REDES NEURONALES ARTIFICIALES APLICADAS A LA VISUALIZACIÓN DE LA INFORMACIÓN

Vicente P. Guerrero Bote
Facultad de Biblioteconomía y Documentación de la UEX.

El aumento exponencial de la información al que se viene haciendo referencia desde hace veinticinco años y al que contribuyen en gran medida el proceso de digitalización, transformación de documentos *"basados en átomos a los basados en bits"*, que se está llevando a cabo, el coste cada vez menor de los medios de almacenamiento y la distribución de información mediante las llamadas *autopistas de la información*, nos sitúa dentro de un universo en desarrollo de información electrónica que puede ser manipulada por medios automáticos. Para acceder a tal cantidad de información se están ideando distintos mecanismos. Algunas redes neuronales artificiales permiten la organización topológica de documentos.

La *Information Retrieval*, en castellano la *Recuperación de la Información* (a partir de ahora R.I.), estudia la forma en que tienen que ser almacenadas grandes cantidades de datos para que puedan ser recuperadas en el momento en que se desee. Actualmente, la R.I. está tomando un mayor auge debido, entre otras cosas, al incremento extraordinario de la información disponible, sin embargo, no es una disciplina de reciente creación, sino que posee una larga existencia, a pesar de que han cambiado los medios y la forma de trabajar. Mientras es posible encontrar trabajos iniciales de esta disciplina que prestan una especial atención a los aspectos básicos de la búsqueda de información, hoy en día la mayoría de las investigaciones se centran en los distintos algoritmos y representaciones documentales que posibilitan la búsqueda de información de forma automatizada sobre soportes digitales.

En cuanto a las tecnologías, como igualmente dice Kantor, la recuperación de la información es una encrucijada de "*caminos tecnológicos*". Así, con el desarrollo de la Documentación que ha entrado en campos como la Ofimática, identificación dactilar, gestión de imágenes médicas, bases de conocimiento, gestión multimedia, etc., la recuperación de información lo ha hecho en el procesamiento del lenguaje natural, la inteligencia artificial, etc.

En el depósito de una biblioteca tradicional, una monografía puede estar ordenada por contenido (a través de una de sus materias). Para compensar esta pérdida de accesibilidad se mantiene un conjunto de catálogos (de autor, título, materias, topográfico, etc.). No obstante, para poder aprovechar la potencia de estas herramientas es necesario que los usuarios conozcan su funcionamiento así como el lenguaje empleado por los bibliotecarios.

Con los ordenadores llegó la posibilidad de indizar el texto completo, lo cual hace posible búsquedas a través de los descriptores y el texto libre (del título, resumen u otros campos). Como consecuencia de ello surgieron los OPACs que gestionan las referencias bibliográficas de las bibliotecas,

♦♦

Con el desarrollo de la Documentación que ha entrado en campos como la Ofimática, identificación dactilar, gestión de imágenes médicas, bases de conocimientos, gestión multimedia, etc., la recuperación de información lo ha hecho en el procesamiento del lenguaje natural, la inteligencia artificial, etc.

♦♦

y en los cuales ya podemos buscar *historia* como un descriptor controlado o bien como una palabra del título o incluso de alguna nota.

El papel de la R.I., por tanto, es gestionar estas inmensas cantidades de datos para poderlas facilitar a los usuarios como respuesta a sus necesidades de información. El resultado ideal sería una ordenación de los documentos por utilidad o relevancia ante una determinada petición, y que esta fuera similar a la que hicieran los usuarios si pudieran examinar todos los documentos disponibles. Sin embargo, la relevancia es muy subjetiva y desde el primer momento ha sido una fuente de conflictos dentro de la R.I., lo que hace que el orden ideal no sea igual para todos los usuarios.

Ese objetivo se ha de lograr en el marco de lo que se conoce como Sistema de Recuperación de Información que tradicionalmente se ha considerado compuesto por:

- La información almacenada.
- El motor de búsqueda.
- El interfaz con el usuario.

