Journal of Technology Management & Innovation
E-ISSN: 0718-2724
ljimenez@jotmi.org
Universidad Alberto Hurtado
Chile

Gaitán Ospina, Carlos Felipe
Vigilancia Tecnológica Cientifica de Ciclos Biogeoquímicos
Journal of Technology Management & Innovation, vol. 4, núm. 2, 2009, pp. 44-53
Universidad Alberto Hurtado
Santiago, Chile

Disponible en: http://www.redalyc.org/articulo.oa?id=84711432004

Journal of Technology Management & Innovation

Received April 24, 2009 / Accepted May 29, 2009 J. Technol. Manag. Innov. 2009, Volume 4, Issue 2

Vigilancia Tecnológica Cientifica de Ciclos Biogeoquímicos

Carlos Felipe Gaitán Ospina [1]

Abstract

The use of new tools is needed in order to study the geosciences, where gathering information related with different variables is required. This activity implies the use of human and financial resources; nevertheless, each year many losses are registered because of the use of resources in the development of previously invented technologies. This fact and the growing amount of information at patent offices and *Proquest* databases, encouraged the design of a tool capable of classifying this kind of information.

Keywords: technological surveillance; SOM, unsupervised learning; clustering; data basesparty.

Resumen

El estudio de las geociencias requiere el uso de instrumentos y desarrollos de última tecnología, que permitan monitorear y recopilar información relacionada con diferentes variables de estudio como precipitación, humedad relativa y radiación, actividad que implica la destinación de grandes recursos financieros y humanos; sin embargo, anualmente se registran millonarias pérdidas por la asignación de recursos a desarrollos ya inventados. Esta situación llevo a plantear el desarrollo de una herramienta que clasificara la información de bases de datos especializadas como las usadas en oficinas de patentes y por *Proquest* para la consulta de artículos científicos.

Palabras clave: vigilancia tecnológica; mapas auto-organizados; aprendizaje no supervisado; bases de datos.

[1] Graduate Research Assistant. 6339 Stores Road. Department of Earth and Ocean Sciences. University of British Columbia. Vancouver, BC. V6T 1Z4, Canada

I. Introducción

Ante la posibilidad de un futuro en el que se presentarían inundaciones, sequias, tormentas, huracanes y un aumento general en la temperatura media del planeta, la humanidad se enfrenta al reto de monitorear y desarrollar dispositivos y técnicas eficientes que permitan estudiar los diferentes componentes del sistema terrestre: agua, suelo, aire y biota. Sin embargo, la invención de todos estos dispositivos requiere la asignación de grandes recursos humanos y financieros, que permitan avanzar constantemente en el estado del conocimiento. Es así como Estados Unidos, Japón, China y Alemania, invirtieron en el 2001, en Investigación y Desarrollo (I+D), 282.000 millones, 104.000 millones, 60.000 millones y 54.000 millones de dólares respectivamente, según información de la *Organización de Cooperación y Desarrollo Económicos* (OCDE).

Ante este escenario, y dada la limitación de recursos económicos en los países en vía de desarrollo para invertir en Investigación y Desarrollo de tecnología, es necesario optimizar el uso de los recursos, y no invertir en "inventar" tecnologías y desarrollos que ya se han creado, ya que, sólo por citar un ejemplo, la *Oficina Europea de Patentes* manifiesta que se registran pérdidas de veinte mil millones de dólares en la Unión Europea por desarrollos y procesos ya inventados. Para reducir estas pérdidas, la solución más simple consiste en investigar el estado actual de la tecnología o el desarrollo de interés a nivel mundial, siendo clave el análisis de la información, obtenida en bases de datos especializadas en publicaciones científicas y el análisis de información en las diferentes oficinas de patentes. Estas actividades, las trata la Vigilancia Tecnológica, herramienta de gestión asociada con acciones de observación, captación y análisis de información, para convertir señales dispersas en tendencias y recomendaciones para tomar decisiones.

