ViBLIOSOM: VISUALIZACIÓN DE INFORMACIÓN BIBLIOMÉTRICA MEDIANTE EL MAPEO AUTOORGANIZADO

Gilberto Sotolongo-Aguilar*, María Victoria Guzmán-Sánchez*, Humberto Carrillo**

Resumen: En este trabajo se describe el uso de una herramienta de visualización que facilita descubrir conocimientos en bases de datos. Se presentan los mapas científico-tecnológicos asociados a los indicadores bibliométricos. Los mapas están basados en el concepto de los mapas auto-organizados (SOM, Self-Organizing Maps) y basados en la tecnología de las redes neuronales no supervisadas. Se incluyen ejemplos que permiten ilustrar la visualización de los resultados.

Palabras clave: visualización de información, mapas autoorganizados (SOM); bibliometría; redes neuronales artificiales.

Abstract: In this paper the use of a visualization tool that helps researchers discover knowledge in databases is described. The paper presents the scientific and technological maps associated with bibliometric indicators. The maps are based on the concept of Self-Organizing Maps (SOM) which is a particularly robust form of unsupervised neural networks. Examples illustrating the visualization of information are included in the paper.

Keywords: information visualization; Self-Organizing Maps (SOM); bibliometrics; artificial neural networks.

1 Introducción

Los profesionales que necesitan manejar grandes volúmenes de información por las características de su trabajo, saben que uno de los retos actuales es conocer y dominar el uso de herramientas que les permitan procesar grandes almacenes de datos. A esta área de trabajo se le ha dado en llamar *minería de datos* y posteriormente *minería de textos*. Muchos problemas de las disciplinas relacionadas con los estudios métricos de la información requieren de la aplicación de estas técnicas.

Las redes neuronales artificiales (RNA) han probado ser de gran utilidad para resolver problemas de minería de datos y texto. Estas han sido útiles particularmente en la organización creativa de información, el descubrimiento de conocimiento y la visualización de información. Esta última es entendida como el «proceso de interiorización del conocimiento mediante la percepción de información» (1). La visualización de información, según la definición anterior, interviene en el paso de datos a información y en la posibilidad de la construcción del conocimiento, al revelar los patrones que subyacen a los datos.

* Instituto Finlay. Habana, Cuba. Correo-e: gsotolongo@finlay.edu.cu; mvguzman@finlay.edu.cu.
** Laboratorio Dinámica no lineal, Fac. Ciencias, UNAM, México. Correo-e: carr@servidor.unam.mx.

Uno de los métodos preferidos para lograr lo anterior es a partir de la «metáfora visual» y su representación en forma de mapas topográficos. En ese sentido, las redes neuronales también han encontrado aplicabilidad, específicamente el modelo desarrollado por Teuvo Kohonen (2), empleado por los autores de este trabajo con resultados interesantes.

Una red neuronal, según Freman y Skapura (3), es un sistema de procesadores paralelos conectados entre sí en forma de grafo dirigido. Esquemáticamente cada elemento de procesamiento (neuronas) de la red se representa como un nodo. Estas conexiones establecen una estructura jerárquica que, tratando de emular la fisiología del cerebro, busca nuevos modelos de procesamiento para solucionar problemas concretos del mundo real. Una definición simplificada sobre los mapas topológicos podría ser que, en una correspondencia que respete la topología, las unidades que se encuentran físicamente próximas entre sí van a responder a clases de vectores de entrada que, análogamente, se encuentren cerca unos de otros. Los vectores de entrada de muchas dimensiones son representados sobre el mapa bidimensional, de tal manera que se mantenga el orden natural de los vectores de entrada [4, 5]. En el procesamiento interno, las RNA realizan una clasificación de los datos; o sea, formación de clusters a partir de la cercanía o similitud entre las variables que son objeto de análisis.

