CO-WORD ANALYSIS FOR THE NON-SCIENTIFIC INFORMATION EXAMPLE OF REUTERS BUSINESS BRIEFINGS

B Delecroix[1*] and R Eppstein[1]

[1*] CESD/ISIS – Université de Marne-La-Vallée –
Email: bertrand.delecroix@wanadoo.fr

ABSTRACT

Co-word analysis is based on a sociological theory developed by the CSI and the SERPIA (Callon, Courtial, Turner, 1991) in the mid eighties. This method, originally dedicated to scientific fields, measures the association strength between terms in documents to reveal and visualise the evolution of scientific fields through the construction of clusters and strategic diagram. This method has since been successfully applied to investigate the structure of many scientific areas. Nowadays it occurs in many software systems which are used by companies to improve their business, and define their strategy but its relevance to this kind of application has not been proved yet. Using 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 overview of the software we used (Sampler) and after an outline of the experimental protocol, we investigate and explain each step of the co-word analysis process: terminological extraction, computation of clusters and the strategic diagram. In particular, we explain the meaning of each parameter of the method: the choice of variables and similarity measures is discussed. Finally we try to give a 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.

Keywords: clustering, co-word analysis, competitive intelligence

1 INTRODUCTION

Companies use competitive intelligence in order to improve their business through a better knowledge of their environment. Many software packages analyse business information, particularly those using co-words Analysis.

There is an extensive literature on co-word analysis (Courtial, Callon & Laville, 1991; Courtial, 1994; Law & Whittaker, 1992). This method reveals patterns and trends in scientific discourse by measuring the association strength of terms representative of relevant publications and patterns. Representative term associations are maps of the conceptual structure for a given scientific field.

However, the primary goal of this method was to analyse scientific documents, and not economic or financial information, particularly business news. We used 700 items of Reuters Business News dealing with information on the DSL's market (Digital Subscriber Line is a high speed internet connection technology). What does "analyzing the evolution of a scientific field" mean for these documents? In this study, we try to evaluate the use of the co-word analysis method by interpreting its results within an economics context.

1.1 Co-Word Analysis

Co-word analysis is related to co-citation analysis (Small, 1973; Small & Griffith, 1974). Co-citation analysis provides a method of mapping the structure of a research field through pairs of documents jointly cited. Co-word analysis deals directly with sets of terms shared by documents instead of shared citations. Therefore, it maps the pertinent literature directly from the interaction of key terms instead of the interaction of citations. While this paper will concentrate on co-word analysis, it would be interesting to investigate how co-citation analysis combined with co-word analysis can be used to represent the actors or knowledge networks that determine a discipline.

The method reduces a large space of related terms (words and phrases) to multiple smaller related spaces that are easier to understand but are also indicative of the actual partitions of interrelated concepts in the literature under consideration. This analysis requires an association measure and an algorithm for searching through a term's space. The analysis is designed to explain how main areas are interrelated.

Metrics for co-word analysis have been studied extensively (Grivel & François, 1995). Two terms, *i* and *j*, co-occur if they are used together in a single document. Take a corpus consisting of *N* documents. Each document is indexed by a set of unique terms that can occur in multiple documents. Let C_k be the number of occurrences of term *k*; i.e., the number of times *k* is used for indexing documents in the corpus. Let C_{ij} be the number of co-occurrences of terms *i* and *j* (the number of documents indexed by both terms).

Different measures of association have been proposed. The basic metric used for this study is the *Association Indice* E_{ij}. The strength of association between terms *i* and *j* is given by the expression:

$$E_{ij}=\frac{C_{ij}^2}{C_i * C_j} \text{ , with } 0\le E_{ij}\le 1 \qquad \text{Eq(1)}$$

This metric provides an intuitive measure of the strength of association between terms, and only indicates that there is some semantic relationship. This metric is easier to understand and utilize in the production and interpretation of term association maps than the so-called inclusion metric. It allows associations of both major and minor terms and is symmetrical in their relationships (Callon, Courtial, & Turner, 1991). *E* can be used as the basis for devising several complementary measures of term interactions and term networks in a unified manner.

