NeuroISOC: un modelo de red neuronal para la representación del conocimiento

Félix de Moya Anegón. *Universidad de Granada*
Purificación Moscoso. *Universidad de Alcalá de Henares*
Carlos Olmeda y Virginia Ortiz-Repiso. *Universidad Carlos III*
Victor Herrero y Vicente Guerrero. *Universidad de Extremadura*

Abstract:

El propósito de esta ponencia es presentar un modelo de red neuronal que se ha desarrollado con el fin de representar el conocimiento expresado a través de la producción científica en el campo de las Ciencias Sociales y las Humanidades. Dicho modelo se ha aplicado al caso concreto de la base de datos ISOC, producida y distribuida por el Consejo Superior de Investigaciones Científicas. Esta aplicación forma parte de un proyecto de investigación cuyo objetivo principal es el desarrollo de una interfaz de realidad virtual basada en motores de clasificación que utilizan técnicas multivariantes o redes neuronales para posibilitar el acceso mediante *browsing* a los registros contenidos en una base de datos. Con el fin de representar las relaciones existentes entre las distintas materias que conforman el área de las Ciencias Sociales y las Humanidades, se han formado conjuntos de documentos a partir de los códigos de clasificación utilizados en la base de datos ISOC. Dichas relaciones se representan mediante matrices de coocurrencia de números de clasificación. Las matrices se forman siguiendo la estructura jerárquica de la propia clasificación. Estas matrices, una vez normalizadas, constituyen la entrada de un proceso de red neuronal que se basa en los mapas autoorganizativos de Kohonen (SOM). De las distintas salidas que produce el simulador de la red neuronal se utiliza la matriz de tasas de activación como entrada de una aplicación ad hoc que genera los mapas cuyas topologías representan el conocimiento extraído de la base de datos. El resultado de la aplicación de la metodología descrita es un árbol de mapas que permite al usuario navegar a través del conocimiento extraído de la base de datos. De esta forma, se genera una interfaz que expresa la topología, entendida como conjunto de vecindades, de las distintas categorías temáticas codificadas en esta base de datos. El resultado final es un árbol de mapas sensibles que se convierte, a su vez, en un sistema de browsing a través del cual el usuario tiene la posibilidad de acceder a los registros de la base de datos ISOC siguiendo una estructura jerárquica de categorías representadas gráficamente.

1. Introducción

El propósito de este trabajo es presentar una interfaz virtual basada en un modelo de red neuronal para la base de datos española ISOC.

El diseño de esta interfaz forma parte de un proyecto de investigación de amplia cobertura cuyo objetivo principal es experimentar en materia de técnicas de análisis multivariante, modelos de redes neuronales y metáforas espaciales, con el fin de analizar sus posibles aplicaciones en el desarrollo de interfaces de realidad virtual para el acceso a información electrónica (1).

Las investigaciones sobre recuperación de información electrónica han sido motivo de preocupación por numerosos expertos, cuyas publicaciones han venido a formar uno de los frentes de investigación con más relevancia en nuestra disciplina (2). Desde que la tecnología permitió desarrollar sistemas de almacenamiento y recuperación de información, los trabajos en esta materia se han centrado en encontrar nuevos medios que faciliten el acceso a la información.

El crecimiento de los recursos electrónicos y el auge del www han tenido dos consecuencias fundamentales. Por un lado, cada vez son más los usuarios de muy diversa procedencia que se enfrentan a los problemas derivados de la búsqueda de información electrónica, tratando de obtener los mejores resultados de las herramientas de búsqueda en el web, en los catálogos bibliotecarios o en una biblioteca digital, por ejemplo.

Por otra parte, bajo el mismo estímulo, profesionales ajenos al campo de la documentación hace tan solo unas décadas han comenzado a analizar los distintos

aspectos relativos a la recuperación de información desde sus propias perspectivas. Así, por ejemplo, la lingüística computacional, la inteligencia artificial o la ciencia cognitiva, han pasado a formar parte de las disciplinas involucradas en nuestra área. Sus trabajos han puesto de manifiesto la necesidad de desarrollar nuevas formas de representación del conocimiento, de naturaleza global, relacional, interactiva y multidimensional.