Respecto a la herramienta desarrollada, se optó usar la Inteligencia Artificial, específicamente aprendizaje no supervisado con Redes Neuronales Artificiales, y la creación de mapas auto-organizados, aplicaciones que son abordadas utilizando el algoritmo de *Kohonen*. Este algoritmo ha sido usado, para la detección de arritmias cardiacas, la clasificación de aminoácidos, identificación de patrones económicos, y la caracterización de descargas atmosféricas. Este trabajo implementa el algoritmo en clasificación de datos multidimensionales, no numéricos; específicamente, para la clasificación de publicaciones técnicas, relacionadas con el ciclo hidrológico y los ciclos biogeoquímicos, en cuatro grandes clústeres de información.

II. Descripción del trabajo

El presente documento aborda someramente dos áreas principales que se trataron en la tesis de posgrado del autor: A) La implementación del algoritmo de Redes Neuronales de Kohonen (RNK) para la creación de clústeres de información, y B) el desarrollo de mapas auto organizados o SOMs (por sus siglas en inglés) de las publicaciones técnicas y científicas relacionadas a los ciclos biogeoquímicos e hidrológico. Así mismo, se muestra a continuación un aparte del estudio de patentes relacionadas a los campos tecnológicos asociados a los ciclos biogeoquímicos e hidrológico, realizado utilizando datos de la USPTO.

Como se observa en la Tabla 1, en el periodo 2003-2007 Norteamérica registró 17957 patentes en las clases tecnológicas asociadas a los ciclos biogeoquímicos e hidrológico, mientras que Centro y Suramérica, sólo registraron 30 patentes en el mismo periodo, para las estas clases tecnológicas. De esas 30 Colombia registró una patente, en la clase tecnológica 374 Pruebas y Mediciones Termales.

ISSN: 0718-2724. (http://www.jotmi.org)

Total patentes en clases tecnológicas asociadas a los ciclos biogeoquímicos e hidrológico						
	2003	**2004**	**2005**	**2006**	**2007**	**Total**
América del Norte	3486	3632	3311	4140	3358	17927
Asia	2102	2313	2143	3124	2689	12371
Europa	1037	1064	942	1101	1003	5147
Oceanía	56	46	46	60	57	265
Lejano y medio oriente	57	64	42	81	77	321
Centro y Sur América	7	5	4	7	7	30
África	8	4	1	4	8	25
TOTAL ANUAL	**6753**	**7128**	**6489**	**8517**	**7199**	**36086**

Tabla 1 Total de patentes en las clases tecnológicas asociadas a los ciclos biogeoquímicos e hidrológico. (*USPTO*, 2007)

El escenario observado en estos últimos 5 años para el registro de innovaciones relacionadas a los ciclos biogeoquímicos e hidrológico, no puede ser más preocupante para la Latinoamérica, dado que sólo aportó 30 de 36086 patentes para este periodo (0.08%). Superando únicamente a África, y patentando 10 veces menos que el Lejano y Medio Oriente.

Respecto a los líderes mundiales, Estados Unidos, Japón y Alemania se ubican en las tres primeras posiciones en cuanto al número de patentes en las clases tecnológicas asociadas a los ciclos biogeoquímicos e hidrológico.

A. Introducción al Algoritmo de Redes Neuronales de Kohonen (RNK)

Esta arquitectura de red, nombrada así por su creador Tuevo Kohonen, varía considerablemente del modelo más usado, que es el de ***feed forward back propagation (FFBP)***; la red neuronal de Kohonen, no solo difiere en cómo es entrenada, sino en cómo recuerda los patrones; a su vez, estas redes neuronales, no usan funciones de activación, capas ocultas, ni predisposición de pesos. (Kohonen, 1984)

La mayor diferencia entre las RNK y la FFBP, es que la red de Kohonen, se entrena de manera no supervisada; esto significa que a la RNK se le presentan datos, pero la salida correcta para los datos, no es especificada; al usar la RNK, estos datos pueden clasificarse en grupos.