Con el algoritmo de los Mapas Auto-Organizados (SOM o *Self-Organizing Maps*), la información de entrada se organiza automáticamente, lo que permite visualizar relaciones importantes entre los datos, a través de mapas bidimensionales de conceptos. Este aspecto es relevante dentro de las disciplinas relacionadas con el descubrimiento de información en grandes bases de datos (*Knowledge Discovery in Databases, KDD*). Al respecto, es importante mencionar el trabajo de Polanco y colaboradores, del *Institut d'Information Scientifique et Technique* (INIST) de Francia. Este equipo de trabajo ha desarrollado un sistema llamado *Neurodoc*, que se basa en el algoritmo SOM de las redes neuronales artificiales, con resultados interesantes (6).

La visualización de la información bibliométrica, hoy por hoy, está en un estadio preliminar. Una consecuencia de ello es el surgimiento de una serie de metodologías y herramientas que carecen de la validación de sus resultados, muchas de ellas con fines muy específicos para un área determinada o un problema; una base de datos señalada, o con un grupo de indicadores limitados (7). A lo anterior se suma lo inaccesibles que pueden ser los software o herramientas de procesamiento para unidades pequeñas de análisis; o bien, con limitados recursos económicos. Esto ha llevado a explorar otros «modos de hacer» y tratar de adaptar sistemas creados con propósitos diferentes al análisis bibliométrico.

En este trabajo se aborda el uso de las redes neuronales artificiales, basadas en el algoritmo de los Mapas Auto-Organizados, con el objetivo de lograr representaciones visuales de los datos que resultan de la aplicación de los indicadores bibliométricos. A este desarrollo le hemos dado en llamar ViBlioSOM por la correspondencia con Visualization – Bibliometrics – Mapas Auto-Organizados (SOM). Al final, se muestran ejemplos que representan la utilidad de esta herramienta y que ilustran lo amigable de la interfase visual para el usuario final.

2 ViBlioSOM

El antecedente del ViBlioSOM es el desarrollo de una metodología (MOBIS-ProSoft) que consiste, en esencia, en un sistema modular abierto basado en diferentes software propietario (8). Esta metodología permitía aplicar una serie importante de indicadores bibliométricos y obtener resultados interesantes, incluyendo una representación visual de algunas variables en forma de mapas (obtenidos a partir de sistemas como el *Statistic*). Aún así, el procesamiento y visualización de datos más complejos, como las correlaciones entre dos o más variables; las representaciones de grandes matrices o de datos, resultado del análisis de textos (título, resumen, *claims* en patentes, etc.) tenía limitaciones.

Es por ello que, como complemento al MOBIS-ProSoft, se comienza la búsqueda de una herramienta, a bajo costo, que permita solventar los problemas antes mencionados. Al respecto se valoraron una serie de alternativas hasta llegar al Viscovery®SOMine, desarrollado por la firma austriaca *Eudaptics Software Gmbh*. Este sistema utiliza el algoritmo SOM para elaborar los mapas topográficos.

Los mapas basados en el algoritmo SOM, en esencia, están inspirados en las propias funciones de la corteza cerebral y ésta es, posiblemente, la estructura más fascinante que existe en la fisiología humana. La corteza es en esencia una capa extensa (aproximadamente de 1 m^2, en humanos adultos) y fina (entre 2 y 4 mm de grosor) que consta de seis capas de neuronas (con un gran nivel de interconexión entre ellas). La corteza está plegada en la forma conocida con el objetivo de maximizar la densidad de empaquetado en el cráneo (3), si esa corteza plegada se extiende se obtendrá una hoja plana con neuronas o elementos de procesamiento. Este hecho natural es tratado de emular por las RNA de forma computacional, constituyendo la inspiración biológica del ya mencionado Teuvo Kohonen (2) para desarrollar en la década de los años ochenta los Mapas Auto-Organizados.

En los mapas cada documento (podría ser una patente) ocupa un lugar en el espacio, en función de sus contenidos temáticos. Cada área del mapa refleja un contenido específico y los tópicos van variando suavemente a lo largo del mismo. Es decir, se establece una correspondencia entre la información de entrada y un espacio de salida de dos dimensiones, los datos de entrada con características comunes activarán zonas próximas en el mapa (5).

