CLASIFICACIÓN INTELIGENTE DE SITIOS WEBS USANDO REDES NEURONALES ARTIFICIALES

J. L. Castillo[1], J. R. Fernández del Castillo[2], L. González Sotos[3]. Departamento de Ciencias de la Computación, Escuela Politécnica. Universidad de Alcalá. Madrid, España.

Recibido enero 21 de 2014 – Aceptado abril 28 de 2014

http://dx.doi.org/10.18566/puente.v8n1.a09

***Resumen*— El creciente número de documentos web en Internet hace que los procesos de búsqueda sobre dichos documentos se dificulten cada día más. Los buscadores tradicionales no estructuran ni categorizan la información encontrada. Para solucionar estos problemas se hacen necesarios agentes web inteligentes. Dichos agentes estarán basados en potentes métodos de categorización que permitan organizar y clasificar los documentos de la forma más automatizada posible. El enfoque de computación neuronal es una de las alternativas más apropiada, debido a las características de adaptabilidad, generalización, potencialidad y robustez de las redes neuronales artificiales (RNA). Basándonos en documentos web obtenidos en buscadores tradicionales a partir de palabras claves relacionadas con "Alcalá", hemos analizado diferentes técnicas de representación de dichos documentos, y utilizando un modelo de RNA que implementa clusterización ciega hemos realizado una clasificación inteligente de sitios web de Alcalá, presentando un estudio de los resultados obtenidos. Proponemos un sistema de procesamiento que incrementa el control que el usuario puede tener sobre la información. Esta propuesta hará posible el análisis e intercambio de información de una forma rápida y fiable, tal que la misma podrá ser catalogada y preparada para futuros procesamientos.**

***Palabras claves*— Redes Neuronales Artificiales, Categorización, Agente Web, Inteligencia Artificial, Documentación.**

***Abstract* — the growing number of Internet web documents makes the search process is difficult on these documents every day. The traditional search engine do not structure or categorize information found. To solve these problems is necessary to use intelligent web agents. These agents will be based on powerful methods of categorization that could classify the documents in a way probably more automatic. The neural computation is one of the most appropriate alternatives, due to its characteristics of adaptability, generalization, power and reliability of artificial neural network. Based on web documents obtained in traditional searchers and starting with keywords related to "Alcalá", we have analysed several technics of documents representations and using a RNA model than implements blind clustering we have make an intelligent classification of these Alcalá web sites. We present a study of the results obtained. We proposed a processing system that increase the control of the user over the information. This propose will make possible the analysis and interchange of information in a very fast way and reliable. It will be categorized and treated for further processing.**

***Keywords*— Artificial Neural Network, Clustering, Artificial Intelligence, and Documentation.**

I. INTRODUCCION

En pocos años, la tecnología de la red de ámbito mundial —*World Wide Web* — se ha convertido en una herramienta universal para todo tipo de actividades culturales, profesionales y comerciales. Los avances tecnológicos de los últimos años han provocado un aumento exponencial de la información mediada por dicha red. El proceso de digitalización y la transformación de documentos que se está llevando a cabo son dos claros ejemplos de la revolución de la información, la cual ha permitido su acceso a un número ilimitado de usuarios. Actualmente, la localización de información en internet es un campo muy abordado por los investigadores y suele tener muchas dificultades, debido principalmente al gran número de documentos presentes en la web y a la poca estructuración semántica de los documentos [7] [8]. Entre las técnicas utilizadas por los usuarios de internet para localizar información podemos destacar principalmente tres: los buscadores, los canales de información y los agentes [2].

Los buscadores son servidores webs especializados en almacenar referencias de gran cantidad de páginas webs y otros documentos presentes en internet, siendo capaces de hacerlas disponibles a los usuarios mediante consultas guiadas por palabras claves y/o estructurando dichos documentos en una jerarquía de categorías. Los

Este trabajo es sometido al Congreso Iberoamericano Soporte al Conocimiento con la Tecnología- SOCOTE 2013.

[1] José Luis Castillo, Dr.en Informática, Área de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Alcalá. E-mail: jluis.castillo@uah.es.

[2] José Raúl Fernández del Castillo es Dr. en Ciencias Física, Área de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Alcalá. E-mail: joseraul.castillo@uah.es.