Donde el *interfaz de usuario* es el encargado de gestionar la comunicación con el usuario, recibe las ordenes y transmite los resultados. La *información* no sólo está en el formato original, sino que se han generado una serie de herramientas y representaciones necesarias, para que el *motor de búsqueda* pueda aplicar una o varias técnicas de R.I. en el cálculo de la relevancia de cada documento. Siendo las *técnicas de R.I.* algoritmos que especifican la comparación que se ha de realizar entre la representación de la necesidad de información y la representación de cada documento en el cómputo de su relevancia.

Algunas técnicas de recuperación llevan a cabo una simple comparación aisladamente entre las representaciones documentales y de la necesidad de información. Otras técnicas como dice Belkin en primer lugar realizan una organización global de los documentos de la base, y en función de ella se lleva a cabo la comparación anterior o simplemente se permite la navegación, el *browsing* por una base organizada.

Actualmente resulta novedosa la aplicación de las redes neuronales artificiales a tal fin. Un tipo particular de ellas, las redes competitivas, se caracteriza por realizar un clustering de los patrones de entrada. Los mapas autoorganizativos de Kohonen, como consecuencia de la interacción lateral existente en la capa competitiva, tienen la particularidad de que, además de hacer el clustering, como todas las redes competitivas, organizan topológicamente los clusters resultantes en una rejilla bidimensional. Para ello en primer lugar se tienen que representar los documentos como patrones que se puedan clasificar. Nosotros aquí vamos a mostrar algunas de las aplicaciones disponibles.

VISUALMAPS DE XIA LIN.

Xia Lin, investigador de la Universidad de Kentucky utiliza estos mapas con el fin de generar una salida *visualizable* de una determinada colección de documentos.

La representación documental la obtiene de la siguiente forma:

1.- Construcción de una lista que incluya todos los términos que aparecen en los títulos y resúmenes de todos los documentos de la colección.
2.- Supresión de términos irrelevantes del lenguaje a partir de una lista de palabras vacías.
3.- Transformación de los términos en la raíz mediante un algoritmo de stemming para reducir la lista.
4.- Eliminación de términos de frecuencias altas y bajas. Como dijo Luhn la significación de un texto está depositada sobre las palabras de frecuencias intermedias.
5.- Construcción de un vector para cada documento con tantas componentes como términos han quedado en la lista. Las componentes se gene-

ran de tres formas diferentes (dependiendo de la que se escoja se obtienen distintas visiones de la base):

a) Dígitos binarios (uno si el término correspondiente está en el documento y cero si no está)
b) Pesos proporcionales a la frecuencia del término en el documento.
c) Pesos proporcionales a la frecuencia del término en el documento e inversamente proporcionales al número de documentos en los que aparece el término.

Terminado el entrenamiento se le asigna a cada neurona el término más cercano a su vector de pesos. Uniendo las neuronas cuyo término asignado es el mismo tenemos la rejilla dividida en distintas zonas (que se pueden etiquetar con el término en cuestión). De esta forma finalmente tenemos los documentos clasificados en una superficie dividida en una serie de zonas etiquetadas por términos.

Lin utiliza esta clasificación como interfaz para el browsing de bases documentales pequeñas, que pueden incluso ser los resultados de una búsqueda realizada. Proporciona prototipos que puede ser consultados en su propia página personal [http://www.ukv.edu/~xLin/]. Una vez etiquetadas las distintas zonas de la base, se puede seleccionar aquella que resulte de nuestro interés. Dichos prototipos están realizados en *java*, de forma que incorpora, además, *dos controles*, barras de desplazamiento:

- La horizontal regula la aparición de puntos que representan a los documentos. Estos puntos aparecen en el lugar que le corresponde a cada documento dentro de la rejilla. Aparecen inicialmente aquellos documentos que logran una mayor activación de la neurona en la que se encuentran clasificados. A medida que se desplaza la barra disminuye el umbral que necesitan superar los documentos para aparecer, de modo que aparece un mayor número.
- La barra de desplazamiento vertical tiene la misma función, pero, en este caso para los términos. Es decir, a medida que se desplaza diminuye el umbral de activación que tienen que superar para aparecer en el mapa. Con ello aumenta el número de los que aparecen, dejando así el mapa más etiquetado.