El avance de la tecnología está modificando, sustancialmente, las formas de concebir los diferentes procesos de recuperación de información. Ha cambiado el planteamiento y se han creado nuevos medios para llevarlo a cabo. El desarrollo de máquinas de cada vez mayor capacidad para almacenar, procesar, visualizar y compartir información; el desarrollo de softwares capaces de integrar múltiples aplicaciones; el desarrollo de nuevos algoritmos de recuperación, así como de interfaces gráficas centradas en el usuario, han hecho evolucionar, de un modo radical, los entornos electrónicos.

El trabajo que ahora presentamos se enmarca dentro de las investigaciones relacionadas con el desarrollo de nuevos algoritmos de recuperación, así como con el de nuevas interfaces gráficas de usuario.

2. La representación del conocimiento y las técnicas de reducción dimensional

Hasta la fecha se han aplicado diferentes métodos que sirven para generar espacios de documentos mediante el uso de algoritmos que permiten reducir el espacio vectorial. De entre ellos, los basados en análisis estadísticos multivariante y los de naturaleza conexionista son los más utilizados en nuestro campo.

Los primeros son básicamente tres: análisis de *cluster*, análisis de componentes principales (en su denominación anglosajona PCA, Principal Component Analysis) y el escalamiento multidimensional (MDS: Multidimensional Scaling) (3). El análisis de cluster permite crear, a partir de objetos representados en varias dimensiones, una representación en dos dimensiones en la que las relaciones entre los objetos se expresan mediante los valores de las matrices. El análisis de componentes principales parte de la premisa básica según la cual las relaciones lineales entre dos variables cualquiera se expresa mejor mediante una línea de regresión. Por último, el escalamiento multidimensional es una técnica de análisis multivariante que posibilita representar la similitud y las diferencias entre las distintas variables. Todos estos métodos permiten generar mapas de representación del conocimiento y, por consiguiente, desarrollar interfaces gráficas de usuario.

Entre los modelos conexionistas, uno de los principales es el de red neuronal de Kohonen (4). Este modelo se basa en el principio de agrupación y autoorganización de vectores de n dimensiones a espacios bidimensionales. Se utiliza para reducir las dimensiones de diferentes espacios de documentos de distinta naturaleza. El modelo de red neuronal de Kohonen ha sido el que se ha utilizado en el desarrollo de la interfaz NeuroIsoc.

3. Metodología

NeuroIsoc es una interfaz que representa las relaciones existentes entre las distintas materias de los documentos contenidos en la base de datos ISOC.

La base de datos ISOC es una base de datos referencial bibliográfica producida y distribuida por el Centro de Información y Documentación Científica del Consejo Superior de Investigaciones Científicas de España. Recoge las referencias de artículos de revistas españolas especializadas en Ciencias Sociales y Humanidades. También, de forma selectiva, recoge, entre otras, series monográficas y congresos. El volumen aproximado de esta base de datos es de 300.000 registros bibliográficos, y su cobertura temporal abarca desde 1975 hasta la actualidad, con una actualización diaria.

La base de datos ISOC se estructura en nueve subficheros que se corresponden, a su vez, con nueve categorías temáticas: América Latina; Economía, Sociología y Política; Historia, Arqueología y Prehistoria; Bellas Artes; Documentación Científica; Derecho; Lingüística y Literatura; Sicología y Educación; y Geografía y Urbanismo.

Para el desarrollo de *NeuroIsoc* se ha utilizado tanto el CD-ROM como el acceso en línea a través de la pasarela web recientemente creada.

Los conjuntos de documentos se han formado a partir de los códigos de clasificación utilizados en la base de datos ISOC. La clasificación de esta base de datos se estructura en diecisiete jerarquías temáticas, que son las siguientes: Antropología, arqueología y prehistoria; Bellas Artes; Información y documentación científica; Derecho; Economía; Educación; Filosofía; Geografía; Historia; Lingüística; Literatura; Sicología; Ciencias políticas; Sociología; Urbanismo; América Latina. Este sistema de clasificación se corresponde con una clasificación sistemática y jerárquica. Cada número consta de seis dígitos que representan a su vez los diferentes niveles jerárquicos. El prototipo desarrollado abarca todas estas áreas, a excepción de la relativa a América Latina.