[3] León González Sotos es Dr. en Ciencias Matemática, Área de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Alcalá. E-mail: leon.gonzalez@uah.es.

canales de información y listas de distribución consiste básicamente en que una vez que un usuario se subscribe a uno de estos servicios, comienza a recibir periódicamente información sobre noticias, recursos, documentos y servicios presentes en internet.

Los agentes son sistemas que, con cierta información sobre el usuario, son capaces de entregarle enlaces de páginas webs que pueden resultar de su interés. Entre los agentes webs distinguiremos los que se les puede llamar inteligentes, porque hacen uso de técnicas de inteligencia artificial para mejorar su rendimiento (que serán los que trataremos principalmente en este artículo) y aquellos que no utilizan estas técnicas. La característica principal es que suelen almacenar una lista de consultas sobre temas de interés que lanzan sobre buscadores clásicos y de los resultados obtenidos realizan cribas sobre aquellas entradas que ya fueron presentadas al usuario, mostrando solo las novedosas. Todas estas técnicas presentan principalmente el problema de que necesitan un excesivo tiempo en la búsqueda de información, y muchas veces no se encuentra lo buscado, a pesar de existir en internet [4].

Habitualmente los buscadores devuelven un elevado número de enlaces, muchos de ellos con información que no tiene que ver con lo que el usuario en realidad busca. Muchos de los canales de información no se ajustan al perfil de lo que el usuario desea, lo que ocasiona que el usuario recibe mucha información irrelevante. Además, debido a la rapidez con la que se ponen al día la gran cantidad de sitios webs, para los más potentes buscadores les resulta muy difícil la reindexación de dichas páginas, con lo que habitualmente se producen resultados erróneos. En los buscadores que estructuran la información por categorías este problema es aún más grave, porque es muy costoso, la catalogación de los sitios webs de forma manual [1]. Mantener por tanto al día estos buscadores es arduo difícil, mientras que la clasificación automática no siempre da resultados satisfactorios. Con los métodos actuales de acceso a la información en internet, no somos capaces de trabajar a un nivel razonable con ella. Este desbordamiento que se produce en el manejo de la información viene en gran medida ocasionada por el hecho de que las páginas webs están diseñadas para ser entendidas por los humanos, no por los computadores, muchas veces hasta la estructuración semántica de los documentos es un aspecto totalmente olvidado, pese a existir estándares al respecto. Por lo tanto, extraer conocimiento automático sobre el contenido de un sitio web se hace muy complicado.

Es por ello, que el uso de sistemas con cualidades similares a los humanos, es una alternativa a tener en cuenta. Sistemas que permitan obtener el significado o al menos poder identificar documentos con significados parecidos dentro de internet, sistemas que sean capaces de navegar, leer y comprender dicha información. Estas y otras tareas son las que asignaremos a los agentes web inteligentes y constituye la razón primordial de éste artículo.

II. ESTADO DEL ARTE

En las siguientes secciones, comentaremos el estado del arte actual, sobre el tema que estamos tratando, empezando por la conceptualización de lo que constituye un agente inteligente, el papel que pueden desempeñar en un entorno web, las tareas que deben de cumplir y las tecnologías existentes para que lleven a cabo sus tareas, describiremos lo que constituye la catalogación de documentos, y los diferentes tipos de agrupación existentes.

A. Agentes Web Inteligentes

Se define como agente autónomo inteligente a un sistema *en* y *parte de un entorno* que *siente ese entorno* y *actúa sobre él*, a través del tiempo, persiguiendo sus propios objetivos. Un agente inteligente no solo actúa respondiendo a un usuario, sino que también puede tener sus propios objetivos y por lo tanto puede decidir en cualquier momento lanzar sus propias acciones para cumplirlos [2]. Por lo tanto, esto implica una cualidad del agente, que sea *persistente*, es decir, aunque el usuario no esté presente, el agente sigue funcionando, recolectando información, aprendiendo y comunicándose.