La forma de proceder para Lin sería:

- Comenzar con muy pocos términos en el mapa, de forma que sea fácil localizar aquellos que describan mejor nuestras necesidades.
- A medida que vayamos necesitando términos más específicos podemos *ir aumentando el número de estos.*
- Con el otro control podemos ver también el número de documentos presentes en la zona para ampliar o reducir la recuperación.
- Si no se encuentran documentos que satisfagan nuestras necesidades de información se puede ampliar la búsqueda en la dirección conveniente.

Con el *ratón* se puede *seleccionar* la zona que se quiera, apareciendo en ese caso una ventana con enlaces que contienen el título, a los distintos documentos.

Una imagen de este interfaz correspondiente a una base generada con todos los documentos clasificados en la categoría Space Science, la podemos ver en la figura 1.

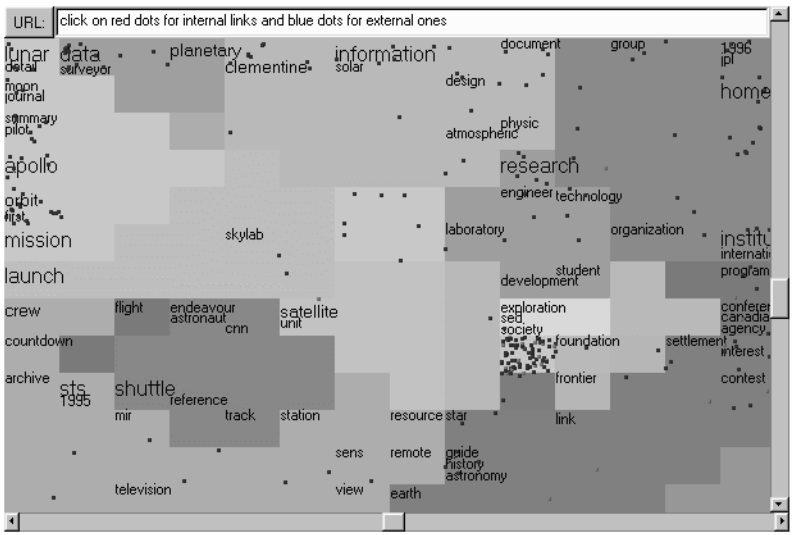

Figura 1: Visual Sitemap realizado por Xia Lin de los documentos pertenecientes a la categoría Space Science de Yahoo [http://lislin.gws.uky.edu/Sitemap/spaceSmall.htm]

ET-MAP

El profesor Chen en el Laboratorio de Inteligencia Artificial de la Universidad de Arizona supervisa la realización de un proyecto llamado ET-Map. En él se ha hecho un proceso similar, de todos documentos pertenecientes a la subcategoría de "Entertainment" del índice Yahoo. Está disponible en Internet [http://ai2.BPA. arizona.edu/ent].

En este caso la interfaz es un mapa sensitivo donde se pueden observar las distintas regiones de la rejilla (generadas de la misma forma que en el caso anterior). En ellas solamente se indica el término ganador en toda la región, así como el número de elementos que contiene, no pudiendo así aumentar el número de términos presentes como permite el interfaz de Lin.

Este mapa tiene dos niveles, es decir, al seleccionar una región aparece otro mapa de los documentos pertenecientes a la misma. Si las regiones resultantes contienen un gran número de documentos se genera otro mapa. Cuando se accede a una región del mapa correspondiente al último nivel aparecen los enlaces a los distintos documentos.

Han realizado experimentos de browsing comparándolo con el correspondiente índice humano de Yahoo. Los resultados son muy parecidos si no se va buscando nada en particular, siendo peores si se busca algo específico.

Podemos ver una imagen del interfaz en la figura 2.

