Cuadernos de Inteligencia Competitiva, Vigilancia Estratégica, Científica y Tecnológica
QUIC&VECT 2 (2009), Edición 1, 23 - 29

Métodos de selección para la extracción de señales débiles

Cristèle ROUX

183, route de Luxembourg L-8077
e-mail: roux.cristelle@tiscali.com

Key words
Weak signals, signal detection, data mining, Tetralogie application, Text analysis, Data analysis

Palabras clave
Señales débiles, detección señales, minería de textos, aplicación Tetralogie, análisis textos, análisis datos

Abstract
We will present a method of extraction of weak signals based on the evolutionary and structural analysis of the semantic fields. This method employs the following tools:

- Matrix of crossovers involving semantic terms which evolve with time,
- Extraction of emergent terms (for posterior normalization, this are last column selections),
- Matrix of concurrence crossing the emergent terms among them,
- Selection of diagonal bars / blocks, on this matrix,
- Extraction of the blocks that represent emergent and coherent concepts.

Resumen
Presentaremos un método de extracción de las señales débiles basado en el análisis evolutivo y estructural de los campos semánticos. Este método hace intervenir las siguientes herramientas:

- Matriz de cruces de los términos semánticos con el tiempo,
- Extracción de los términos emergentes (por normalización posterior, son elecciones de la última columna),
- Matriz de concurrencia cruzando los términos emergentes entre ellos,
- Selección barras / bloques diagonales, de esta matriz,
- Extracción de los bloques que representan conceptos emergentes y coherentes.

1. Algoritmos de elección de grandes matrices

1.1. Elección por bloques sobre las relaciones absolutas

Esta técnica tiene numerosas aplicaciones:

- Investigación en clases conexas,
- Para cada una de las clases, una selección interna por bloques permite reagrupar directamente los elementos más unidos,
- Reorganización de una matriz conexa en bloques diagonales.

Su utilización en análisis de textos permite detectar las clases semánticas emergentes más marcadas, llevándonos hacia la matriz de cruces de los nuevos términos. Esta terminología emergente puede formar grupos correspondientes a conceptos emergentes. Con un solo término no hay suficiente, ya que se puede tratar de una evolución terminológica que consagra un concepto ya antiguo que, manteniéndose, se beneficia de un vocabulario específico (a menudo una palabra sencilla reemplaza así a una expresión o una palabra compuesta).