Las capacidades de comunicación de los agentes, no solo con los usuarios, sino también con otros agentes conduce a la creación de colonias de agentes, que asociados de esta forma logran beneficios entre todos, facilitándoles la tarea de lograr cada uno sus propios objetivos. En los agentes inteligentes, por tanto, la percepción, el procesamiento de la información es fundamental, pero además para incrementar su adaptabilidad se hace necesario la capacidad de modificar su comportamiento con la experiencia, es decir, la capacidad de aprendizaje.

Los agentes webs inteligentes estarán situados en la web [7][8], y por tanto han de tener capacidades para moverse en dicho medio, por ello, deberán:

- Navegar por la web, saltando entre enlaces.
- Lanzar consultas en los buscadores de páginas web.
- Introducirse en foros de discusión en busca de información.
- Movilizarse entre los ordenadores de la web, de forma que pueda ejecutarse en cualquier lado, siguiendo a su usuario o ubicándose en lugares donde le resulte más fácil obtener sus recursos.

Los objetivos que pueden perseguir pueden ser de distinto índole. En este trabajo, nos centraremos en aquellos que tienen como objetivo filtrar, categorizar y recomendar páginas webs a sus usuarios. Para poder cumplir de manera satisfactoria dichos objetivos, el agente podrá llevar a cabo, alguna de las siguientes tareas [4]: Ver Fig. 1.

- Obtener conocimiento sobre las preferencias de los usuarios: Realizando un proceso de monitorización constante de la navegación del usuario en internet, analizando los hábitos de navegación, las páginas que visita, su lista de favoritos, etc.
- Navegar frecuentemente por internet, buscando nuevas páginas interesantes para su usuario: Incluiría lanzar consultas en los buscadores, revisar los enlaces de que dispone y navegar a través de sus hipervínculos en busca de modificaciones y contrastar la información encontrada con las preferencias del usuario.
- Atender peticiones de búsqueda de su usuario sobre temas concretos: Contrastando la información obtenida en la tarea anterior con la petición de su usuario.
- Comunicarse con agentes similares y compartir información con ellos: Podrá encontrar agentes pertenecientes a usuarios con preferencias similares a las del suyo, compartir información con estos de manera regular.

Fig. 1. Esquema de funcionamiento de un agente

- Enviar recomendaciones a su usuario: De forma que le presente la información obtenida, sin llegar a saturarle, tratando de ofertarle documentos acordes a las preferencias.

- Organizar la lista de enlaces favoritos de su usuario: Eliminando aquellos enlaces que haya dejado de existir, o buscando su nuevo emplazamiento. Organizaría por categorías y cada vez que el usuario decidiera añadir uno nuevo le sugeriría la categoría más adecuada.

- Trasladar su ejecución a otros computadores: De forma que siga a su usuario en sus nuevos emplazamientos.

Las tecnologías necesarias para que un agente lleve a cabo las tareas anteriores son de muy diversa índole. Por un lado hay las que implican poder utilizar los protocolos y capacidades como la comunicación vía HTTP, el análisis sintáctico de documentos HTML, la comunicación entre agentes, todas ellas tecnologías que se encuentran bien desarrolladas, disponiéndose de un amplio conjunto de librerías y métodos. Por otro lado, tenemos las tecnologías tendentes a dotar a los computadores de capacidades inteligentes, entre las cuales podrían estar:

- Extracción de conocimiento a partir del análisis de documentos.
- Representación del conocimiento.
- Catalogación de documentos.
- Navegación en internet guiada por intereses.

- Planificación de acciones a emprender.
- Establecimiento de políticas de colaboración con otros agentes.
- Generación de resúmenes de documentos.

Para obtener las capacidades anteriores se hace necesario el uso de técnicas de inteligencia artificial. Por ello, nos centraremos en las tres primeras proponiendo como tecnología base para la obtención de dichas capacidades el uso de redes neuronales artificiales (RNAs), por sus características de adaptabilidad, potencialidad y robustez [5][6].

B. Catalogación de documentos

La catalogación de los documentos obtenida en el proceso de búsquedas, posibilita ir más allá de la simple búsqueda de páginas similares, ya que permite dotar a los agentes de la capacidad de estructurar y organizar sus datos, adquirir conocimiento sobre las preferencias de su usuario y tal vez de usuarios de otros agentes.