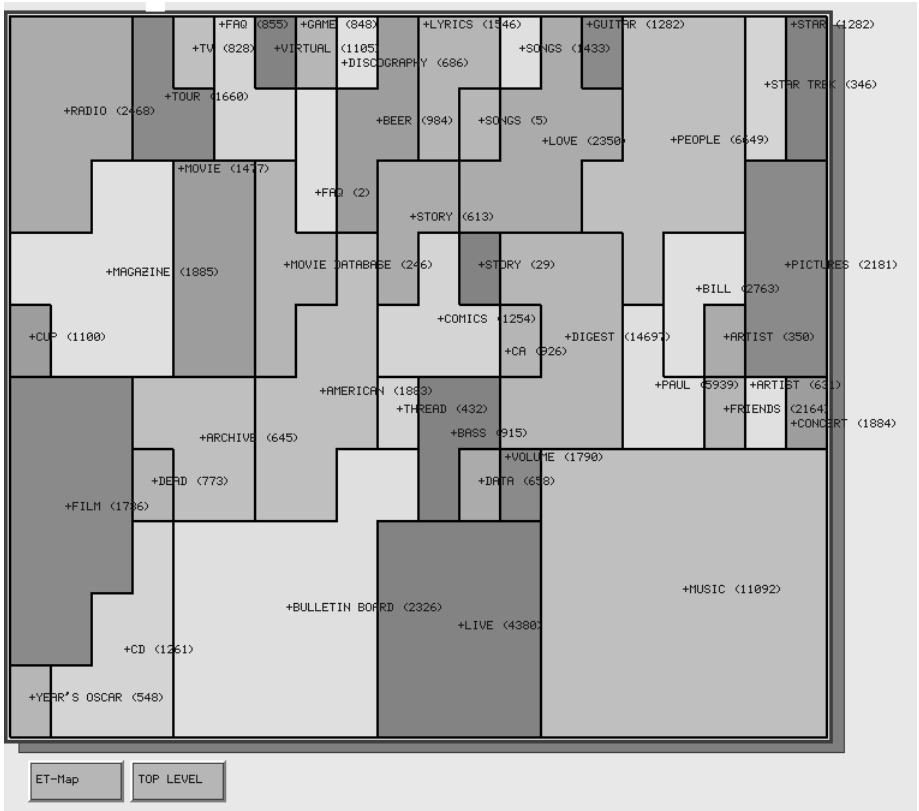

Figura 2: Mapa sensitivo general realizado en el proyecto ET-Map bajo la dirección del profesor Chen [http://ai2.bpa.arizona.edu/ent/entertain1/].

WEBSOM

El mismo Teuvo Kohonen dirige un grupo finlandés perteneciente al Centro de Investigación en Redes Neuronales de la Universidad Tecnológica de Helsinki que está utilizando este tipo de redes tanto para hacer clasificaciones de términos como de documentos.

El *Word Category Map* (mapa de categorías de palabras) se forma clasificando los términos con una red de Kohonen. Lo más innovador con respecto a otras aplicaciones similares es la forma de representar las palabras incluyendo un pequeño contexto. En primer lugar, asignan una clave a cada término, compuesta por un vector de noventa componentes (que toman valores aleatorios entre 0 y 1). Y para formar la representación de un término unen tres claves, la del término que le antecede en el texto, la suya propia y la correspondiente al término que le sigue en el texto. Estos vectores de 270 componentes son los que se utilizan para entrenar una red de Kohonen. En la figura 3 se pueden observar unos resultados obtenidos para lengua inglesa que muestran una llamativa agrupación de los nombres en una zona, de los verbos en otra y de palabras estruc-