En términos generales, al presentarse un patrón de entrada (input) a la red neuronal de Kohonen (RNK), solo una, de las neuronas de salida (output) es seleccionada como ganadora. Esto se denomina un aprendizaje no supervisado competitivo, ya que no existe ninguna salida objetivo hacia la cual la red neuronal deba tender, a su vez, al competir las neuronas por activarse, queda solo una como neurona vencedora y el resto son forzadas a sus valores de respuesta mínimos. El objetivo de este aprendizaje es categorizar los datos que se introducen en la red. Se clasifican valores similares en la misma categoría y, por tanto, deben activar la misma neurona de salida (Heaton, 2005).

Es importante comprender las limitaciones de las RNK, donde al igual que las redes neuronales con dos capas, presentan mejores resultados al usarse en problemas que pueden ser descompuestos linealmente (Heaton, 2005). De otro lado, las redes neuronales de Kohonen, proporcionan ventajas como facilidad de construcción y agilidad en el proceso de entrenamiento.

ISSN: 0718-2724. (http://www.jotmi.org)

1. Estructura de la RNK

A diferencia de las redes FFBP, las redes neuronales de Kohonen, solo contienen dos capas de neuronas, una de entrada y otra de salida, no tienen capas ocultas. A continuación, se mostraran los conceptos de entrada y salida de datos para las reden neuronales de Kohonen. (Kohonen, 1984)

2. Entrada de datos

La entrada de datos a la RNK, está dada por las neuronas de entrada; estas neuronas tienen asignados números de punto flotante, que corresponden al patrón de entrada de la red. La red neuronal de Kohonen, requiere que estas entradas estén normalizadas, usualmente entre 0 y 1 (0,1) o en el rango entre -1 y 1 (-1,1). Este patrón de entradas a la red, ocasionará que las neuronas de salida reaccionen. (Heaton, 2005)

3. Salida de datos

La salida de una red neuronal de Kohonen, varía de la salida de una red tipo FFBP, en que solo una de las neuronas de salida producirá un valor, adicionalmente, este valor puede ser Verdadero o Falso; al presentarse un patrón a una RNK, solo una neurona es escogida como neurona de salida, motivo por el cual, usualmente la salida de la RNK es el índice de la neurona que se activó (ej. La neurona ganadora es la número 3). (Heaton, 2005)

4. Procesamiento de la información

Para examinar como es el procesamiento de información, se debe comprender cuál es el proceso de cálculo que lleva la red neuronal. Estos pasos incluyen normalización del vector de entradas, cálculo de la neurona de salida, mapeo de números en formato bipolar (opcional), selección de la neurona ganadora, entrenamiento de la red, determinación de la razón de aprendizaje y ajuste de pesos.

5. Normalización la Entrada

Las redes neuronales de Kohonen, requieren que sus entradas sean normalizadas, siendo este requerimiento uno de los mayores limitantes de las RNK. El rango de las variables de entrada, debe ser [-1,1], y cada una de las variables de entrada debe poder usar este rango libremente; se ha encontrado que si una o varias neuronas de entrada usan solo el rango entre [0,1], se perjudica el desempeño de la red neuronal. (Heaton, 2005)

Usando la metodología propuesta por Heaton (2005), para normalizar la entrada, primero se debe calcular la magnitud del vector de los datos de entrada sumando los cuadrados del vector de entrada (ej. $0.5^2+0.1^2 = 0.26$); con el valor de esta magnitud, es posible determinar el factor de normalización; siendo el factor de normalización igual al reciproco de la raíz cuadrada de la magnitud. (ej. $1/(0.26)^{0.5} = 1.96$) El proceso de normalización, será usado al calcular la capa de salida.

6. Calculo de la salida de cada neurona

Usando como ejemplo la Figura anterior, el algoritmo de Kohonen establece que se debe tener en consideración, el vector de entradas y los pesos de conexión entre la neurona 1, de la capa de entradas y los pesos entre esta neurona y cada una de las neuronas de la capa de salida. Una medida usual para relacionar las entradas y los pesos, es calcular la distancia entre los mismos; generalmente se utiliza el concepto de Distancia Euclidiana para este fin. (Buckland, 2005)

Sean P = (x1, y1, z1) y Q = (x2, y2, z2), dos puntos en el espacio. La distancia PQ entre P y Q está dada por:

$$\overline{PQ} = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2}$$

Donde el cálculo de este resultado, debe repetirse para cada neurona de salida y posteriormente normalizarse, una vez calculada y normalizada la salida.