La catalogación consiste en, dado un conjunto de documentos, poder identificar las distintas categorías basadas en sus contenidos, en las que se puede particionar el conjunto documentos, y luego poder asociar cada documento a una categoría [1]. Es decir, a partir de un conjunto de M elementos (documentos) no etiquetados: $\{X_i,\ i = 1, 2,..,M\}$ encontrar K agrupamientos $S_j,\ j = 1, 2,..., K$. Ello, sujeto al ajuste de algunos parámetros (ver Fig. 2.)

Fig. 2. Esquema funcional de un algoritmo de agrupamiento exclusivo

Es posible generalizar más este concepto, incluyendo una jerarquía de categorías. Para ello, debemos destacar dos aspectos exigibles a los procesos de catalogación de documentos [2]:

- Detección del número y tipo de categorías existentes.
- Asignación de categorías a los documentos.

C. Tipos de agrupación

Las formas de agrupación de objetos [4], tales como asignar clases predeterminadas a cada elemento son susceptibles de dividirse según el esquema de la Fig. 3.

- No exclusivas: Un mismo objeto puede pertenecer a varias categorías o grupos.
- Exclusivas: Cada objeto pertenece solamente a una categoría, clase o grupo.

a) Extrínsecas (supervisadas): Las clases a las que pertenecen los objetos están predefinidas, y se conocen ejemplos de cada una, ó algunos de los objetos ya están agrupados y son utilizados por el algoritmo para aprender a clasificar a los demás.

b) Intrínsecas (no supervisadas): La agrupación se realiza en base a las características propias de los objetos, sin conocimiento previo sobre las clases a las que pertenecerán.

i. Jerárquicas: Los métodos jerárquicos consiguen la agrupación final mediante la separación (métodos divisivos) o la unión (métodos aglomerativos) de grupos de documentos. Así, estos métodos generan una estructura en forma de árbol en la que cada nivel representa una posible agrupación.

ii. Particionales (No jerárquicas): Los métodos particionales, o de optimización llegan a una única agrupación que optimiza un criterio predefinido o función objetivo, sin producir una serie de anidaciones de los grupos.

La agrupación automática de documentos se encuentra en la categoría *intrínseca*, ya que los criterios de agrupamiento se basan en la información contenida en los mismos que será utilizada para determinar sus similitudes. En la Fig. 3. se muestra una clasificación de las técnicas existentes para agrupar cualquier tipo de objeto:

Entre los muchos algoritmos utilizados para la categorización de los sitios webs, destacamos:

- Principal Component Divisive Partitioning (PCDP)
- Document Cluster Tree (DC-Tree)
- Agglomerative Hierarchical Clustering (AHC)