Para el primer nivel se generó una matriz que representa las relaciones existentes entre las dieciséis jerarquías. Las relaciones entre los números de clasificación; esto es, entre los documentos clasificados con números de clasificación pertenecientes a más de una jerarquía, se han vectorizado mediante matrices de co-ocurrencia. Como ya se ha explicado, dichas matrices se han formado siguiendo la estructura jerárquica de la propia clasificación. Para generar los vectores correspondientes representados por las matrices, y con el fin de obtener los conjuntos de documentos, se han realizado las búsquedas necesarias por cada nivel de clasificación temática.

Posteriormente se generaron las matrices para el segundo nivel de cada una de las jerarquías, con el fin de obtener el vector correspondiente a la co-ocurrencia de, por ejemplo, el 050100 y el 050200, 050300, 050400, etc. Se ha descendido hasta el tercer nivel en las categorías relativas a Geografía, Sicología y Arqueología.

Las matrices resultantes se normalizaron con el objeto de eliminar las diferencias de escala así como los valores nulos. Una vez normalizadas pasaron a constituir las entradas de un proceso de red neuronal basado en una variante de los mapas autoorganizativos de Kohonen (SOM: Self-Organizing Maps) (5) desarrollado para el proyecto E.T. (Entertainment Thesaurus) (6). El mapa resultante de la aplicación del algoritmo expresa, en forma de áreas rectangulares, la co-ocurrencia de los números de clasificación de las diferentes materias, y, por consiguiente, las relaciones de similitud y

diferencia entre los documentos indizados en la base de datos. El tamaño, la forma y la vecindad de las áreas representan estas relaciones.

Para la presentación de los mapas resultantes, en forma de manchas y colores, se ha utilizado el modelo desarrollado por Xia Lin (7).

4. Resultados

El prototipo resultante de la aplicación de la metodología descrita es un árbol de mapas que permite al usuario navegar a través del conocimiento extraído de la base de datos ISOC. Es importante enfatizar el hecho de que los mapas generados son el reflejo del sistema de clasificación del conocimiento utilizado para esta base de datos concreta. Recordamos, a este respecto, que la indización y la clasificación de los documentos de cada base de datos responde, fundamentalmente, a las necesidades de sus usuarios potenciales, así como a la naturaleza de los documentos. Hay que tener en cuenta, por consiguiente, el carácter multidisciplinar de la base de datos en cuestión, así como las características específicas del lenguaje de indización y clasificación de los documentos de las bases de datos de Ciencias Sociales (8).