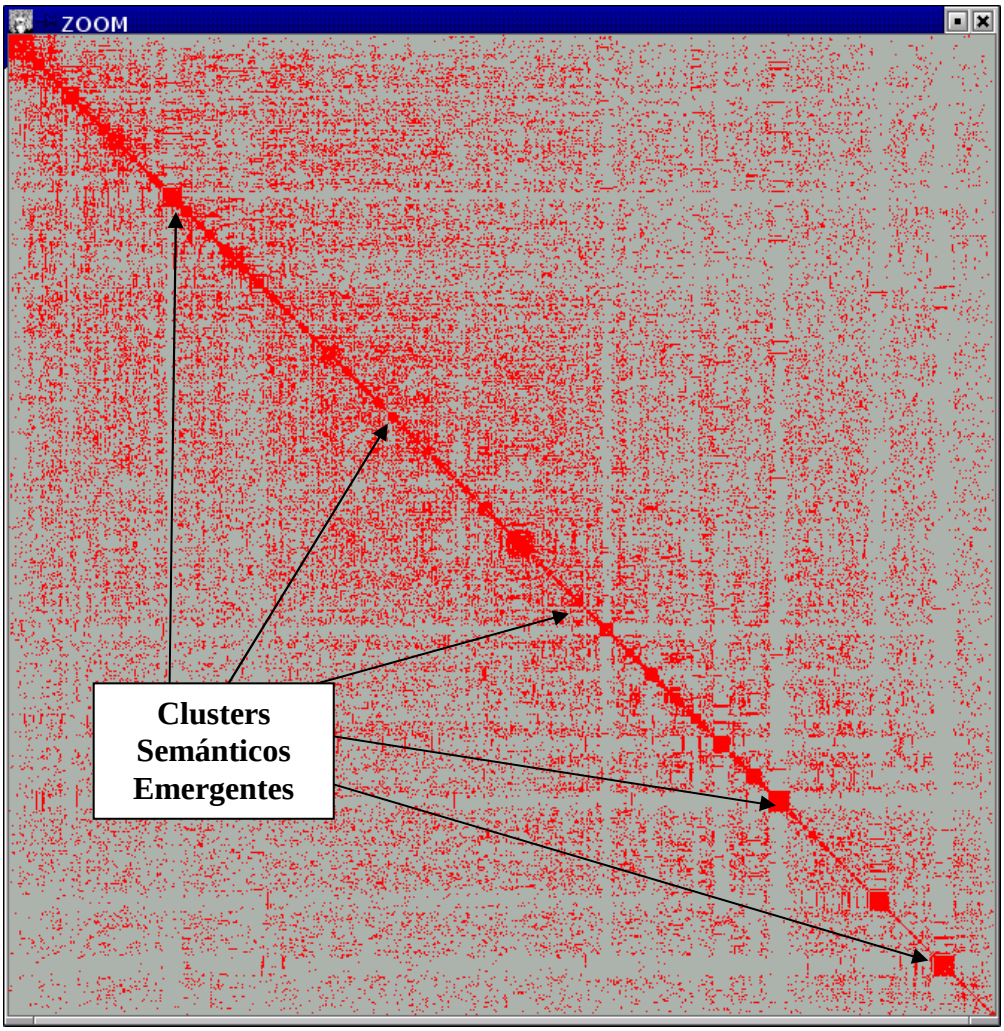

Figura 1. : Gráfico de bloques diagonales sobre una matriz de coocurrencia semántica.

La matriz de arriba reagrupa cerca de 25 millones de celdas (5000 x 5000) y ya hemos trabajado en matrices de hasta 10000 líneas y columnas.

1.2 Elección por bloques sobre las relaciones relativas

Esta técnica es utilizada cuando los términos cruzados tienen frecuencias muy diferentes. En efecto, en los textos se mezclan términos corrientes o muy utilizados en el ámbito con otros mucho más precisos que tienen especificidades. Si queremos descubrir los grupos semánticos que corresponden a estos asuntos emergentes o raros, tenemos que pasar previamente por la modalidad relativa antes de hacer la elección. Fijémonos que, para las matrices de coocurrencias simétricas, cruzando modalidades exclusivas (por ejemplo: autores o palabras clave), los elementos diagonales representan de hecho las frecuencias en el corpus. Tenemos que proceder igualmente para cruces asimétricos entre dos variables diferentes planteando los mismos problemas de dispersión de frecuencias. Proponemos diversos métodos para pasar a modalidad relativa:

- División de cada elemento de la matriz por la raíz cuadrada de los elementos diagonales que le corresponden. Obtenemos entonces una matriz diagonal unitaria (caso simétrico de manera única). Este principio funciona bien sobre las matrices semánticas y tiene en cuenta las relaciones débiles.

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

- División del cuadrado de cada elemento por los elementos diagonales, obtenemos entonces una matriz de equivalencia que también será diagonal unitaria (caso simétrico de manera única). Este método es muy utilizado para analizar las redes semánticas, pero tiene tendencia a penalizar las relaciones débiles, porque de hecho se trata del cuadrado de la similitud precedente y por lo tanto un valor de ½, que no representa en absoluto más que ¼.

- La similitud de Kulzinsky es de igual orden y equivalencia, pero la media de las frecuencias se reemplaza por uno de los factores del numerador. Se utiliza en la detección de las redes semánticas asociadas a las señales fuertes.

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

- Podemos difuminar el efecto reductor de las dos proposiciones precedentes utilizando el índice de proximidad, que es obtenido dividiendo cada término de la matriz por los elementos diagonales asociados (caso simétrico de manera única).

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

- Siempre en el marco de las matrices simétricas, podemos utilizar la similitud de inclusión que nos informa, si se acerca a 1, que un término siempre está enlazado con otro o que un autor pertenece exclusivamente a un equipo del cual el director firma todas las publicaciones. Esta métrica es muy útil para hacer la diferencia entre los elementos específicos de un grupo y los que interfieren con los otros grupos.