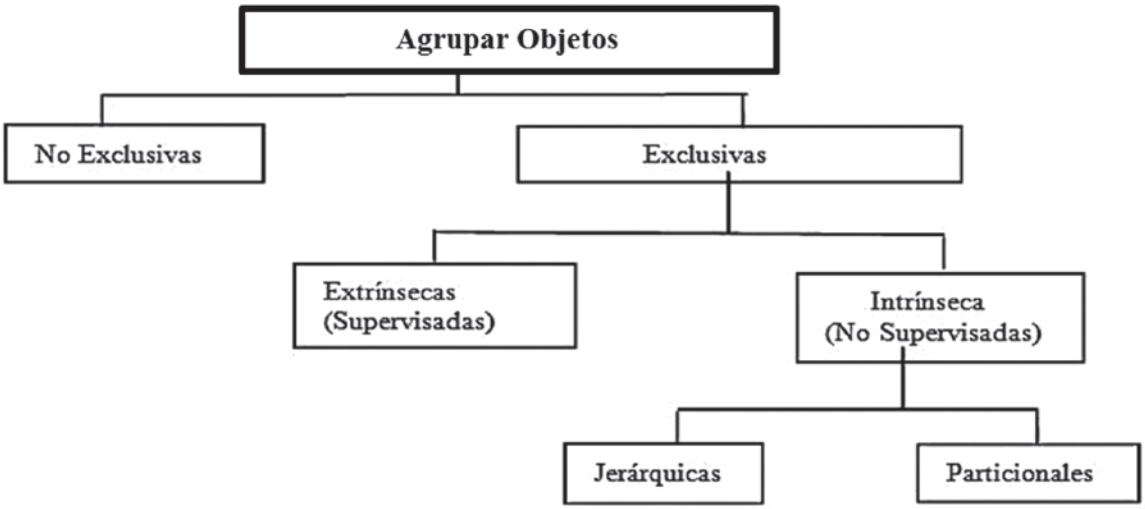


Fig. 3. Formas de agrupar cualquier tipo de objeto

Algunos algoritmos de agrupación requieren conocer previamente el número de agrupamientos a realizar y en general todos requieren unos parámetros específicos. Para hacer un agrupamiento se considera un conjunto *P* de patrones, cada uno de los cuales es característico de los agrupamientos, es decir: Sea M el número de elementos $X_1, X_2,..., X_M$ de *P*. Un *proceso de agrupamiento* o catalogación exclusivo consiste en buscar *K* grupos (subconjuntos de *P*) $S_1, S_2, ..., S_K$ tales que todo X_i, $i = 1, 2,..., M$ pertenece a uno y solo uno de estos grupos, ósea:

$$S_1 \cup S_2 \cup \cdots \cup S_K = \{X_1, X_2,..., X_M\} = P$$

$$S_i \cap S_j = 0 \quad \forall i \neq j$$

es decir, $\{ S_1, S_2, \ldots\ldots S_k \}$ constituye una partición de *P*.

Estos algoritmos, están basados en computación clásica, dentro de la computación neuronal podemos destacar los mapas autoorganizados (SOM) [5], [6] que han sido utilizados para este propósito.

III. CLASIFICACIÓN INTELIGENTE DE SITIOS WEB'S UTILIZANDO RNA'S, UN ENFOQUE PRÁCTICO

En este apartado, comentaremos, cada uno de los procesos que nos han permitido realizar una catalogación de un sitio web con un agente neuronal, orientado a catalogar las páginas webs del "Corredor del Henares". Por ello, empezaremos con el primer proceso que nos ha permitido extraer y representar los documentos de los sitios documentales, para luego llegar a obtener los vectores característicos del sitio web, y catalogarlo con nuestro agente neuronal.

A. Extracción y representación del contenido de sitios Web's

En primer lugar, es importante una representación adecuada y normalizada de los documentos que determine el contenido de un sitio web, ésta servirá al agente cuando esté navegando para poder identificar los lugares y compararlos con otros sitios. Además, conociendo los sitios webs que explora su usuario podrá por ende, obtener información de las preferencias de éste para poder luego compararlas con el resto de representaciones. La representación ha de ser adecuada para procesos de categorización de los sitios webs, es por ello, que hemos tenido que estudiar un procedimiento de selección de términos representativos de la base, así como de representación vectorial de los mismos [3]. La falta de estructuración semántica en los documentos, une el hecho de que mucha información puede venir codificada en formato multimedia y gran parte de las páginas webs existentes apenas poseen texto no siguiendo una estructura gramatical concreta. En la Recuperación de Información (RI) normalmente este problema se aborda mediante la selección semiautomática de un grupo de palabras claves [1] y la generación en base a éstas de histogramas de frecuencias ponderados para cada documento.

La selección del grupo de palabras claves habitualmente se realiza a partir de un gran conjunto de documentos webs, extrayendo todas las palabras que lo contienen. Luego, se eliminan aquellas semánticamente menos relevantes, como etiquetas de HTML, artículos, adverbios y las palabras pocas y excesivamente frecuentes. A continuación, las palabras restantes pasan otra criba, realizándose un proceso de lematixación [11], para finalmente ser seleccionadas y ponderadas, eliminándose las que obtengan peores valores, dejando la cardinalidad del conjunto de palabras claves en unos rangos tratables. En la Fig. 4., mostramos el conjunto de procesos que se toman en cuenta para llevar a cabo la extracción y representación del sitio Web [3].