7. Selección de la neurona ganadora

La neurona ganadora es aquella cuya distancia euclidiana entre las entradas y los pesos es la menor. Como se pudo observar, los pesos entre las neuronas de entrada y las de salida, determinan el valor de la salida; a continuación se procederá a explicar el proceso en el que esos pesos son ajustados para producir salidas más adecuadas para una tarea dada; este proceso es el entrenamiento de la red. (Hecht-Nielsen, 1990)

8. Entrenamiento de la red

En términos generales, el entrenamiento de las redes neuronales de Kohonen, es competitivo, ya que para

ISSN: 0718-2724. (http://www.jotmi.org)

cada juego de parámetros de entrenamiento, solo una neurona resulta ganadora; a esta neurona le será reajustado su peso, de manera que reaccione más fuertemente a las entradas la próxima vez; como diferentes neuronas ganan para diferentes patrones, su habilidad para reconocer un patrón específico será aumentada. (Kohonen, 1995)

El proceso de entrenamiento de una RNK, involucra pasar por diferentes épocas (iteraciones), hasta que el error (o distancia entre las entradas y los pesos) de la RNK sea menor a un valor aceptable dado. Este proceso, incluye calcular el error de la red, ajustar los pesos en cada época y establecer cuando no es necesario continuar el entrenamiento.

La RNK es entrenada mediante la repetición de épocas hasta que uno de dos eventos pasa, si el error calculado es aceptable se termina el proceso de entrenamiento, si por el contrario, el error permanece alto respecto al valor aceptable, existen dos opciones, que la razón de cambio de este error sea muy pequeña, ocasionando que este ciclo individual sea abortado y no se realicen épocas adicionales; de ocurrir esto, los pesos se inicializan con valores aleatorios y se iniciara un nuevo ciclo. Este entrenamiento continuara el ciclo previo y realizará de nuevo la verificación de si los pesos producen errores aceptables. (Hassoun, 1995)

Como se puede inferir, el proceso iterativo para reducir los errores es clave en el entrenamiento de la red.

9. Razón de aprendizaje

La razón de aprendizaje es un valor numérico usado por el algoritmo de aprendizaje, puede ser constante o variable en el proceso, teniendo en cuenta que siempre debe ser un número positivo menor que 1; usualmente la razón de aprendizaje es un número entre 0.4 y 0.5 y se representa con la letra griega alpha (α). Generalmente valores más altos de *alpha*, causan que el proceso de aprendizaje sea más rápido, sin embargo pueden causar que la red nunca converja; esto a causa que las oscilaciones en los vectores de pesos, pueden ser tan grandes que impidan a los patrones de clasificación manifestarse. (Kohonen, Self-Organizing Maps, 1995)

Una variante a la técnica anterior, es iniciar *alpha* con valores altos e ir decreciendo su valor conforme avanza el entrenamiento, permitiendo un entrenamiento inicial más rápido de la RNK e ir depurando el proceso a medida que avanza el mismo. Independientemente de si la razón de aprendizaje se toma como variable o constante, esta razón se usa como parte integral del algoritmo que calcula los pesos de las neuronas. (Heaton, 2005)

10. Ajuste de los pesos

La memoria de las RNK, es almacenada dentro de las conexiones ponderadas entre las capas de entrada y de salida; estos pesos son ajustados en cada época o iteración, que busca que la red neuronal presente una respuesta más favorable la próxima vez que el mismo juego de datos de entrenamiento se le presente; estas iteraciones continúan al ingresar nuevos datos a la red y ser los pesos reajustados. Eventualmente, el reajuste de los pesos disminuirá hasta que no sea importante continuar con este juego de pesos, cuando esto ocurre la matriz de pesos se reinicia con valores aleatorios y se crea un nuevo ciclo. (Heaton, 2005)

La matriz de pesos definitiva que será usada, corresponderá a la mejor matriz de pesos determinada en cada uno de los ciclos.