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

- División por la raíz cuadrada de los marginales: este procedimiento es aplicable a las matrices asimétricas. Como los marginales son siempre superiores a los elemento diagonales, este método tiene tendencia a penalizar los términos más frecuentes (palabras, herramientas, términos generales, términos de la ecuación de búsqueda), por eso, privilegia a los términos raros que son frecuentemente en coocurrencia. Es pues posible detectar ciertas señales débiles (grupos coherentes de términos poco extendidos).

$$S_{ij} = \frac{a_{ij}}{\sqrt{a_{i\bullet}\ a_{\bullet j}}} \qquad teniendo: \qquad a_{i\bullet} = \sum_j a^{ij} \qquad y \qquad a_{\bullet j} = \sum_i a^{ij}$$

- División por la norma de las líneas (o de las columnas). Este método de reducción permite uniformizar una de las dos variables, las modalidades son entonces de igual corte y el efecto frecuencial es difuminado.

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

- División por el máximo de la línea (o de la columna) como en el caso de la norma n°3 encima de. Se denota pues que para una matriz simétrica, la diagonal se hace unitaria ya que, de manera inicial, es dominante a nuestras matrices.

Hemos conservado dos técnicas de Tetralogie.

- La primera consiste a normalizar la matriz, después modificarla, y finalmente escogerla. Tiene la ventaja de la elección de la normalización, pero destruye los valores iniciales de la matriz.

- La segunda se basa en una normalización compatible con las matrices no simétricas, escoge la matriz en función de los nuevos valores, pero conserva los antiguos. Así pues sólo la estructura de la matriz cambia pero no los valores.

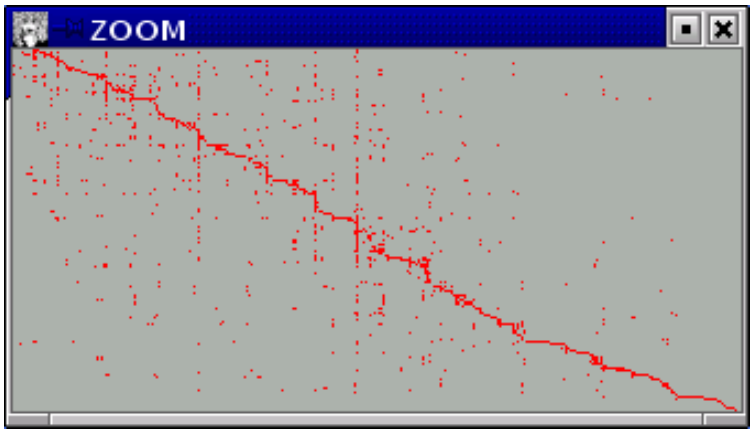

Figura 2. : Selección por bloques diagonales de una matriz asimétrica Autores - Diarios.

En el ejemplo de encima, detectamos los rastros de un ámbito de investigación a partir de una matriz Autores - Diarios, escogida por bloques en moda relativa.

1.3 Extracción automática de las clases

Considerando la grandeza de algunas de las matrices analizadas y el número importante de clases (clusters) puesto en evidencia, nos ha parecido oportuno buscar una técnica automática que permita aislar a cada una de ellas. Como aquí los elementos a agregar consiguen secuencialmente formar la diagonal o la pseudo-diagonal dominante de la matriz, es suficiente detectar los saltos de semblanza para aislar cada clase de la clase adyacente. Una bajada de esta medida traduce, en efecto, la ausencia en el resto de los elementos no clasificados de elementos susceptibles de venir a completar la clase en el curso de elaboración. Un umbral convenientemente escogido permite entonces realizar un recorte eficaz, sólo de las clases que tienen bastantes elementos, que serán pues analizadas.