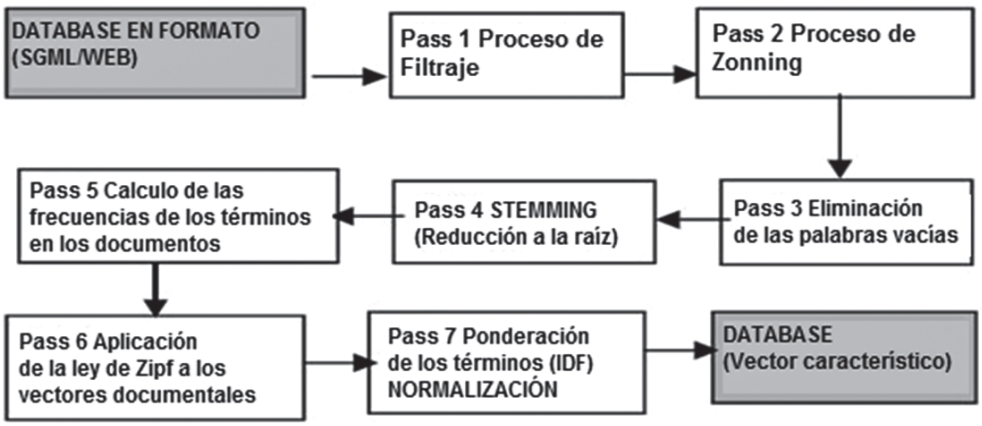


Fig. 4. Procesos desarrollados para obtener del sitio web sus vectores característicos

Los métodos de ponderación de palabras claves se basan principalmente en relacionar la frecuencia de éstas y los grupos de documentos, entre los que podemos destacar: Term Frecuency- Inverse Document Frecuency (TF-IDF) y la ley de Zipf [9][10].

De esta forma, para obtener la representación de un sitio web, a partir de un subconjunto de páginas webs, se extraen las palabras existentes de dicho subconjunto, se obtiene un histograma de frecuencias de las palabras claves seleccionadas, y luego se lematiza [11] y pondera los términos restantes, con lo que finalmente se genera el vector de características de cada documento que representa al sitio web. Habitualmente se obtiene una alta dimensionalidad para los vectores de características generados por estas técnicas, lo que suele limitar en gran medida los métodos de categorización aplicables a posteriori.

B. Red Neuronal Artificial propuesta para catalogar sitios Web's.

Se propone la utilización de un agente neuronal auto organizado, el cual ya ha sido probada con éxito en otro tipo de aplicaciones de categorización ciega. Este agente procesará la información a través de una capa de entrada que recibirá los vectores característicos del sitio web y utilizando una capa de competición generará las categorías (Ver Fig. 5.)

Por lo tanto, el agente neuronal, tiene dos capas y se caracteriza por:

- Cada neurona de competición es una categoría.
- Cada neurona de entrada está conectada con cada una de las células de la capa de competición (que se distribuyen inicialmente de forma aleatoria).
- Para cada ejemplo se calcula la salida de cada neurona de competición y nos quedamos con la mejor (ganadora).

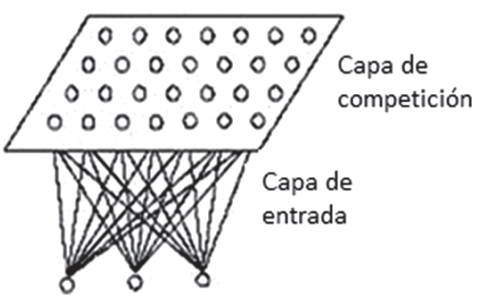


Fig. 5. Esquema del agente neuronal autoorganizado

El agente neuronal procesa adecuadamente los documentos con los vectores obtenidos de los procesos de: Extracción de características [13][14]. Luego él realizará la Generación de prototipos y el Etiquetado. El problema de catalogación de sitios webs, lo abordamos a continuación.

IV. PROCESO DE CLASIFICACIÓN DE PÁGINAS "ALCALAÍNAS" UTILIZANDO LA RNA

Para probar las capacidades de la RNA en la clasificación de sitios webs se decidió en primer lugar (Proceso de Filtraje: paso 1 de la figura 4), crear un entorno de datos basado en páginas relacionadas con la comunidad de Madrid, ciñéndose exclusivamente a aquellas páginas que sean de la zona del "Corredor del Henares". Para ello se lanzó una búsqueda en Google preguntando por páginas web en idioma español con la palabra clave "Madrid" y con la palabra clave "Alcalá de Henares". Se obtuvieron las 85 primeras páginas de enlaces devueltas por el buscador. De los enlaces encontrados se eligieron 45, eliminándose todas las páginas que no correspondían con documentos raíces de sitios webs, así, nos quedamos con los 45 enlaces del tipo: http://<direccion_servidor_web>/.