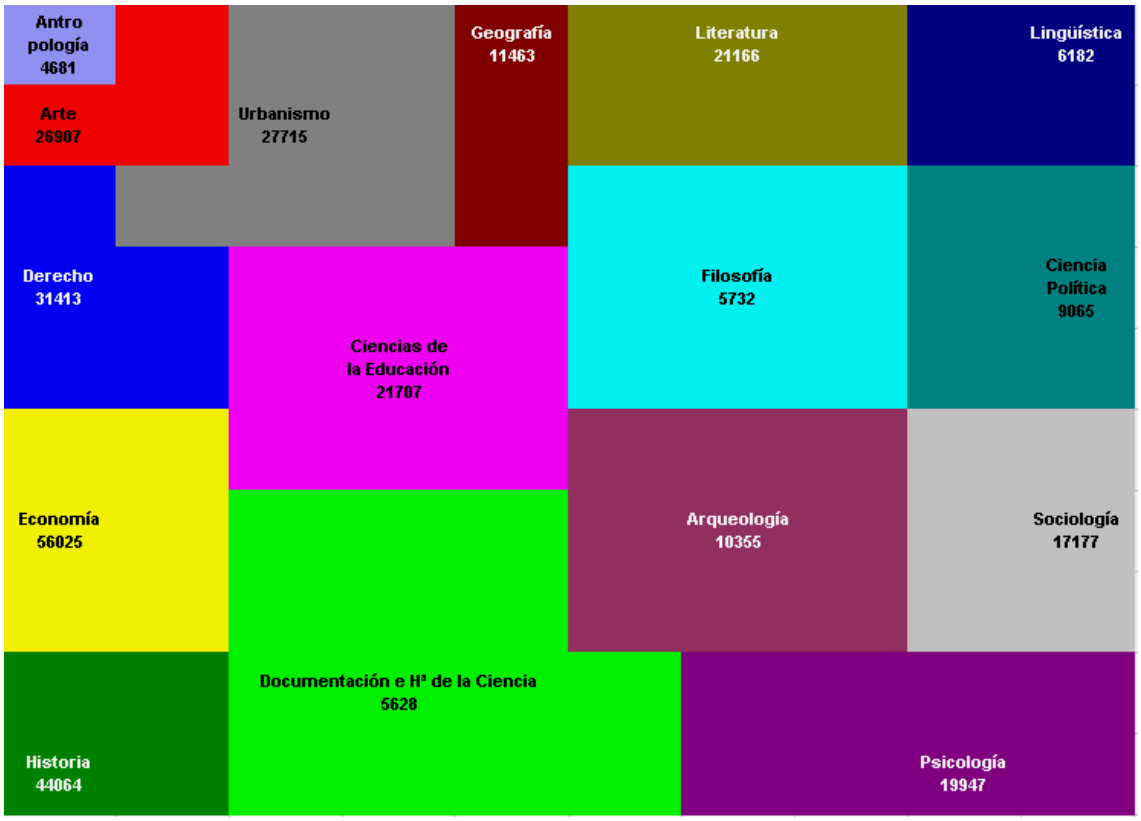

Figura 1. Primer nivel

La interfaz gráfica generada expresa la topología, entendida como conjunto de vecindades, de las distintas categorías temáticas codificadas en la base de datos. El resultado es un árbol de mapas sensibles que se convierte, a su vez, en un sistema de browsing a través del cual el usuario tiene la posibilidad de acceder a los registros de la base de datos siguiento una estructura jerárquica de categorías representadas gráficamente. El prototipo desarrollado representa hasta el segundo nivel de todas las

categorías, y, en algunas de ellas, hasta el tercero. Con el fin de facilitar al usuario la comprensión de los mapas, sobre cada área temática aparece una etiqueta identificativa de su contenido y del número de documentos asociados con el número de clasificación correspondiente.

La figura 1 muestra el mapa generado para el primer nivel de clasificación. Las áreas del mapa se corresponden con las dieciséis categorías de la clasificación ya mencionadas. Estas áreas, determinadas por las matrices de co-ocurrencia generadas, visualizan el conocimiento contenido en la base de datos ISOC.

El tamaño de las áreas representa la amplitud del concepto, o, lo que es lo mismo en este caso, la frecuencia con la que el número de clasificación pertinente co-ocurre con otras clases en la base de datos en su totalidad. De esta forma, tamaño del área y vaguedad del concepto a efectos de recuperación de los documentos se convierten en sinónimos. La topología del conjunto expresa la totalidad del contenido de la base de datos, contenido que se entiende como el contacto entre las distintas áreas temáticas. De forma que cuantas más relaciones se establecen entre un área y las restantes, mayor es el tamaño del área en cuestión. Así, por ejemplo, el área correspondiente a Documentación e Historia de la Ciencia es la mayor de todas puesto que es la que más relaciones mantiene con el resto de las áreas. Es decir, su número de clasificación es el que más veces co-ocurre en relación con los número de clasificación relativos a las otras categorías temáticas. Por el contrario, el menor tamaño del área de Antropología se debe a que su número de clasificación co-ocurre en menor medida.