El método original para calcular los cambios en los pesos, o método aditivo, fue propuesto por T. Kohonen (1984) y usa la siguiente ecuación:

$$w^{t+1} = \frac{w^t + \alpha x}{\left\| w^t + \alpha x \right\|}$$

Donde la variable x es el vector de entrenamiento que fue presentado a la red, la variable w^t es el peso de la neurona ganadora, y la variable w^{t+1} es el nuevo peso. Las líneas verticales dobles representan la magnitud del vector. (Kohonen, 1984)

Aunque el método aditivo usualmente trabaja bien con las RNK, existen ocasiones en las que el método es excesivamente inestable y no converge, siendo necesario utilizar un método alterno, como el método sustractivo. (Heaton, 2005)

El método sustractivo usa las siguientes ecuaciones, para transformar los pesos de la red:

ISSN: 0718-2724. (http://www.jotmi.org)

$$e = x - w^t$$

$$w^{t+1} = w^t + \alpha e$$

a) Calculo del Error

El propósito de las RNK, es clasificar datos de entrada en varios juegos, luego el error para las redes neuronales de Kohonen, debe ser capaz de medir que tan bien se están clasificando los datos de entrada. Esta característica permite que diversos cálculos de errores hayan sido propuestos para las RNK, sin ser ninguno oficial.

Es importante resaltar que al ser el entrenamiento no supervisado, se debe replantear la formula de calcular el error, siendo aceptable la cotejar el valor esperado del entrenamiento y el valor actual; esto debido a que no existen salidas anticipadas que permitan la comparación entre lo observado y lo simulado (Fort, Letremy, & Cottrell, 2002).

En términos generales, el error es mínimo cuando la distancia euclidiana entre las entradas y los pesos es cero.

B. Mapas Auto – Organizados de *Kohonen* (SOM)

Los SOM proveen una forma de representar datos multidimensionales en espacios dimensionales menores, usualmente en una o dos dimensiones; esta proceso es básicamente una técnica de compresión de datos, conocida como Cuantización Vectorial; adicionalmente, esta técnica crea una red que almacena información, de manera tal que las relaciones topológicas entre los elementos del conjunto se mantienen. (*Kohonen*, 1995)

1. Proceso de entrenamiento:

El entrenamiento ocurre en varios pasos y numerosas iteraciones, de la siguiente manera (Buckland, 2005):

1. Los pesos de cada nodo son inicializados.
2. Un vector se escoge del juego de datos de entrenamiento y es presentado a la malla.
3. Cada nodo se examina para calcular que juego de pesos es más similar al vector de entrada. El nodo ganador se determina como la Mejor Unidad de Coincidencia (MUC).
4. El radio de la MUC es calculado; este valor inicia con un valor elevado, generalmente el radio de la malla, pero disminuye en cada paso de tiempo. Todos los nodos que se encuentren dentro de este radio, se consideran que pertenecen a la vecindad de la MUC.
5. Para cada uno de los nodos encontrados en el vecindario de la MUC, se ajustan los pesos con el fin de hacerlos "mas similares" al vector de entrada. Entre más cerca se encuentre un nodo a la Mejor Unidad de Coincidencia, sus pesos serán alterados de mayor manera.
6. Repetir el paso 2 por N iteraciones.

2. Metodología

A continuación se listan los aspectos fundamentales considerados para la creación de los SOMs. Para el lector interesado en el particular, se recomienda consultar el trabajo de grado en el que se basa el presente documento (Gaitan Ospina, 2008).

- Se trabajó con la base de datos *ProQuest* que brinda acceso a mas de 125 millones de documentos electrónicos
- Criterios de búsqueda : BIOGEOCHEMISTRY AND HYDROLOGY
- Se crearon 4 familias de palabras: ECO, GEO, CHEM e HYDRO
- Cada familia representa un clúster de datos en el SOM
- Se utilizó el SOMPack para Matlab (The MathWorks Inc., 2008) , desarrollado por la Universidad de Helsinki para crear los mapas.

III. Resultados

La siguiente tabla muestra la familia predominante en cada una de las 72 celdas del SOM generado. Como se observa los clústeres están ubicados en las zonas correspondientes a la familia predominante.