Debido al orden en que devuelve los resultados el buscador Google dichos sitios webs corresponden a los más referenciados relacionados con las palabras claves utilizadas, para luego finalmente quedarnos con 23 sitios webs debido al proceso de zonning que aplicamos al tener sitios no accesibles o caídos.

A. Representación de los contenidos Web's

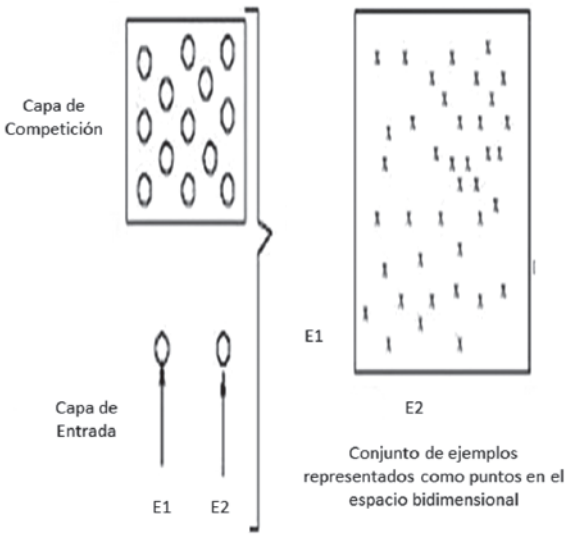
Para representar el contenido de los sitios web hallados nos basamos en conseguir un grupo de palabras claves relevantes y ponderarlas según su utilidad. Para ello obtuvimos todas las palabras existentes en las páginas webs de hasta dos niveles de profundidad, es decir, la página raíz más páginas enlazadas directamente con la raíz de los sitios mencionados, para ello se utilizó la herramienta ***wget*** [12], que es un spider de dominio público. Del conjunto de palabras distintas obtenidas se eliminaron aquellas semánticamente menos relevantes, siguiendo el procedimiento descrito en la figura 4. Al final se obtuvieron los resultados que mostramos en la tabla 1 para todos los sitios webs seleccionados.

Hubo que eliminar muchos sitios obtenidos inicialmente (Proceso de Zonning: Paso 2 de la figura 4), debido a que en el momento de calcular los vectores de características, estos no estaban accesibles, pues pertenecían a enlaces perdidos o servidores caídos, y en algunos casos el número de palabras claves que contenían dichos sitios era muy bajo. Por lo tanto, los sitios web se redujeron a 23, que procesamos para generar los vectores característicos.

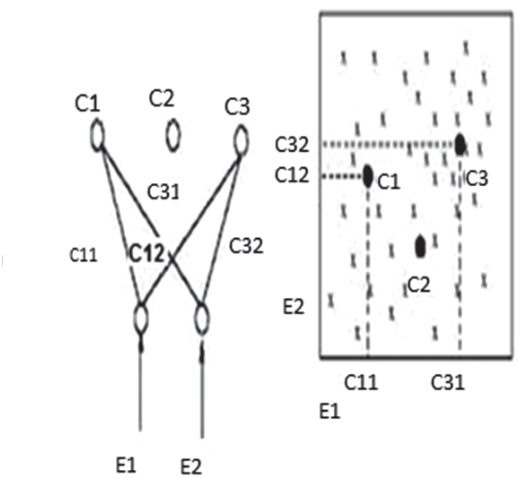
B. Procedimiento utilizado por la RNA

El procedimiento seguido para procesar los documentos con el mapa auto organizado con los vectores característicos consistió en:

1. Recibimos la entrada de cada documento del sitio web ejemplo. Cada ejemplo son representables como puntos en un espacio "*n-dimensional"*.



2. Se propaga por las conexiones hasta llegar a la capa de competición. Los prototipos también se pueden representar en el espacio y sus coordenadas quedan determinadas por los pesos de las neuronas de la capa de competición.



