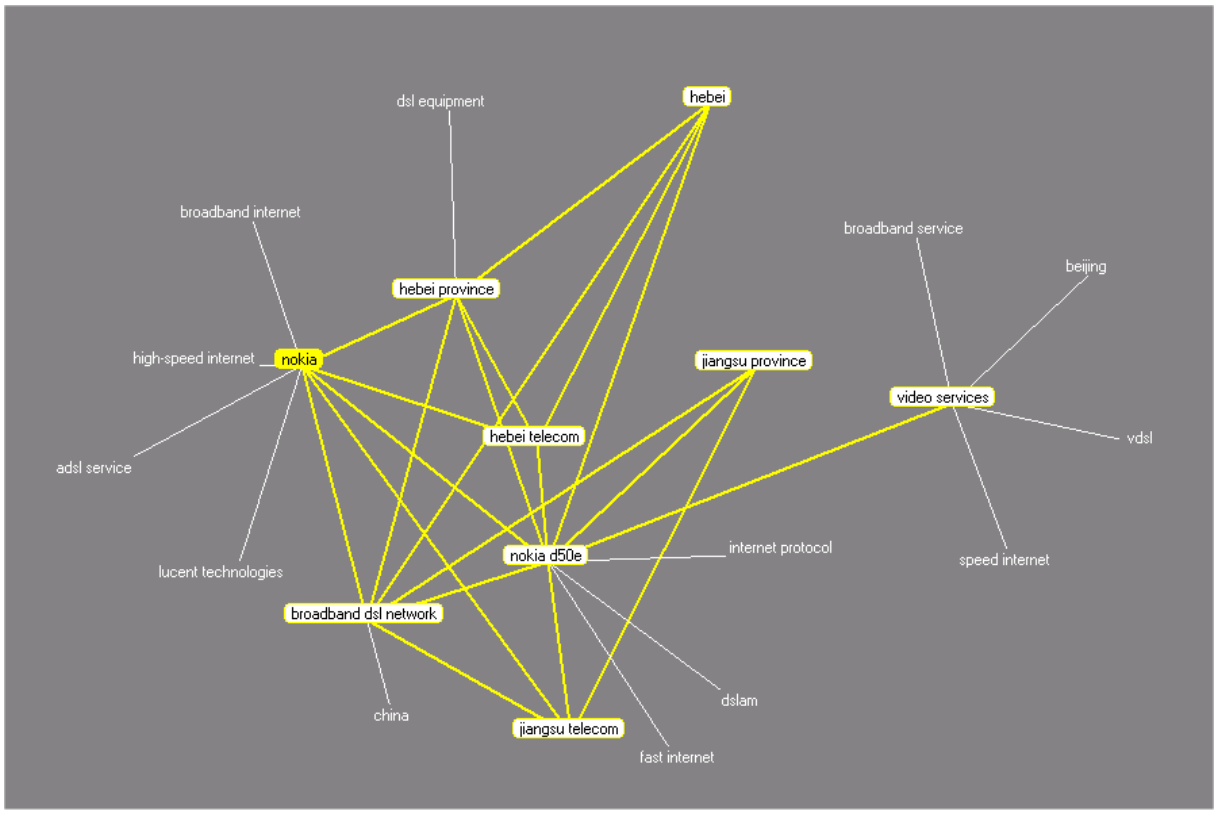

Fig 3. Cluster Nokia

Si Nokia est reliée à des termes relatifs à son activité (*broadband Internet, high-speed Internet, adsl service…*), cette société est également connectée à deux provinces chinoises et aux opérateurs locaux correspondants : *hebei province* and *hebei telecom*, *jiangsu province* and *jiangsu telecom*.
Cette activité récente de Nokia en Chine a été confirmée par une étude de marché *Idate.* Le cluster permet de mettre en évidence l'activité du constructeur durant cette période

Dans le deuxième exemple, on observe que le terme "*Federal Communication Commission"* est reliée aux termes *Telecom Act* et *Cable Providers* . Cette remarque est intéressante pour l'expert qui, en se référant aux dépêches d'origine, a pu confirmer un changement dans la réglementation concernant les câblo-opérateurs susceptible d'avoir un impact fort sur l'activité des opérateurs DSL.