Las relaciones de vecindad entre las áreas indican la frecuencia de las co-ocurrencias de los números de clasificación. A este respecto hay que señalar que la red neuronal busca la topología óptima. Esto implica que al tener que reducir a dos dimensiones la representación, las áreas se despliegan y ocupan su lugar en función del mayor o menor contacto entre ellas, por lo que las relaciones de co-ocurrencia que se establecen entre dos áreas condicionan la ubicación en el mapa del resto. La cercanía/distancia entre las áreas viene a determinar la frecuencia de co-ocurrencias, sin que ello signifique que los números de clasificación de dos áreas alejadas físicamente entre sí no co-ocurran de forma absoluta. Así, por ejemplo, según se muestra en la figura 1, el área de Urbanismo está más relacionada con el Arte, el Derecho, la Geografía y la Educación con la Lingüística.

Con respecto a la forma de las áreas, éstas se encuentran también determinadas por la co-ocurrencia de sus números de clasificación con el resto de las áreas. Las relaciones de co-ocurrencia no siempre se pueden representar mediante áreas puramente rectangulares, sino que es necesario recurrir a áreas de más lados que permitan visualizar el contacto entre las áreas, como ocurre, en el caso del mapa del primer nivel (figura 1) con el área de Arte.

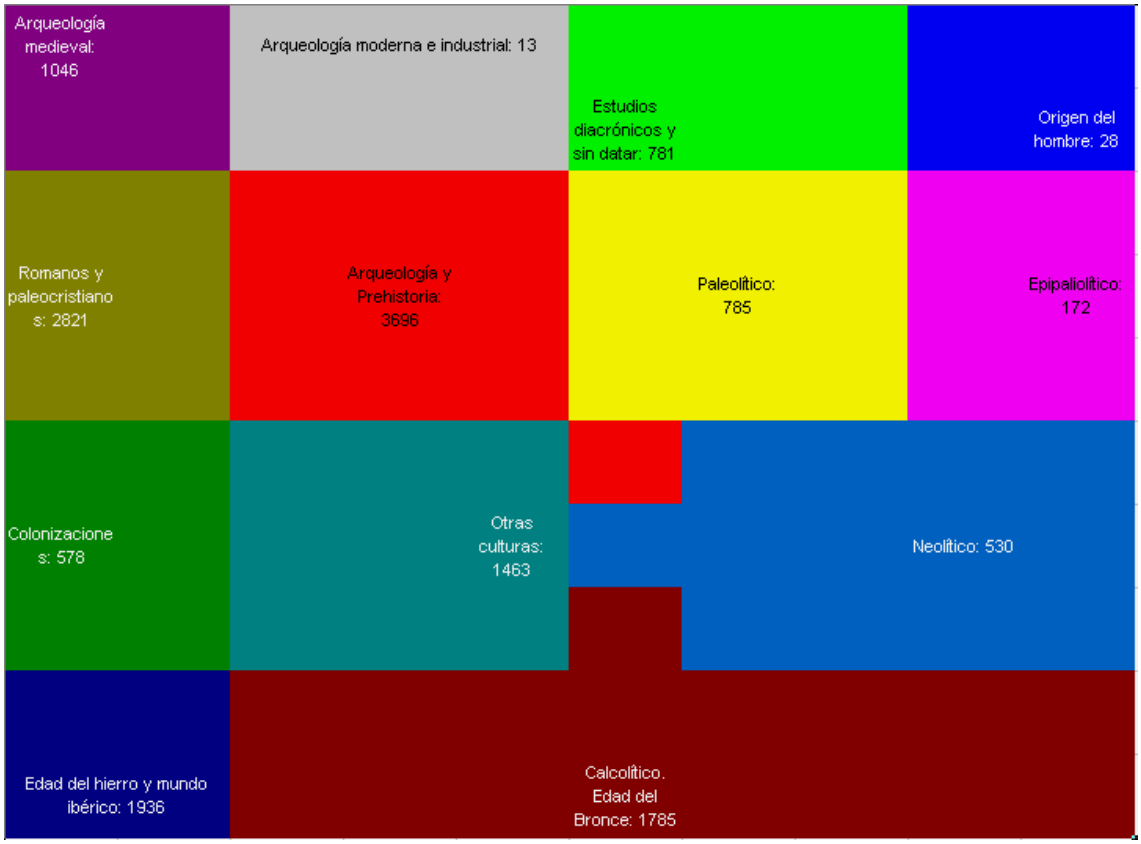

Figura 2. Mapa correspondiente al área de Arqueología

La interpretación del mapa, comenzando por el ángulo superior derecho, a partir de Estudios diacrónicos sin datar, y en el sentido de las agujas del reloj, permite diferenciar de forma ordenada cronológicamente, los diferentes periodos históricos: Origen del hombre; Paleolítico; Epipaleolítico; etc., para terminar en la Arqueología moderna e industrial.

La extensión, por ejemplo, de la clasificación correspondiente a Calcolítico y edad del bronce indica la frecuencia de co-ocurrencias con los números de clasificación relativos al Neolítico, Otras culturas y Edad del hierro.

Las áreas que representan a los documentos clasificados con Arqueología y Prehistoria, y Otras culturas se sitúan en una posición central, ya que son las áreas que mayor número de co-ocurrencias tienen con el resto de las áreas. Por el contrario, la posición aislada y esquinada del área Origen del hombre viene determinada porque sólo presenta mínimas co-ocurrencias con otras áreas.

5. Conclusiones