El Viscovery®SOMine no está concebido por sus desarrolladores como un sistema de análisis y visualización de información bibliométrica. Sin embargo, la fortaleza del algoritmo de las redes neuronales artificiales permite visualizar cualquier tipo de dato: el único requisito es que estos datos estén distribuidos según el formato de una hoja de calculo.

El ViBlioSOM es muy útil para realizar análisis de correlación entre variables o datos complejos y en la clasificación de información. Con relación a esto último, permite realizar filtrajes de clusters ya formados y profundizar en el análisis de las variables que lo componen. Las ventajas alcanzadas con este método consisten en que ha permitido organizar visualmente la información bibliométrica y patentométrica (análisis bibliométrico de los documentos de patentes). Ha sido de ayuda, además, para percibir la estructura del conjunto de los datos y para realizar análisis de información con «ruido».

Todo lo anterior ha permitido enriquecer el procesamiento, visualización y análisis de los indicadores bibliométricos, con una metodología propia y a bajo costo. Adicionalmente, se ha considerado la validación metodológica del sistema (7).

Este método puede ser aplicado a cualquier campo del saber y tiene un vínculo muy estrecho con los procesos de inteligencia empresarial, vigilancia científico-tecnológica, gestión del conocimiento y evaluación de proyectos. Igualmente, el método puede ser aplicado en servicios bibliotecarios e informativos y en observatorios de ciencia y tecnología (9, 10, 11, 12, 13).

3 Aplicaciones

Las ventajas estratégicas y económicas que representan los resultados obtenidos han sido identificadas en las propias aplicaciones prácticas hechas por el equipo de trabajo que lo desarrolló; así como por otras instituciones de diferentes sectores económicos del país, que ya la han implementado. A partir de estos trabajos se han identificado situaciones estratégicas que no han sido divulgadas, como son las líneas tecnológicas en las que trabajan los competidores, alianzas entre empresas, tecnologías emergentes y en declive, etc. Se ha evaluado la situación científico-tecnológica de aspectos importantes dentro de la investigación o la producción y se ha medido la relación entre la investigación y la innovación.

Es muy útil para la gestión de los activos intelectuales de una empresa. Se tiene la experiencia concreta de las empresas farmacéuticas, que necesitan contar con un dispositivo orgánicamente estructurado que les permita, entre otras cosas, justipreciar su capital intelectual (conocimiento científico-tecnológico); así como hacer una mejor gestión del mismo con la competencia como horizonte (14).

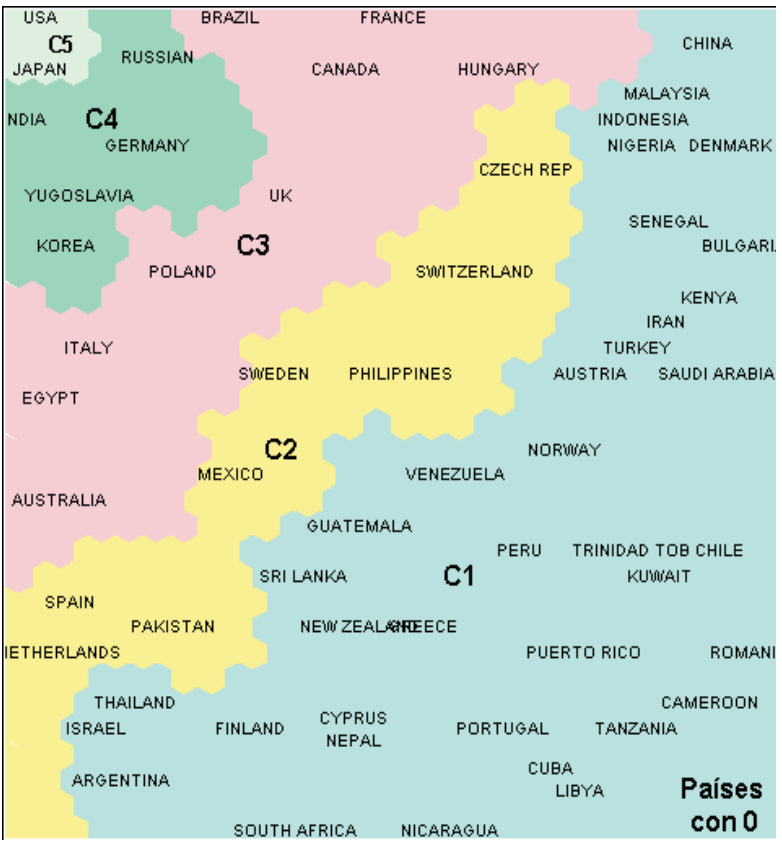
El ViblioSOM puede ser aplicado en cualquier campo del conocimiento; por ejemplo, en un reciente estudio realizado en la temática de la soja aplicada a la alimentación humana (15) (Industria de los Alimentos), se obtuvieron los resultados que aparecen en las figuras 1 y 2.