ISSN: 0718-2724. (http://www.jotmi.org)

FAMILIA PREDOMINANTE EN CADA CELDA DEL MAPA AUTO ORGANIZADO									
		Columna No.							
		1	2	3	4	5	6	7	8
Fila No	1	CHEM	CHEM	CHEM	CHEM	CHEM	CHEM	CHEM	HYDRO
	2	CHEM	CHEM	HYDRO	HYDRO	CHEM	CHEM	HYDRO	HYDRO
	3	CHEM	CHEM	CHEM	HYDRO	HYDRO	HYDRO	HYDRO	HYDRO
	4	CHEM	CHEM	HYDRO	HYDRO	HYDRO	HYDRO	HYDRO	HYDRO
	5	GEO	GEO	HYDRO	HYDRO	HYDRO	HYDRO	GEO	GEO
	6	ECO	ECO	ECO	ECO	HYDRO	HYDRO	GEO	GEO
	7	ECO	ECO	ECO	ECO	ECO	ECO	ECO	GEO
	8	ECO	ECO	ECO	ECO	ECO	ECO	HYDRO	HYDRO
	9	ECO	ECO	ECO	ECO	ECO	ECO	ECO	HYDRO

Tabla 2 Familia predominante en cada celda del mapa auto organizado

La Tabla 3 muestra el porcentaje de ocurrencias en cada una de las 72 celdas del SOM generado. Se aprecian valores superiores al 4% en la esquina superior izquierda e inferior izquierda, correspondientes a las celdas con 11 y 10 ocurrencias respectivamente. En el caso de la celda del clúster CHEM, corresponde a los títulos cuyos valores dimensionales en cada una de las cuatro familias es bajo, sugiriendo para futuros desarrollos ampliar el número de palabras asociadas a cada familia, para mejorar así la clasificación. Por el contrario la celda del clúster ECO-BIO, representa 10 títulos que fueron catalogados principalmente en esta categoría.

ISSN: 0718-2724. (http://www.jotmi.org)

Porcentaje de ocurrencias por celda del SOM generado								
		Columna No.						
		1	2	3	4	5	6	7	8
Fila No	1	5.023	1.370	1.370	2.740	0.913	1.370	2.740	0.913
	2	3.196	1.370	0.913	1.826	1.370	1.370	0.457	0.457
	3	0.000	1.826	3.196	1.370	0.913	0.457	1.826	0.913
	4	1.370	0.913	1.826	1.370	0.913	1.370	3.196	0.457
	5	0.913	0.913	0.457	1.826	2.740	0.913	1.370	0.913
	6	2.740	1.370	0.457	1.370	2.283	3.196	0.000	1.370
	7	1.370	0.000	0.457	1.826	0.457	0.457	2.283	0.457
	8	0.457	0.000	2.283	0.457	1.370	0.000	0.913	0.457
	9	4.566	2.283	0.000	2.283	0.457	0.913	2.283	1.826

Tabla 3 Porcentaje de ocurrencias por celda del SOM generado

La tabla 4 muestra el porcentaje de títulos catalogados en cada familia o clúster, usando el SOM generado, como se observa los clústeres ECO, CHEM e HYDRO aportan más del 30 % de los títulos clasificados cada uno y el clúster GEO sólo clasificó el 5.94 % de los mismos. Esto indica una tendencia similar en cuanto a la temática de las obras científicas y técnicas publicadas, con casi iguales porcentajes de categorización.

FAMILIA	PORCENTAJE
ECO	30.14
GEO	5.94
CHEM	30.14
HYDRO	33.79

Tabla 4 Porcentaje de títulos catalogados en cada familia usando el SOM

De otra parte las tablas anteriores muestran que la clasificación de obras en estos 3 clústeres es efectiva y por lo tanto valida la importancia de cada una de estas familias; sin embargo, al encontrar valores inferiores al 10 % en la distribución total de celdas del SOM y en el porcentaje de títulos clasificados en la familia GEO, surge el interrogante de la relevancia de esta categoría como clúster principal. Se sugiere para siguientes desarrollos utilizar GEO como subcategorías de otra familia, probablemente ECO por la similitud en sus temáticas.