Two terms that appear many times in isolation but only a few times together will yield a lower *E* value than two terms that appear relatively less often alone but have a higher ratio of co-occurrences. Terms with relatively high *E* values form the networks' links. A term network consists of nodes (terms) connected by links. Each node must be linked to at least one other node in a network.

The co-word algorithm then proceeds in two steps to produce the paired connection of terms. The first step builds networks that can identify areas of strong focus. The second step can identify terms that associate in more than one network and thereby indicate overlapping issues.

The first step generates the primary associations among terms; these terms are called internal nodes and the corresponding links are called internal links. The second step generates links between first-step nodes across networks, thereby forming associations among complete clusters. Second-step nodes and links are called external links.

Without some minimum constraints, terms that appear infrequently but almost always together could dominate clusters; hence a minimum co-occurrence C_{ij} value is required to generate a link. At the same time, some maps can become cluttered due to an excessive number of legitimate links (generally of decreasing *E* values); hence, restrictions on numbers of nodes and sometimes links are required to help discover the major partitions of concepts. However, in many cases only the number of qualifying nodes limits many term networks.

1.2 Sampler Software

Sampler is a lexico-statistical analysis software developed by Cisi (Jouve, 1996) which is now a subsidiary of CS communication & Systemes. It is based on the research of the *Service d'Etude et de Réalisation de Produits d'Information Avancés (SERPIA)* and *Centre de Sociologie de l'Innovation (CSI) at "Ecole des Mines"*(Callon, Courtial, Turner & Bauin, 1983) on the co-words analysis theory. In Sampler software, this approach is combined with a morpho-syntaxic analyzer that works on uniterms but also on expressions or multiterms, but avoids polysemy (Peyrichoux, 2000).

Associated terms are aggregated to produce association networks. Each network enables graphical navigation into documents. Graphics represent lexical networks also called "clusters". These networks do not correspond to established semantic structures but to contextual associations.

This bottom-up model reveals relations between terms to the user without any preliminary knowledge of the topic and allows the discovery of new expressions.

Sampler’s clusters are not oriented but contain two types of links. Internal links represents associations where the occurrence of each term is strongly linked to co-occurrence. External links represent relations between terms that appear in different contexts. Cluster construction is done with an Ascendant Hierarchical Clustering algorithm. This simple method builds clusters but doesn't permit the representation and the visualization of their relative positions, which are allowed with methods such as the multidimensional scaling algorithm or self-organizing maps.

1.3 Experimental protocol

A total of 800 news items concerning "Digital Subscriber Lines" over a period of six month (from October 2001 to March 2002) were extracted from Reuters Business Briefing database. The exact extraction equation we used was *"dsl OR adsl OR xdsl OR digital subscriber lines"*.

To improve the reliability of our analysis, we completely cleaned this set of documents: First we removed repeated news (Reuters Business Briefing gathers news from various press agencies) to obtain 700 unique documents. Later on these documents were cleaned up to eliminate interfering words and tags such as name of the author, town of origin, name of the press agency, etc.

We fixed the Sampler parameters as follows:

- minimum number of co-occurrence: 3
- minimum number of occurrence: 3
- maximum number of inner links: 20
- maximum number of outer links: 20
- maximum number of word per cluster: 10

These choices were made taking into consideration both default parameters, our experience in analyzing documents with co-word analysis and related work on scientific material (Ding, Chowdhury & Foo, 2000; Grivel & François, 1995)

After the first terminological extraction using Sampler's extractor the resulting descriptors were standardized to eliminate spelling differences and variations of the same terms. This operation was done under the guidance of DSL's experts. One must note here the importance of this step in the whole process. This experiment confirmed that a trusted index was obtained after several steps. The total time spent cleaning data was nearly 3 days!

2 EXPERIMENT AND RESULTS

2.1 Index examination

Sampler computes an index by extracting terms from all documents. This index contained 1394 terms, whose occurrence ranged from 3 to 1590. Terms with an occurrence of 1 or 2 are not presented. Terms with very high frequency were eliminated using an empty word stoplist.