En la figura 1 se aprecia cómo en el mapa, que representa la situación de la actividad científica de la temática en la década de los ochenta, Cuba denota una modesta actividad. Sin embargo esta situación comienza a cambiar en el mapa siguiente (cluster 2, Figura 2) donde se observa cómo la situación de Cuba mejora y se separa de los países con baja actividad. La década de los noventa coincide con la ruptura de la ayuda soviética a Cuba, afectándose sensiblemente el campo de la alimentación humana y animal. Quizás por ello Cuba se siente presionada a investigar sobre nuevos productos y procesos. En esta etapa (1990-1999) se detecta una actividad similar a Cuba en países como España y Polonia (cluster 2, figura 2).

En el plano internacional los países que han mantenido una actividad científica más estable han sido Estados Unidos (USA) y Japón (cluster 5, figura 1 y 2). En Latinoamérica destaca la posición de Brasil (cluster 4, Figura 2), considerado el mayor exportador de la región después de Estados Unidos (16).

Este ejemplo es importante para evaluar la dinámica de una línea de investigación, la evolución de un país a través de los años y/o establecer el ciclo de vida de un producto. Se ha trabajado además en otras aplicaciones como (a) en la identificación y caracterización de los procesos de mejoramiento del petróleo pesado; (b) en identifi-

Figura 1
Distribución de la actividad científica por países según los años (1980-1989). Análisis de la soja en la alimentación humana

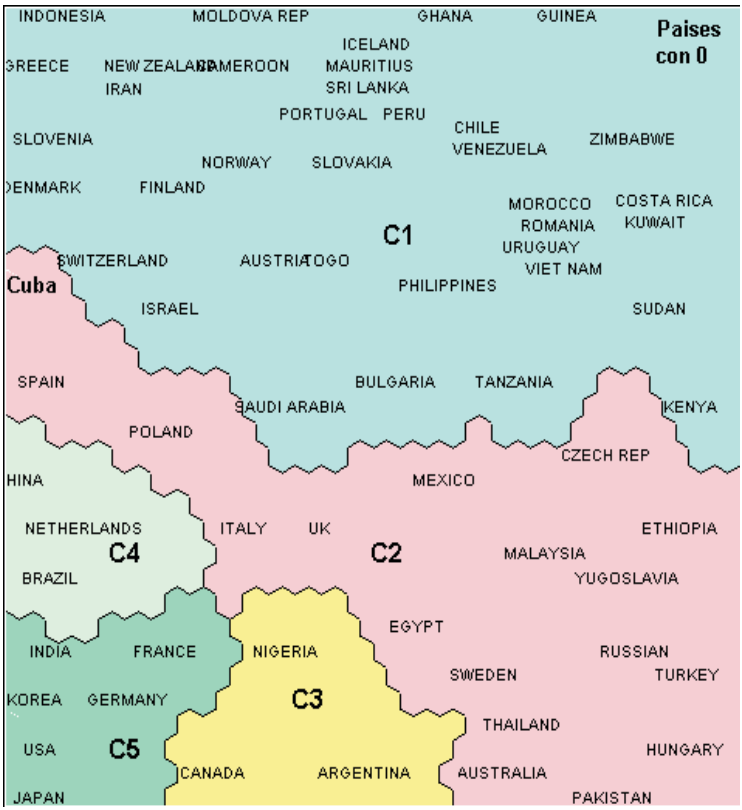


car las líneas y tendencias de la investigación relacionada con la *Neisseria meningitidis*; y (c) en la identificación de alianzas tecnológicas no públicas entre empresas del sector farmacéutico.