Finalmente la Ilustración 1 muestra los diferentes SOM generados para cada una de sus dimensiones vectoriales (BIO, GEO, CHEM, HYDRO), donde la superposición de estas graficas dimensionales crea un SOM análogo al mostrado en la Tabla 2. En la Ilustración las celdas con colores cálidos concentran los registros relacionados a cada uno de los grupos.

ISSN: 0718-2724. (http://www.jotmi.org)

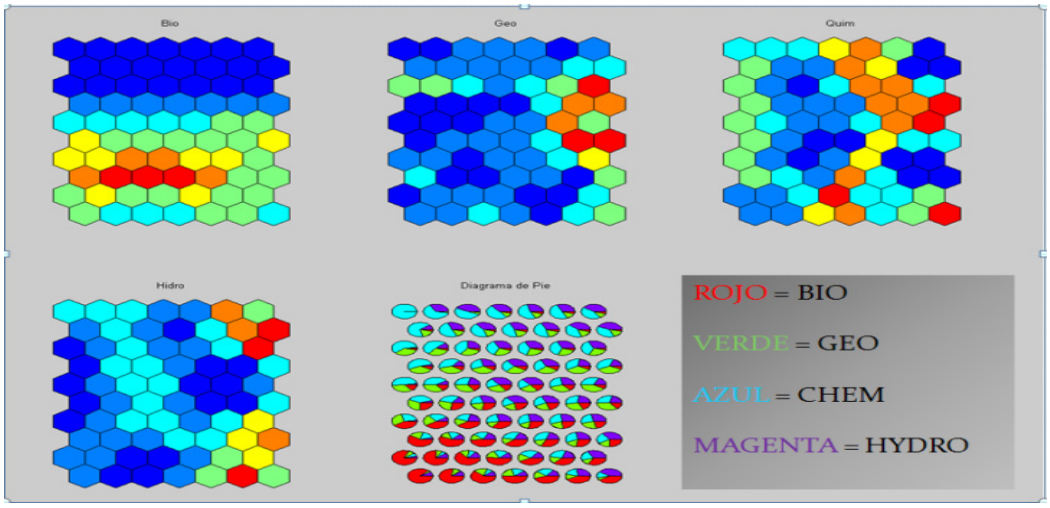

Ilustración 1. SOMs para cada dimensión vectorial.

La tabla 5 muestra la distribución de celdas del SOM entre las cuatro familias o clústeres creados. Se observa un predominio de las familias ECO e HYDRO con cerca de un tercio de las celdas asignadas para cada uno, seguido por CHEM y un pequeño clúster de la familia GEO compuesto por 7 celdas (9.7 %).

FAMILIA	Número de Celdas	Porcentaje
ECO	24	33.33
GEO	7	9.72
CHEM	16	22.22
HYDRO	25	34.72
TOTAL	72	100

Tabla 5 Distribución de celdas dentro de la clasificación obtenida por el SOM

La tabla 6 muestra el porcentaje de títulos catalogados en cada familia o clúster, usando el SOM generado, como se observa los clústeres ECO, CHEM e HYDRO aportan más del 30 % de los títulos clasificados cada uno y el clúster GEO sólo clasificó el 5.94 % de los mismos. Esto indica una tendencia similar en cuanto a la temática de las obras científicas y técnicas publicadas, con casi iguales porcentajes de categorización.

FAMILIA	PORCENTAJE
ECO	30.14
GEO	5.94
CHEM	30.14
HYDRO	33.79

Tabla 6 Porcentaje de títulos catalogados en cada familia usando el SOM

De otra parte las tablas anteriores muestran que la clasificación de obras en estos 3 clústeres es efectiva y por lo tanto valida la importancia de cada una de estas familias; sin embargo, al encontrar valores inferiores al 10 % en la distribución total de celdas del SOM y en el porcentaje de títulos clasificados en la familia GEO, surge el interrogante de la relevancia de esta categoría como clúster principal. Se sugiere para siguientes desarrollos utilizar GEO como subcategorías de otra familia, probablemente ECO por la similitud en sus temáticas.