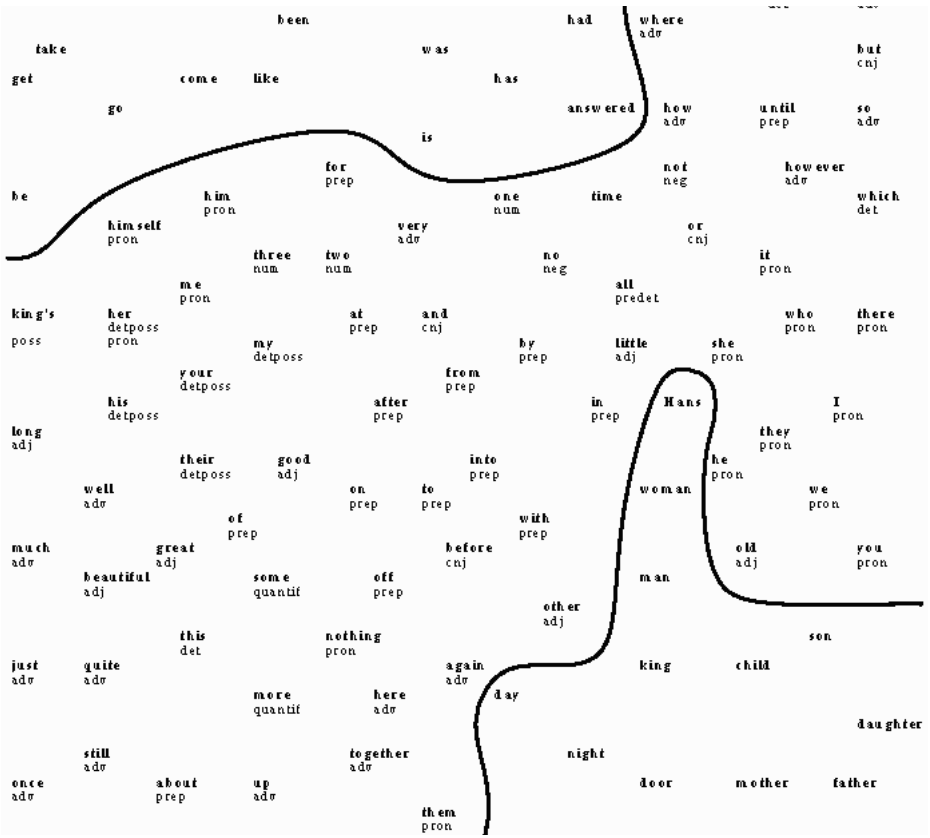

Figura 3: Visión Parcial del Mapa de Características de Palabras generado por el equipo de Kohonen a partir de una colección de cuentos de los hermanos Grimm.

turales en otra. Sin embargo, estos resultados no son extrapolables a otras lenguas, ya que en gran medida pueden deberse a la estructura del inglés.

Como he dicho anteriormente también han realizado un sistema de clasificación documental, que se puede utilizar como interfaz para acceder a la información. Este lo han denominado *WEBSOM*, y es un sistema pensado para clasificar un gran número de documentos, que lo han aplicado a Internet.

Para la representación de cada documento se siguen los siguientes pasos:

1.- Se eliminan las palabras de alta y baja frecuencia, con el fin de reducir el procesamiento computacional y eliminar ruido.
2.- Con estas palabras se genera *un mapa de categorías de palabras* (Word category map).
3.- Se construye un histograma para cada documento indicando el número de palabras de cada categoría que contiene.

La diferencia con respecto a los otros es que para aliviar el proceso de cálculo reduce el número de componentes de los vectores documentales representados en función de las categorías en lugar de las palabras.

Los vectores generados se utilizan también para entrenar una red de Kohonen que los organiza temáticamente en dos dimensiones. Tiene un interfaz gráfico bastante cuidado, en el que también se han incorporado etiquetas descriptivas generadas automáticamente (puede consultarse un prototipo del mismo en internet [http://websom.hut.fi/websom/]). En la figura 4 podemos ver el mapa general realizado para una colección de 4600 artículos de USENET pertenecientes al grupo de discusión de *comp.ai.neural-nets*. Tenemos que tener en cuenta que junto a este mapa se facilita un índice donde se especifica el significado de cada etiqueta.

En la figura 5 podemos ver un zoom de la esquina superior derecha, del mismo mapa.

♦♦

Resulta novedosa la aplicación de las redes neuronales artificiales a tal fin. Un tipo particular de ellas, las redes competitivas, se caracteriza por realizar un clustering de los patrones de entrada