2. Extracción de informaciones estratégicas

2.1. Extracción interactiva de información: las emergencias (conceptos emergentes)

Además de la visualización en 4D, una de nuestras aportaciones más apreciadas a nivel de métodos de análisis multidimensionales es la introducción de la variable tiempo a numerosos niveles de exploración. Así pues tenemos un método de extracción de conceptos emergentes utilizando las manipulaciones interactivas sobre una AFC realizada en función de la variable tiempo:

- Cruzar la variable a analizar con el tiempo expresado en periodos efectivos lo bastante homogéneos (relación entre 1 y 2),
- Hacer una AFC de la matriz obtenida,
- Visualizar el mapa de las modalidades temporales (columnas solas),
- Por rotaciones, manipular la nube hasta aislar el último componente temporal en un rincón de la ventana (en la figura 3 siguiente: 1997 en alto a la izquierda),
- Visualizar el mapa global (variable a analizar más el tiempo),
- Exportar, hacia este mapa, el acimut encontrado en la primera,
- Extraer los elementos que se encuentran más allá o cerca del icono, asociados al último periodo (en naranja sobre el mapa 4D),
- Generar el filtro que contiene todas las modalidades emergentes de la variable analizada.

Este filtro puede entonces ser reutilizado para cruzar las emergencias entre ellas y encontrar así los conceptos emergentes.

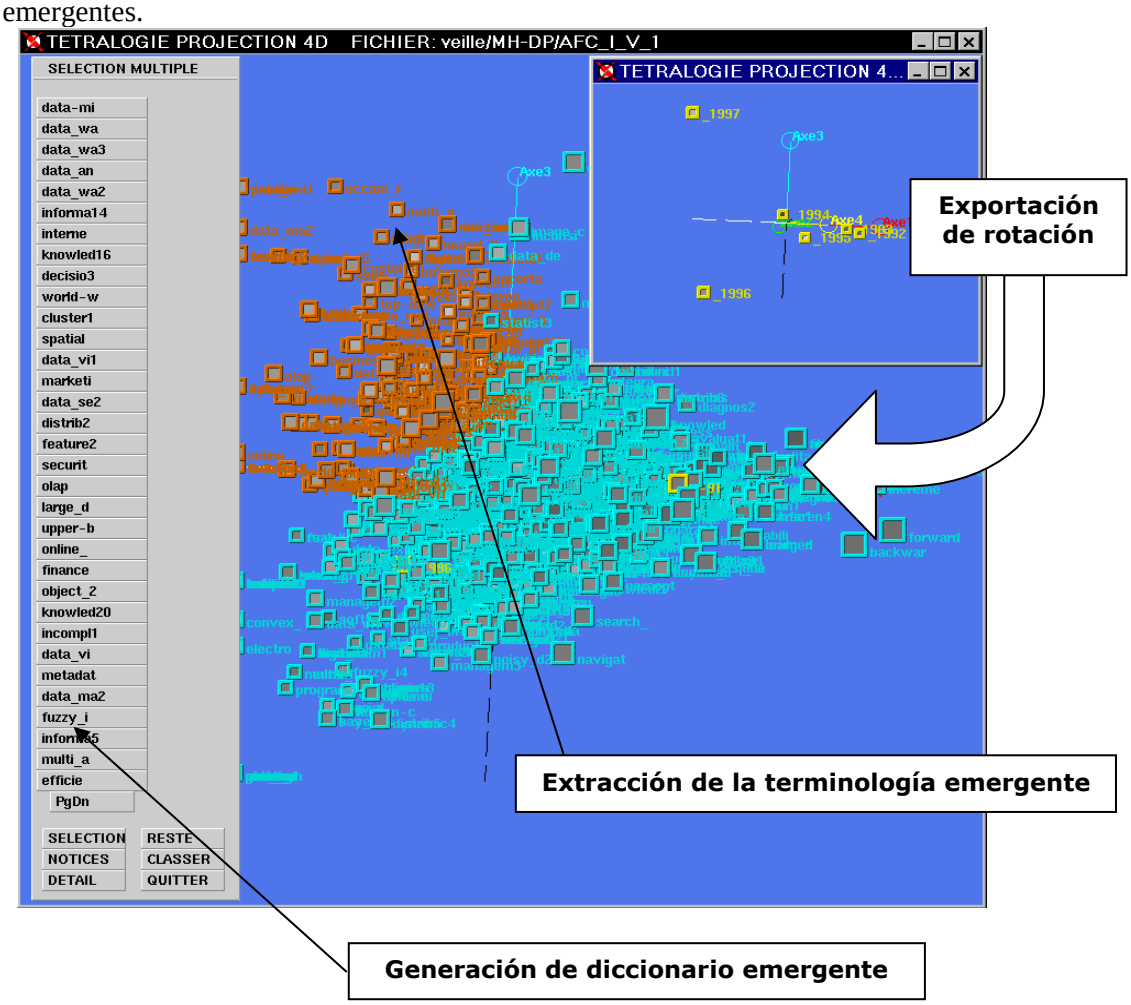

Figura 3. : Extracción de elementos emergentes basada en una AFC Temática – Tiempo.