A quick examination of the index allowed the treated domain to be repaired. The ten first terms were:
Dsl, adsl, broadband, customers, Internet, business, data, users, dsl services, alcatel.

If we look at further terms, we can find: *communication, adsl service, high speed, Internet service, high speed Internet, bandwith…*

Furthermore, it allowed us to identify the major actors of the sector: *Alcatel* (10th term), *sbc communications* (15th), *verizon communications* (25th, *bellsouth* (34th), *lucent technologies* (36th), *chunghwa telecom* (46th), *France telecom* (49th), *deutsche telekom* (59th) …

One will notice that if the occurrence (number of appearances of a pattern in the corpus) of names of companies is relatively high, the frequency of these terms (number of documents in which a pattern appears) is much lower: on average, a company is cited three times in a document. The ratio is not so high with more generic terms (1 or 2 on average).

Actually, Reuters News item are often inspired by *news releases*, and financial news will focus on *dsl* if a company is involved. One consequence is that company names will become over-represented, and more weighted in the index and clusters than other terms.

2.2 Cluster analysis

Sampler computed 72 clusters using the parameters we fixed. One can distinguish two different kinds of clusters:

Generic term clusters

Generic term clusters are, by definition central but not dense (example: adsl), and terms forming these clusters belong to subtopics (*dsl equipment*, *telecom*, *dslam…*). These clusters provide little information for experts of the domain, but are useful for those who want to familiarise themselves with the topic being studied.