♦♦

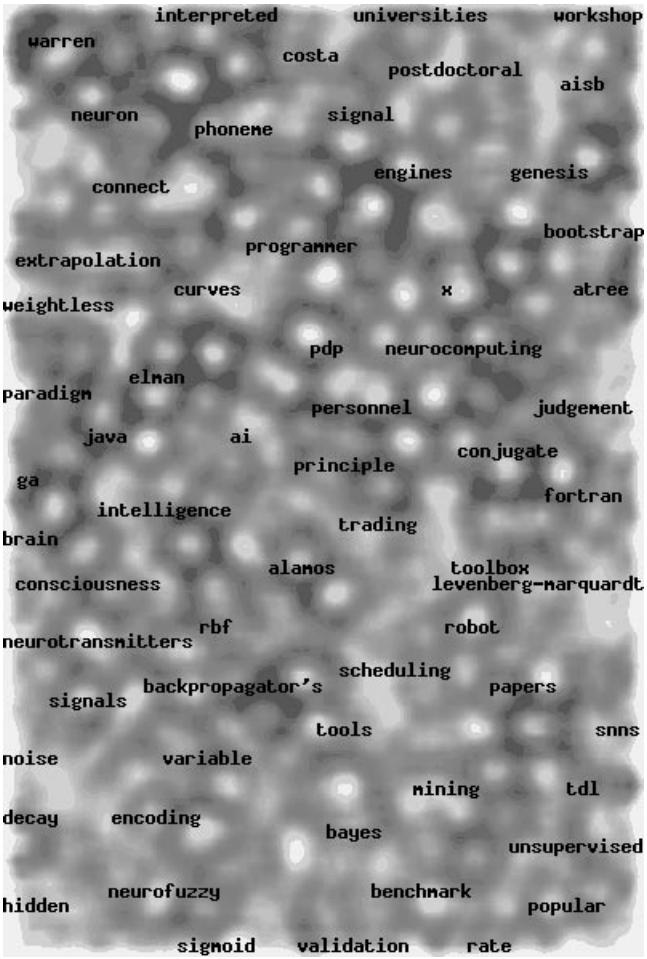

Figura 4: Mapa de primer nivel correspondiente al WEBSOM aplicado al grupo de discusión de usenet d comp.ai.neuronal-nets [http://websom.hut.fi/websom/comp.ai.neural-nets-new/html/root.html]

Mediante la intensidad del color, que se puede ver en las dos figuras, se indica las distancias entre los prototipos de los distintos nodos de cada zona. De este modo se puede interpretar de manera similar a un mapa topográfico:

- Una superficie rugosa donde los colores oscuros representan las montañas, donde las distancias entre las poblaciones son mayores, y donde estas tienen un menor número de habitantes.
- Los colores claros representarían los valles, zonas con menores impedimentos físicos (con ciudades más cercanas y pobladas).

Según el equipo de desarrollo es en los valles donde se pueden encontrar las discusiones más intensas.

En el zoom también se aprecian como pequeñas estrellas blancas los nodos existentes. Estas son las poblaciones donde se agrupan los habitantes (que son los documentos). Si se selecciona uno de estos nodos (o ciudades) se obtiene una página con enlaces a cada uno de los documentos (habitantes) ubicados en ella.

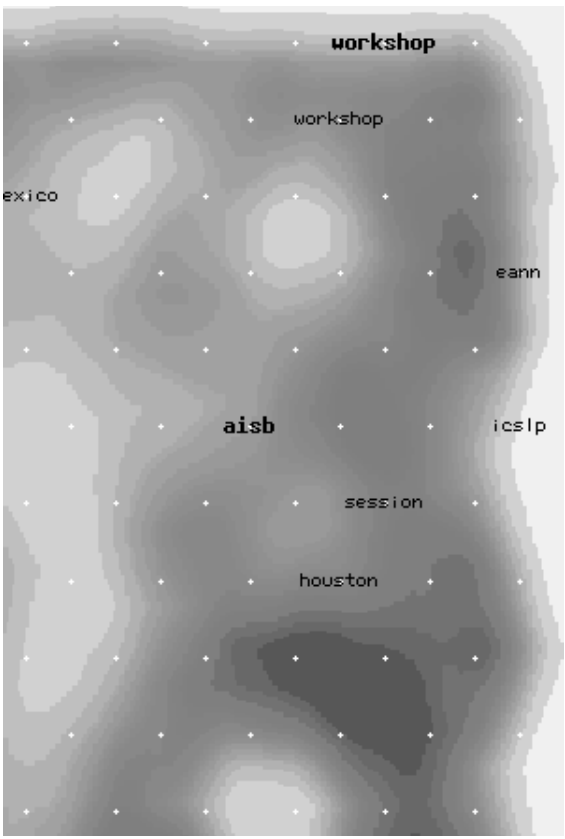

Figura 5: Zoom de la esquina superior derecha de la figura anterior.