ISSN: 0718-2724. (http://www.jotmi.org)

IV. Conclusiones y recomendaciones

Se encontró muy útil la capacidad de crear mapas auto organizados con el algoritmo de *Kohonen*, ya que uno de los principales retos de la Vigilancia Tecnológica, es lidiar con volúmenes de información cada vez mayores, como número de publicaciones y número de patentes. En este orden de ideas, resulta ser una solución adecuada para visualizar y organizar información obtenida de grandes bases de datos.

De otra parte, los mapas auto organizados ofrecen una manera alterna de visualizar la información, dado que es posible asignarle color, intensidad, brillo u otras características a los clústeres de información formados luego del proceso de entrenamiento de la red; estas mismas características lo han llevado a ser usado en la clasificación de moléculas químicas o de células con cáncer. En el caso de su aplicación a la Vigilancia Tecnológica, puede clasificar cualquier vector de datos n-dimensional, independientemente de si sus valores son numéricos o alfanuméricos. Esta funcionalidad abre un gran espectro de posibilidades para clasificar, tecnologías, patentes, autores, empresas, inventores o incluso países, según determinadas características de interés para la Vigilancia Tecnológica que se quiera llevar a cabo.

Un aporte adicional es la implementación de un sistema de clasificación dimensional para determinar el grado de pertenencia de una oración (título) a una determinada clase o familia. Generalmente esta clasificación es booleana y solo determina existencia o no existencia de una palabra

Se propone realizar futuros estudios para probar la herramienta en la creación de 5 o más clasificaciones, así como desarrollar un método para determinar las familias de los clústeres automáticamente. Se sugiere implementar búsquedas que cuenten todas las palabras en los registros y luego determinar las de mayor ocurrencia entre los sustantivos.

Referencias

BUCKLAND, M. (2005, Abril 20). *Neural Network Tutorial in Plain English.* AI Junkie: http://www.ai-junkie.com/ann/som/som1.html (accesed in May 13, 2008)

FORT, J.-C., Letremy, P., & Cottrell, M. (2002). Advantages and drawbacks of the Batch Kohonen algorithm.

GAITAN Ospina, C. F. (2008). *Desarrollo e implementación de procesos de Vigilancia Tecnológica asociados al estudio de los ciclos biogeoquímicos e hidrológico.* Bogota.

HASSOUN, M. H. (1995). *Fundamentals of Artificial neural Networks.* Cambridge, Massachusetts: MIT Press.

HEATON, J. (2005). *Introduction to Neural Networks with JAVA.* St. Louis: Heaton Research.

HECHT-NIELSEN. (1990). *Neurocomputing.* Massachusetts: Addison-Wesley.

KOHONEN, T. (1984). *Self-Organization and Associative Memory.* Berlin: Springer-Verlag.

KOHONEN, T. (1995). *Self-Organizing Maps.* Berlin: Springer-Verlag.

ProQuest. (2008). *ProQuest.* http://www.proquest.com/division/aboutus/ (accesed in May 20, 2008)

The MathWorks Inc. (2008). Neural Network Toolbox™ User's Guide. In H. Demuth, M. Beale, & M. Hagan. Natick: MathWork.

USPTO. (2008). *USPC-to-IPC Reverse Concordance*. http://www.uspto.gov/go/classification/international/ipc/ipc8/ipc_concordance/ipcsel.htm (accesed in July 08, 2008)

Acerca del Autor

Carlos Felipe Gaitán Ospina es Ingeniero Civil y Magister en Hidrosistemas de la Pontificia Universidad Javeriana de Bogotá, Colombia, donde desarrolló su tesis de posgrado en "Desarrollo e implementación de procesos de Vigilancia Tecnológica asociados al estudio de los ciclos biogeoquímicos e hidrológico", usando herramientas de Inteligencia Artificial y técnicas de aprendizaje no supervisado como las Redes Neuronales de *Kohonen*. Actualmente es Asistente de Investigación en la Universidad de British Columbia, Canadá, donde cursa sus estudios doctorales.

ISSN: 0718-2724. (http://www.jotmi.org)