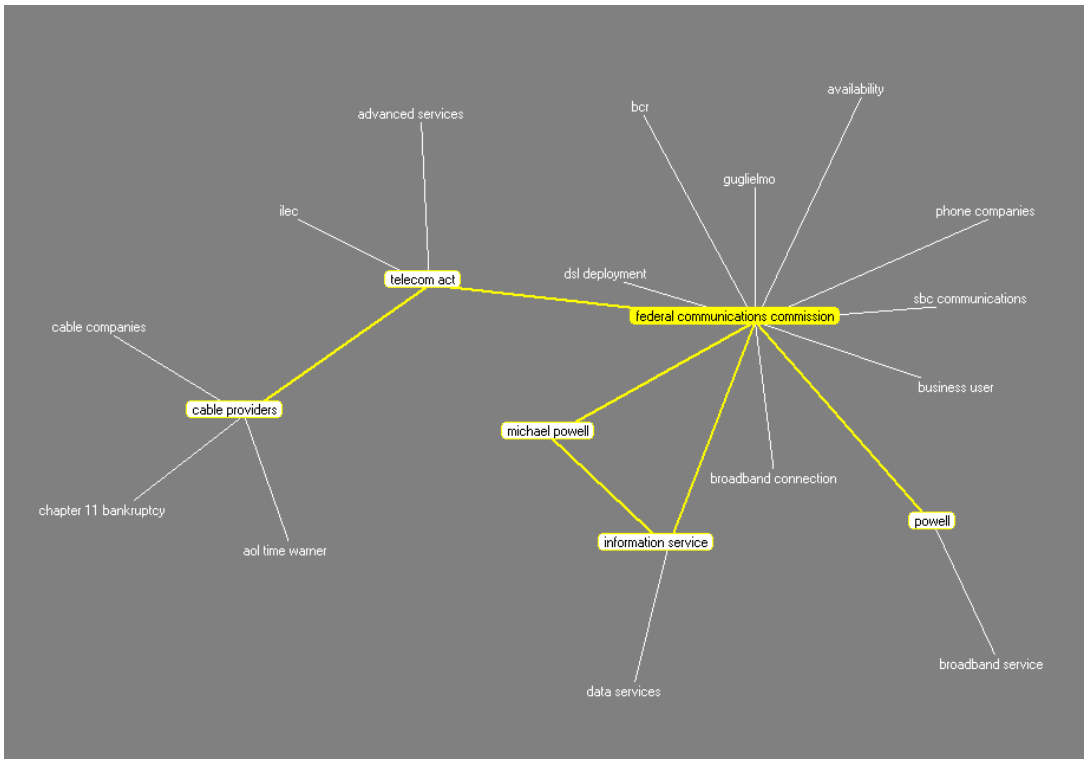

Fig 4. Cluster «Federal communication commission »

Synthèse sur l'analyse des clusters

Il existe deux types de cluster, chaque type ayant un intérêt pour une catégorie différente d'utilisateurs :
Les clusters qui permettent une bonne vue d'ensemble du domaine. Ces clusters permettent de découvrir les principaux acteurs, les technologies en jeux ainsi que les zones géographique où l'activité commerciale est la plus forte. Au sein de ces clusters, la cooccurrence est variable et l'indice d'équivalence souvent faible.
Les clusters recelant un signal faible. Ces clusters reflètent de faibles cooccurrences mais un indice d'équivalence fort. Ils contiennent une information à forte valeur ajoutée pour l'expert et mettent en scène les acteurs du domaine.

CONCLUSION

Cette expérience a été réalisée dans le but de valider ou d'invalider l'utilisation de la méthode des mots associés dans un contexte économique et financier. Elle nous a également permis de vérifier la pertinence de la méthode sur des dépêches de type Reuters.