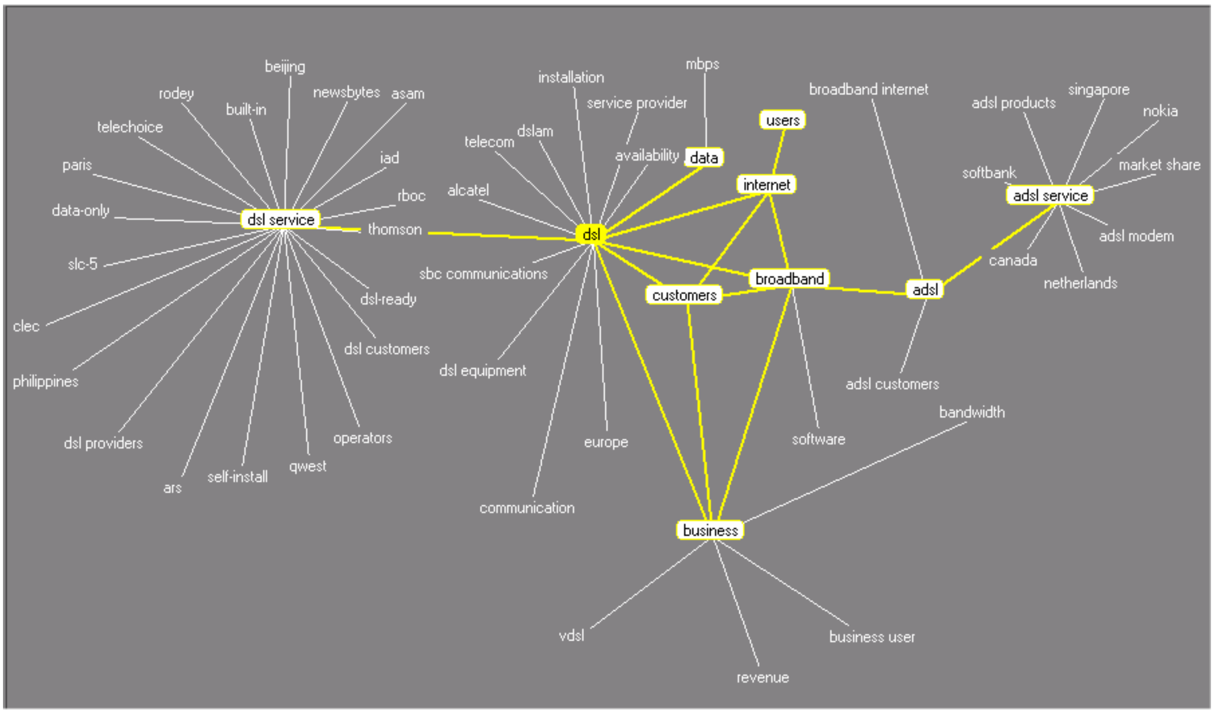

Figure 1. Example of *adsl* cluster

Another example is the Alcatel cluster which shows the five major actors in the DSL market. This does not reveal any agreement or particular link between these companies. The associations only reflect citations as comparisons or as references in the sector, like *"Alcatel is the first company behind siemens"*

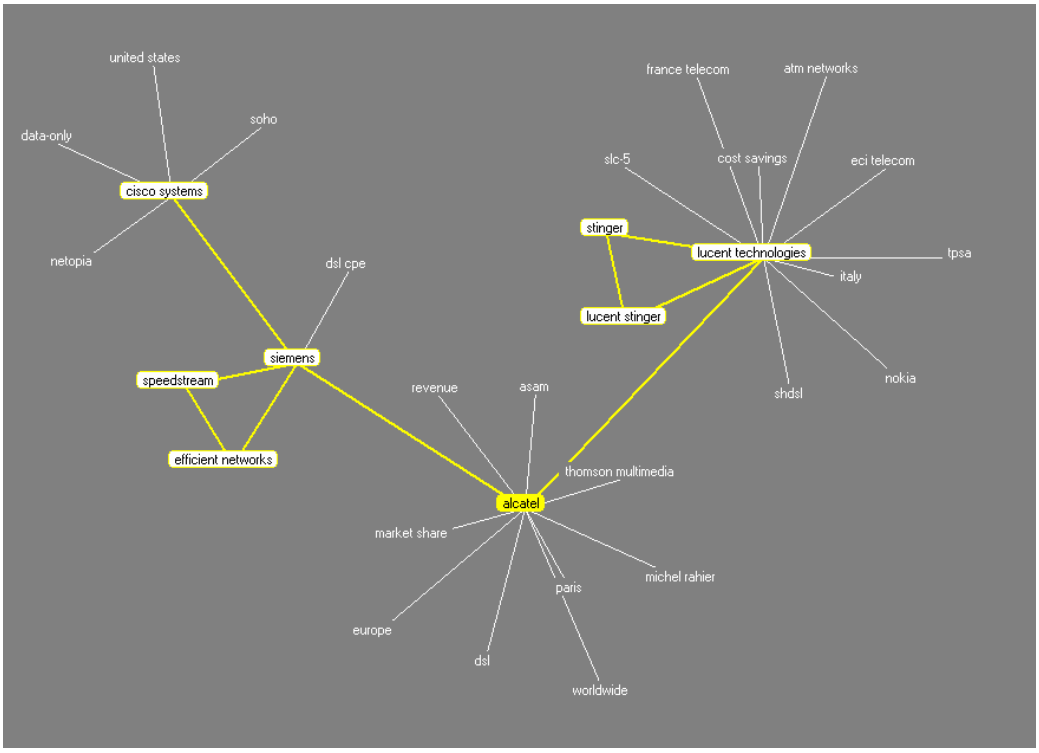

Figure 2. Example of *alcatel* cluster

Weak signal detection clusters

A small number of clusters reveal a weak signal, and were determined with the help of an expert. For example, the *Nokia* cluster *illustrates* its geographical implementation in China.