NeuroIsoc es una interfaz gráfica de usuario que representa un nuevo concepto de visualizar el conocimiento recogido en los documentos de una base de datos bibliográfica referencial, y por consiguiente, modifica sustancialmente la forma en la que el usuario lo recupera.

Esta nueva forma de representación y visualización del conocimiento influye substancialmente en la manera en la que el usuario se enfrenta al proceso de búsqueda y recuperación de la información.

El algoritmo utilizado para el desarrollo de la red neuronal demuestra que es posible tratar grandes volúmenes de información. Asimismo, el uso de este algoritmo posibilita obtener interfaces intuitivas y gráficas.

La diferencia fundamental con respecto a otros tipos de interfaces radica en el hecho de que lo que se representa son relaciones entre conceptos, lo que, sin lugar a dudas, enriquece las perspectivas desde las que el usuario afronta el proceso de acceso a la información.

En definitiva, el fin último de la investigacione llevada a cabo es desarrollar un modelo de representación del conocimiento que se asemeje a la forma en la que opera la mente humana, desterrando los viejos modelos lineales, potenciando el sistema de browsing y posibilitando la interactividad y las relaciones multidimensionales en la recuperación de información.

Referencias

(1) *Interfaz de realidad virtual para el acceso a información electrónica (IRVAE).* Proyecto de investigación subvencionado por la Comisión Interministerial de Ciencia y Tecnología. Plan Nacional, CICYT. Código TEL 97-1131. Investigador principal: Félix de Moya Anegón.

(2) MARCHIONINI, G. *Information Seeking in Electronic Environment*. Cambridge: University press, 1997.

(3) MOYA, F.; HERRERO, V.; GUERRERO, G. Virtual reality interface for accessing electronic information. *Library and Information Research News*, 1998, vol. 22, nº 71, p. 34-39.

(4) MOYA, F.; HERRERO, V.; GUERRERO, G. La aplicación de redes neuronales artificiales (RNA) a la recuperación de la información. *Anuario SOCADI De Documentación e Información*, 1998, no. 2: 147-164.

(5) KOHONEN, T. *Self Organizing Maps*. 2nd edition. Berlín: Springer, 1997.

(6) CHEN, H. Et al. Internet browsing and searching: user evaluations of category map and concept space techniques. Journal of the *American Society for Information Science*, 1998, vol. 49, nº 7, p. 582-603.

(7) LIN, X. Maps displays for information retrieval. *Journal of the American Society for Information Science*, 1997, vol. 48, n 1, p. 40-54.

(8) EXTREMEÑO, A.; MOSCOSO, P. El control de calidad en las bases de datos de ciencias sociales. *Boletín de la ANABAD*, 1998, vol. XLVIII, nº 1, p. 231-253.