Uno de los trabajos que se ha retomado recientemente está relacionado con la identificación de las posibles aplicaciones de la dinámica no lineal, investigación que se desarrolla en conjunto con el Laboratorio de Dinámica no Lineal de la Universidad Nacional Autónoma de México (UNAM). En los resultados preliminares se identificó que las aplicaciones a partir del año 1996, se habían dirigido hacia el campo de la biomedicina (M5, figura 3). Se pudieron obtener como aplicaciones más significativas, los modelos biológicos y cardiovasculares.

El ViBlioSOM se ha aplicado también al campo de las finanzas y al estudio de clientes; incluso se tiene experiencia sobre su aplicación para optimizar la gestión de los fondos bibliotecarios.

Figura 2.
Distribución de la actividad científica por países según los años (1990-1999). Análisis de la soja en la alimentación humana



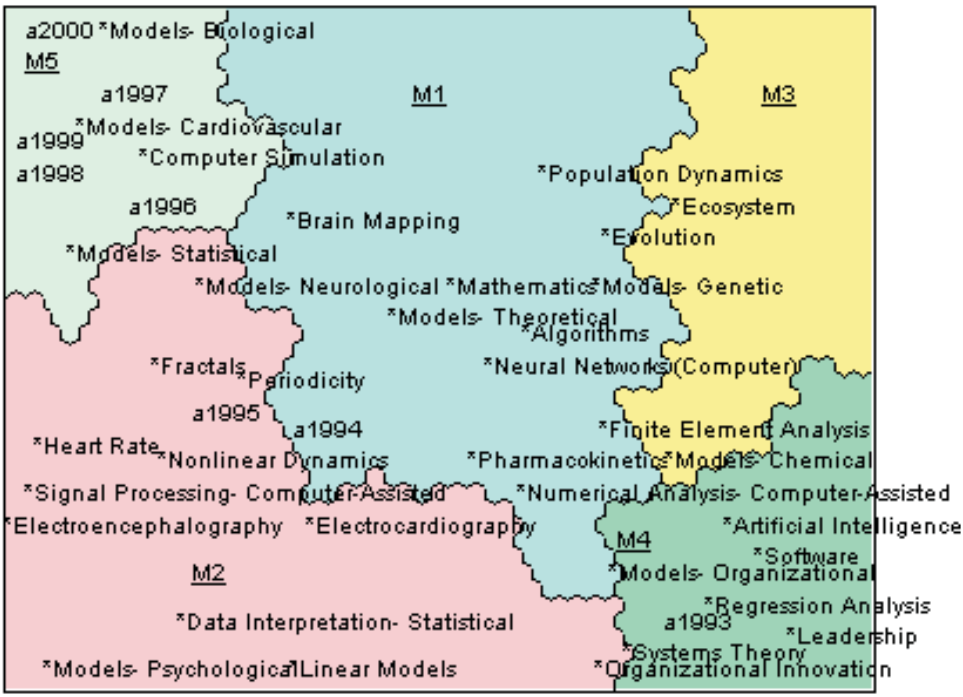
4 Conclusiones

Entre las perspectivas de trabajo con relación al ViBlioSOM se anotan: desarrollar mapas científico-tecnológicos interactivos y completar el ViBlioSOM con otras herramientas, o mejorar las ya existentes. Se tiene particular interés en explorar otras aplicaciones dentro del ámbito empresarial y sobre todo profundizar en la validación del sistema.

La disponibilidad de una tecnología avanzada de minería de textos podría ser usada dentro de los sistemas de vigilancia científico-tecnológica empresariales; o en los observatorios de ciencia y tecnología, a nivel de país o región.