Nous devons tout d'abord souligner que l'analyse est fortement dépendante de l'extraction terminologique, une extraction différente menant à un autre index et à la constitution de clusters différents. Nous avons essayé l'extracteur du logiciel *Leximine* sur le même corpus avec un résultat sensiblement différent (une meilleure extraction des multitermes mais des possibilités de modification manuelle de l'index obtenu moindre)

Peu de clusters sont directement interprétables (une quinzaine sur 72) De plus, les liens internes n'ont souvent aucun sens. L'interprétation, lorsqu'elle est possible, conduit à deux types de raisonnement :

- Le cluster est une interconnexion de termes du même ordre (pays, acteurs, technologies). Ces clusters sont utiles pour un profane souhaitant découvrir le domaine.
- Le cluster contient un signal faible ou un épiphénomène : Il est alors d'une grande aide pour l'expert.

La structure du corpus ne permet pas d'envisager l'évolution du domaine considéré. Les signaux faibles ne traduisent pas l'émergence de tendances fortes, mais traduisent une actualité ponctuelle.

En conclusion, la méthode des mots associés expérimentée au travers du logiciel Sampler fait sens, mais de façon différente. Si elle permet la détection de signaux faibles, elle n'est globalement pas aussi pertinente. De plus le temps de mise en œuvre est relativement important par rapport aux résultats obtenus.

D'autre travaux sur ce sujet sont en cours. Ils permettront une analyse plus fine des résultats obtenus et permettront à terme une généralisation des conclusions. Une étude « en dynamique » est actuellement en cours pour observer l'évolution des résultats présentés ici.

BIBLIOGRAPHIE

Eppstein R., (2001), Création d'un système d'information stratégique dans le domaine des technologies de l'information et de la communication, PhD Thesis – Université de Marne-la-Vallée.

Eppstein R., Datchary F. (1999), Complementarities between statistical and semantical analysis. in *17th Codata Conf Proceedings*.

Small, H. (1973) Co-citation in the scientific literature: a new measure of the relationship between two documents, JASIS, vol 24 p 265-269.

Small, H., Griffith, B.C., (1974) The structure of the scientific literature: identifying and graphing specialities, Science Studies, vol 4, p17-40

Turenne N., Rousselot F., (1998) Evaluation of four clustering methods used in text mining, *ECML'98 text mining workshop*

Y. Ding, G. Chowdhury, S. Foo, (2000) Bibliography of information retrieval research using co-word analysis" in *Information Process Management* 517, p1-26, Elsevier Ed.

Courtial J.P., Callon M., Laville F. (1991). Co-words analysis as a tool for describing the networks of interaction between basic and technological researches : the case of polymer chemistry, *Scientometrics*, Vol 22, Nº 1,

Callon M., Courtial J.P., Turner W., Bauin S., (1983) From translation to problematic networks : an introduction to co-word analysis, *Social Science Information* n°22, 1983.

Callon M., Courtial J.P., Turner W., (1991) La méthode Leximappe, un outil pour l'analyse stratégique du développement scientifique et technique, *Gestion de la recherche : nouveaux problèmes nouveaux outils, ed. de Boeck.*

Grivel L.,François C., (1995) Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique, *Solaris n°2, Presses Universitaires de Rennes*.

Peyrichoux, I., (2000) Sampler, un logiciel d'analyse textuelle : de la conception aux usages, *Internal Report CNAM/INTD*.

Polanco X., (1993) Analyse stratégique de l'Information Scientifique et Technique : Construction de clusters de mots-clés, *Sciences de la société,* n° 29.

Jouve O., (1999) Les outils d'analyse et de filtrage d'information : l'exemple du projet Sampler, *IDT'99* - Paris.

Jouve O., (1996) Sampler Manual, Cisi - Paris.

Turenne N., Rousselot F. (1998) Application of clustering in a system of query reformulation, *KAW'98.*

Law, J., Whittaker, J., (1992) Mapping acidification research: a test of the coword method, Scientometrics, Vol 23, p 417-461.

Courtial, J.P (1994) A coword analysis of scientometrics, Scientometrics, VOL 31, p251-260.