Así pues, ampliaremos este tipo de paso hacia otras estrategias de descubrimiento de conocimientos esencialmente basadas en la interacción. Utilizando la herramienta de visualización para duplicar las facultades sensoriales del usuario, quien, por sus capacidades de deducción y su dominio del asunto, es el único que puede dirigir el análisis en cuestión.

2.2. Detección de las señales débiles

Este método, muy apreciado por los responsables, consiste al extraer de las clases semánticas emergentes, las novedades en un ámbito dado. Por eso, procedemos a:

- Partir de una matriz Palabras Clave - Fechas o mejor Multi-términos - Fechas,
- Extraer la terminología emergente,
- Cruzarla con ella misma (matriz cuadrada de coocurrencias),
- Escoger esta matriz por bloques diagonales,
- Extraer las clases más visibles,
- Generar el detalle (la lista de los términos conectados entre ellos).

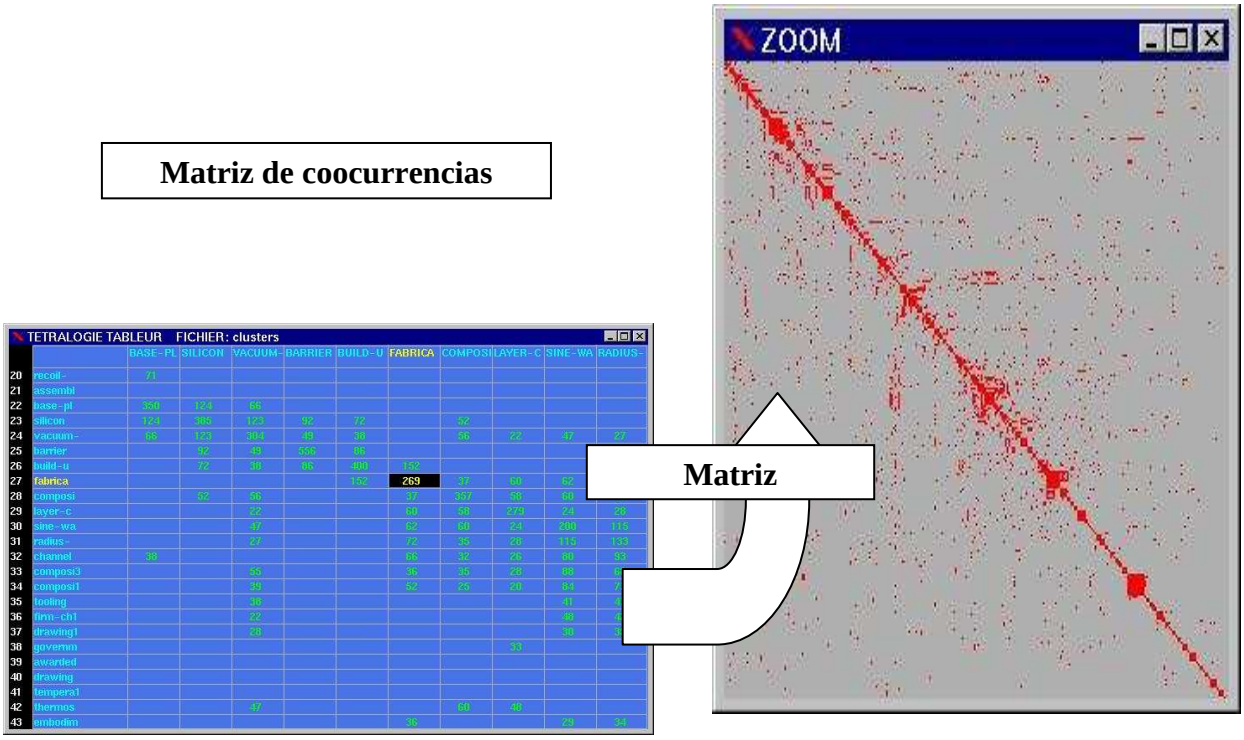

Figura 4.: Ilustración del método de extracción de señales débiles

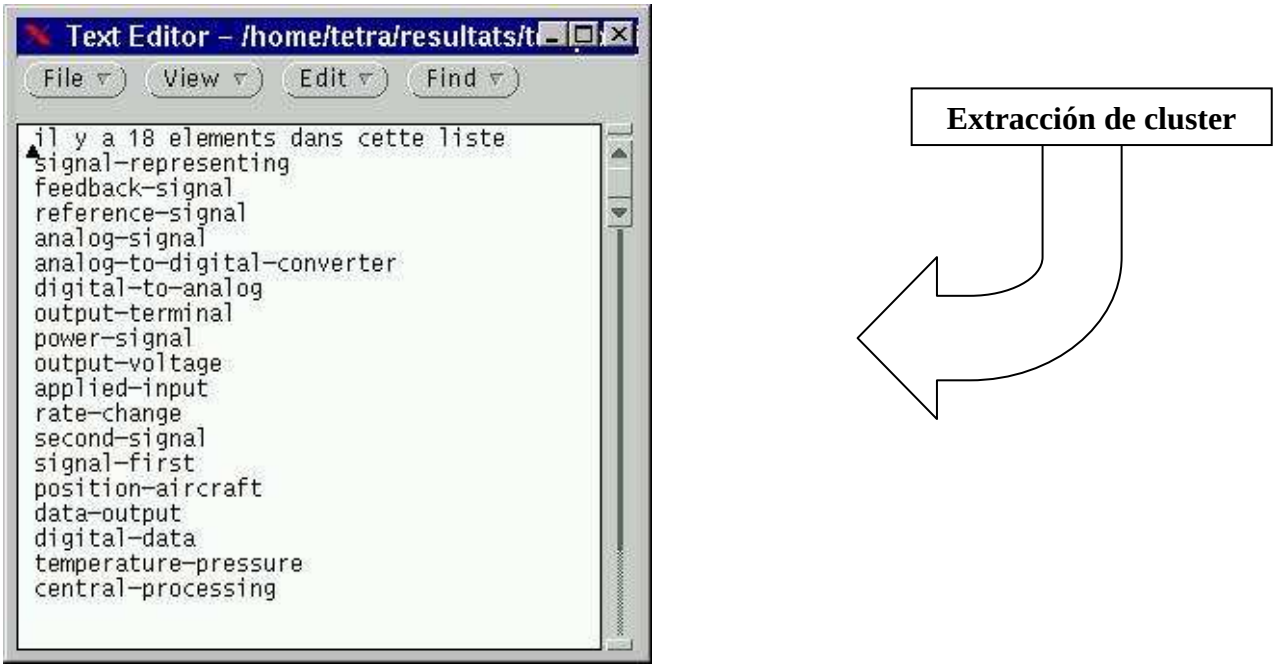

Figura 5.: Ilustración del método de extracción de señales débiles.

El resultado supera a menudo toda previsión, ya que los conceptos subyacentes son completamente nuevos, lo cual desestabiliza a los expertos, que se reconocen bien a menudo incompetentes en la materia (Roig, 1998). Los nuevos asuntos, así detectados, tienen que ser objeto de un zoom detallado, que puede ser obtenido cruzando su terminología específica con los actores del ámbito y los otros conceptos próximos. Es deseable re-interrogar las bases de información sobre este nuevo tema (del cual la ecuación de búsqueda nos es dada), con el fin de completar su perfil de identidad y delimitar mejor la potencialidad.

2.3. Fenómenos de ruptura

La rápida desaparición de un subámbito o subdominio determinado, de un equipo, o de un actor mayor, puede ser una información estratégica. La consulta de una matriz teniendo tiempo como segunda variable es a menudo suficiente (histograma de evolución, clasificación en función del tiempo, elección de una columna por consistencia). En cambio, cuando se trata de poner al día una reorientación temática, un cambio de alianza o simplemente el fin de una colaboración, es necesario hacer intervenir dos variables y el tiempo. Uno se gira pues hacia el análisis de las matrices 3D y el conjunto de los métodos asociados.