Sería importante que estas herramientas y métodos de análisis tengan un uso e impacto regional. Los países de Ibero-América contarían con una herramienta mucho menos costosa y viable para realizar investigaciones bibliométricas; o bien, podrían incorporarla dentro de los servicios o productos de valor añadido de los centros de información o bibliotecas.

Figura 3
Aplicaciones de la dinámica no lineal. Período 1993-2000



5 Bibliografía

1. CARD, S. K., MACKINLAY, J. D., SHNEIDERMAN, B. *Reading in information visualization*. San Francisco; Morgan Kaufmann Publishers, Inc., 1999.
2. KOHONEN, T. *Self-organizing maps.* Berlin: Springer, 3. ed; 2001.
3. FREEMAN J. A.; SKAPURA D. M. *Redes Neuronales. Algoritmos, aplicaciones y técnicas de programación*. México; Addison-Wesley, 1993.
4. KORNIKOV, A. R. Intelligent technologies new opportunities for modern industry. *Information Technology*, 1997, vol. 3, p.1-14.
5. SOTOLONGO, G.; GUZMÁN, M. V. Aplicaciones de las redes neuronales. El caso de la bibliometría. *Ciencias de la Información*, 2001, vol. 32, p. 27-34.
6. POLANCO, X., FRANCOIS, C., KEIM, J. P. Artificial neural network technology for the classification and cartography of scientific and technical information. *Proceedings of the sixth Conference of International Society for Scientometrics and Informetrics*. 1997, June 16-19, Israel, p. 319-330.
7. SOTOLONGO, G.; GUZMÁN, M. V. SAAVEDRA, O.; CARRILLO, H. A. Mining Informetrics Data with Self-organizing Maps. *Proceedings of the 8 th International Society for Scientometrics and Informetrics*. 2001, July 16-20, Australia, p. 665-673.
8. SOTOLONGO, G; SUÁREZ, C. A.; GUZMÁN, M. V. Modular Bibliometrics Information System with Propietary Software (MOBIS-ProSoft): a versatile approach to bibliometric research tools. *Library and Information Science Electronic Journal (LIBRES),* 2000, Vol.10. http://libres.curtin.edu.au/

9. GUZMÁN, M.V.; SANZ, E.; SOTOLONGO, G. Bibliometrics Study on Vaccines (1990-1995) Part I: Scientific Production in Iberian-American Countries. *Scientometrics*, 1998, vol. 43, p. 189-205.
10. SAAVEDRA, O.; SOTOLONGO, G.; GUZMÁN, M.V. Medición de la producción científica en América Latina en el campo agrícola y afines: un estudio bibliométrico. Revista Española de Documentación Científica. 2002; Vol. 25, p 151-161.
11. SAAVEDRA, O; SOTOLONGO, G., GUZMÁN, M.V. Mapeo autoorganizado de las revistas científicas y técnicas de América Latina y el Caribe. Aprobado. ACIMED.
12. SOTOLONGO, G; GUZMÁN, M. V.; GARCÍA, I.; SANZ, I. Retos de la bibliometría: la vigilancia y evaluación de la actividad científico y tecnológica. *Reencuentros,* 1998, vol. 21, p. 39-44.
13. GUZMÁN, M.V., SOTOLONGO, G. Mapas Tecnológicos para la Estrategia Empresarial. El caso de la *Neisseria meningitidis*. Aceptado. ACIMED.
14. SOTOLONGO, G.; GUZMÁN, M.V. La vigilancia tecnológica y la gestión de activos intelectuales. Opciones, 2002, año 9, no. 50, p. 2.
15. SALGADO, D. *Sistema de Vigilancia Científico – Tecnológico. Aplicación en el Instituto de Investigaciones de la Industria Alimentaría*. [Tesis de Master]. La Habana; Universidad de la Habana, 2002. Tutor: Maria Victoria Guzmán.
16. GARCÍA URIARTE, A. *La soja en la alimentación humana. Experiencia cubana*. [Tesis doctoral]. Valencia, España: Universidad Politécnica de Valencia; 1998. Tutor: Pedro Fito Maupoey.