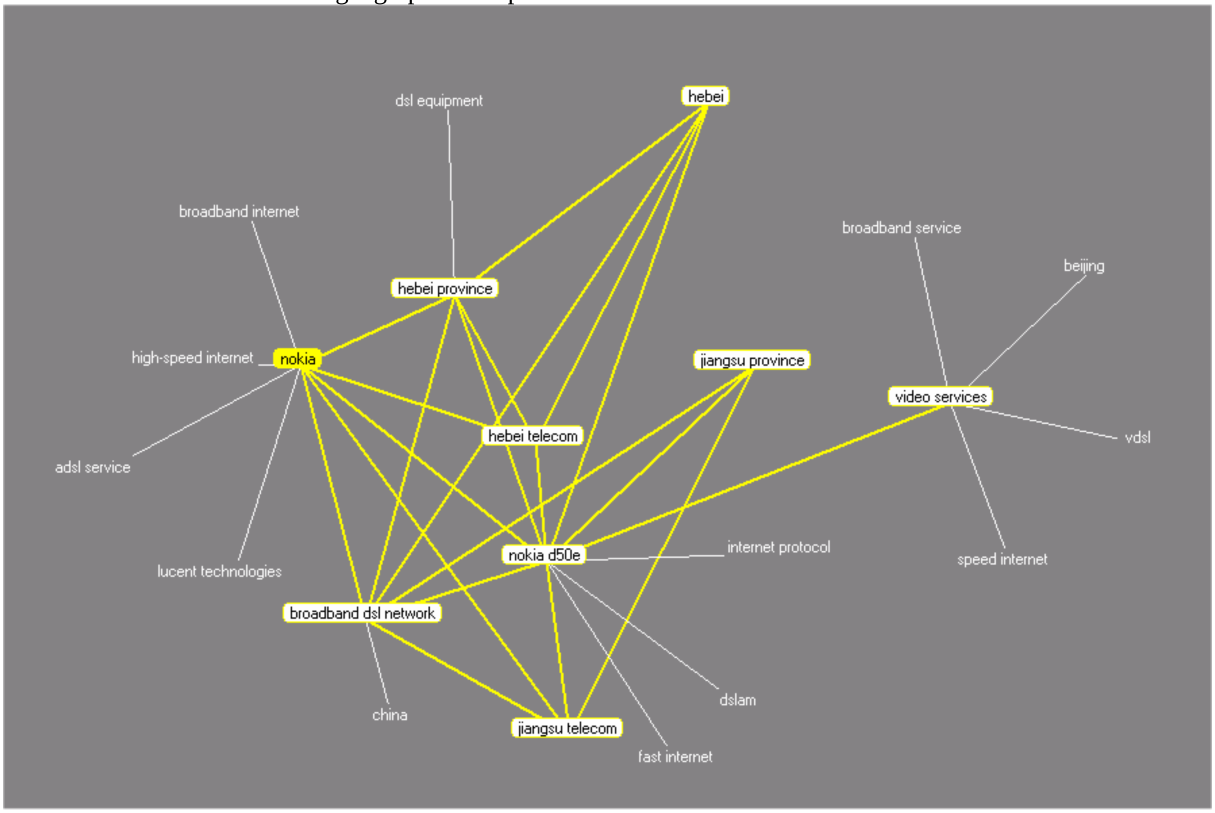

Figure 3. Example of the *Nokia* cluster

As well as *Nokia* being linked with terms related to its activity (*broadband Internet*, *high-speed Internet, adsl service…*), it's also linked with Chinese province names and local operators: *hebei province* and *hebei telecom*, *jiangsu province* and *jiangsu telecom*.

Actually Nokia's activity in China was confirmed by an *Idate Market Study*. The cluster reveals that Nokia was very active in this geographical sector in the period studied.

Another example is the *Federal Communication Commission* cluster, which is linked with the *Telecom Act* and *Cable Providers*. This observation is useful for the expert who has deduced, by referring to the original documents that the regulation concerning cable providers was under discussion, which could have had a great impact on the activity of dsl providers.

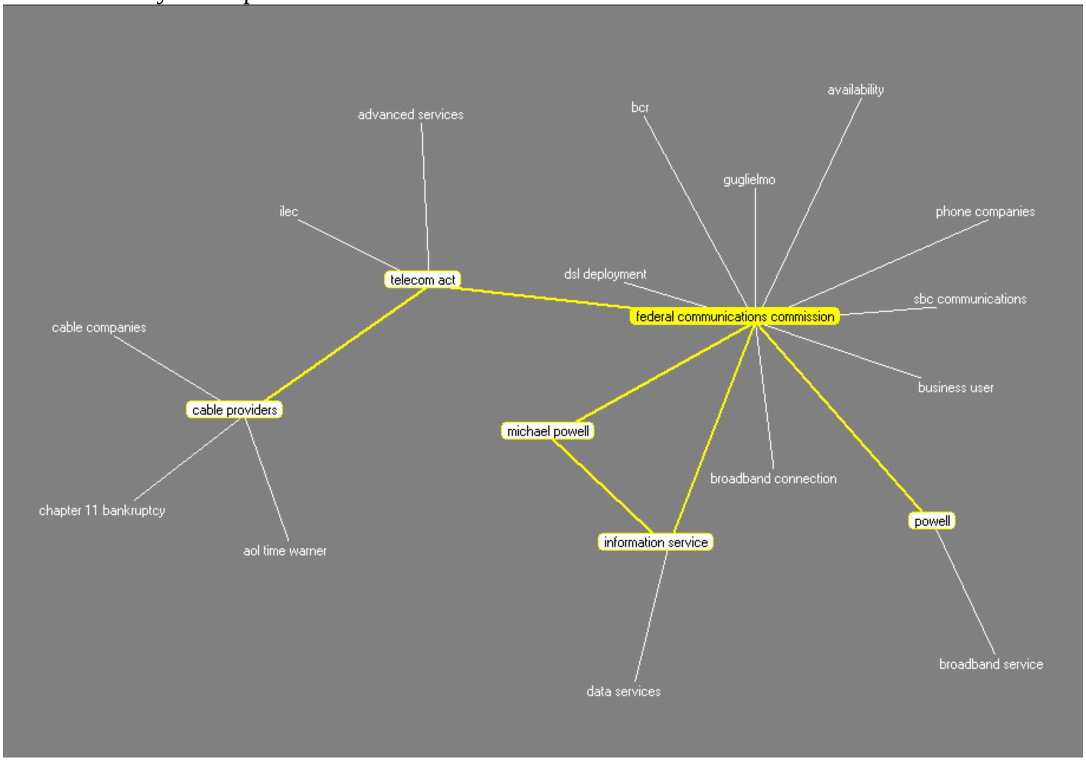

Figure 4. Example of *Federal Communication Commission* cluster

Conclusion on the clusters

Two types of clusters exist, which are useful for two types of users:

1- Clusters allowing a wide view of the domain. The clusters enable the discovery of the principal actors, geographical areas and technologies. In these clusters, the association relation can be much lower than in the other ones.

2- Weak signals: these clusters are computed with terms of low co-occurrence, but a large association relation. They contain finite and timely information. They are useful for the expert, allowing actors to be linked with events.

2.3 Strategic diagram

In scientific studies, clusters allow the discovery of the different thematic components of a scientific field.
The strategic diagram shows the emergence of trends in the studied domain. Every cluster is represented on a diagram according two criterions: centrality and density.

- Centrality is measured with the mean of the *Association Indice E* of all external links. The higher the value of this mean, the more closely the cluster represents a *reference* topic (central) in the corpus.

- Density, or internal cohesion, is measured with the mean of the *Association Indice E* of all internal links of the cluster. The higher it is, the more coherent is the cluster and the more likely it is to contain inseparable expressions.

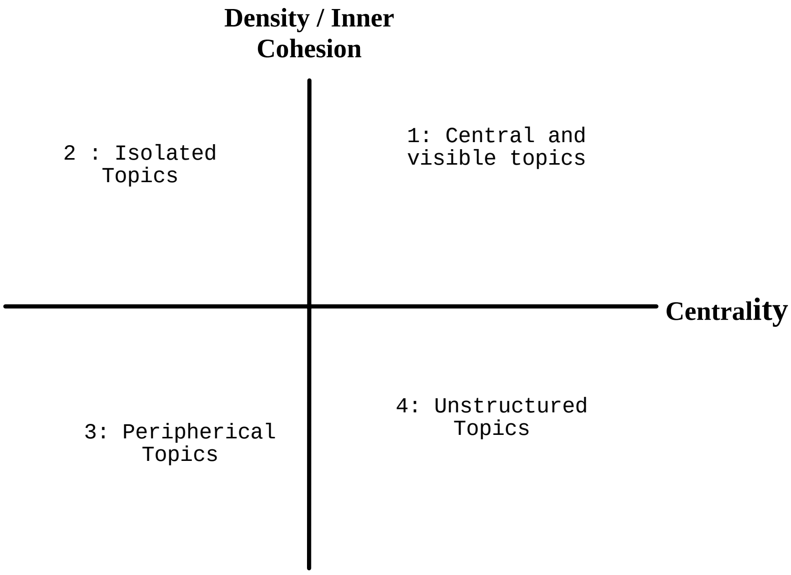

Figure 5. Strategic Diagram

In the first quadrant, we must locate the most central and visible clusters, i.e. the *heart of the network*. In quadrant 2, we find the densest and most isolated topics. Quadrant 3 contains clusters weakly linked with the rest of the network. This could mean that these topics are becoming marginalised, or on the contrary are interesting domains for the future, for which a structured vocabulary is not well defined yet. Quadrant 4 contains peripheral domains and a few structured domains, which can sometimes illustrate a future view of the domain.

It is hard to generalise from a single experiment, but one could assert that in the quadrants with high density, weak signals are found; while in the quadrants with high centrality, the vocabulary of the domain is found.

3 CONCLUSION

This experiment was made in order to evaluate the use of co-words analysis in an economic and financial domain, even though this method was originally used for studying the evolution of scientific fields. However, now software packages using this method are sold to companies claiming economic benefits. We tried to evaluate the use of the method with Reuters news briefings.

Firstly, the analysis is very dependent on terminological extraction. Actually, using another terminological extractor would lead to another index, and then to other clusters. We tried to compute the same corpus with another tool, *Leximine* (Lexiquest, 2001) The index was very different to the one we obtained (we achieved better recognition of multiterms, but found it impossible to customize dictionaries).

After computing the clusters, a small number of them could be interpreted: the connected inner terms made no real sense. If we ignore these clusters, we can identify two types of clusters:

- Clusters that connect terms of the *same level* (countries, actors, technologies), without hierarchy. These clusters are very useful for those who wish to explore the domain of the corpus.
- Clusters that allow the detection of weak signals, i.e. of epiphenomenon.

The structure of the corpus didn't allow us to visualize or to predict the evolution of the domain in this type of analysis, even though it's one of the most valuable applications to scientific fields. The weak signals were not emergent events, they're only timely, and probably won't become *the rule*.

In conclusion, in the corpus we used, the co-words theory makes sense, but in a different way than for a scientific corpus. Even if it enables the detection of the weak signals, this is not so pertinent globally in that it doesn't help us identify, thanks to the strategic diagram, emerging trends. More importantly, it is very costly in time if we want to detect weak signals (because of corpus and index cleaning).

Further studies will be conducted in order to generalize the results of this study. A dynamic study on a larger period of time will be done, in order to see whether the evolution of events can be inferred. Furthermore, metrics other than the equivalence indice will be experimented with in the computation process.

4 REFERENCES

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

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*, 191-235.

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, 22(1)*, 155-205.

Courtial J.P (1994) A coword analysis of scientometrics, *Scientometrics, 31*, 251-260.

Ding, Y., Chowdhury, G. & Foo, S., (2000) Bibliography of information retrieval research using co-word analysis" in *Information Process Management*, *517,* 1-26.

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, 2*, Rennes: Presses Universitaires de Rennes.

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

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

Lexiquest (2001) Lexiquest product white paper – Lexiquest Mine. Retrieved June 14 2004, from the Lexiquest website: www.lexiquest.fr/products/ LexiQuest%20Mine%20White%20Paper.pdf

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

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

Small, H. & Griffith, B.C., (1974) The structure of the scientific literature: identifying and graphing specialities, *Science Studies,4,*17-40