3. Cada neurona de esta capa de competición produce una salida al comparar el ejemplo con sus pesos.

4. Se selecciona el prototipo cuya distancia al ejemplo sea menor (neurona ganadora).

5. Los pesos de la célula ganadora se modifican para acercarse ligeramente al ejemplo modificando así el mapa de prototipos inicial, formándose así:

El proceso que lleva a cabo durante el prendizaje cada vez que se le presenta un vector, se puede resumir en:

- Seleccionar como nodo ganador (cada nodo de la RNA está representado por un vector de pesos de la misma dimensión que la entrada) el más cercano a la entrada presentada.
- Modificar el vector de pesos del nodo ganador y los nodos correspondientes a su vecindad acercándolo hacia el vector de entrada. (a veces el refuerzo es igual para toda la vecindad, en otros decrece al aumentar la distancia al ganador).

Tras el entrenamiento las neuronas topológicamente cercanas resultan ganadoras [5] con clusters de vectores que son cercanos en el espacio de entrada.

V. RESULTADOS OBTENIDOS CON LA CLASIFICACIÓN MEDIANTE LA RNA

Los resultados permitieron realizar una fácil navegación y búsqueda de los contenidos web.

Navegación y Búsqueda de Documentos: Si realizamos un clic en la imagen o en un punto blanco se puede investigar el contenido de la unidad de mapa individual. Las flechas permiten moverse a las unidades adyacentes, donde es probable encontrar textos similares. Se puede leer los textos haciendo clic en los títulos.

La imagen del mapa contiene las etiquetas que son las palabras claves o descriptores que hemos procesado del sitio web [14]. De esta forma, las etiquetas permiten dar una idea general de los temas de la colección de documentos en todo el sitio web. El ***color*** de las distintas áreas del mapa representa la densidad de los documentos en esa zona, es decir las áreas claras contienen más documentos de una etiqueta determinada.

TABLA I.
LA RELACIÓN DE LOS 23 SITIOS WEB PROCESADOS

Sitio Web	Cantidad de Documentos	Total de Palabras Procesadas	Total palabras diferentes	Términos diferentes luego del *Stop List*	Términos Lematizados (Vectores Caracteristicas)
http://www.uah.es/	511	51069	12577	5113	3883
http://www.ayto-**alcaladehenares**.es/	436	49688	12358	4928	3729
http://www.**alcala**turismo.com/	455	52160	13213	5357	4081
http://www.hotelcampanile**alcala**.com/	150	8794	4246	837	413
http://www.obispado**alcala**.org/	246	27692	13967	5545	2152
http://www.turismo**alcala**.com/	498	65543	15270	5985	4472
http://www.fgua.es/	106	807	391	212	89
http://www.institutofranklin.net/	175	3397	1249	430	65
http://www.icaah.com/	152	2372	762	305	110
http://www.alcine.org/	257	39573	14452	5792	3367
http://www.lacallemayor.net/	485	54784	13469	5477	4167
http://www.botanicoalacala.es/	415	9878	4418	773	367
http://www.esn-uah.org/	269	6022	3971	1782	381
http://www.corraldealcala.com/	199	26936	7781	3630	2807
http://www.**alcala**delasartesylasletras.es/	179	9747	4339	3081	2430
http://tuna.**alcala**.org/	141	7625	5640	4961	2753
http://www.20minutos.es/	488	56441	13160	5188	3917
http://www.museo-casa-natal-cervantes.org/	198	5996	2408	543	285
http://www.**alcal**ingua.com/	128	4589	2875	474	235
http://eoi.**alcala**.**alcala**.educa.madrid.org/	216	9515	5579	3907	2757
http://www.ruah.es/	102	1752	905	481	908
http://www.**alcala**joven.org/	273	21673	9651	5164	2527
http://www.diariodealcala.es/	1402	91708	63677	51406	30337

Tabla 1: La Relación de los 23 sitios web procesados

Fig. 6 Resultado final obtenido por el SOM, luego del entrenamiento

La TABLA I muestra las características de los sitios procesados y la Fig. 6. el resultado final obtenido con el SOM, los tiempos de generación de categorías y catalogación de documentos ejecutados sobre una máquina SUN fueron alrededor de 5 minutos, tiempos relativamente

buenos si se tiene en cuenta que no se emplearon modificaciones a la RNA para acelerar el aprendizaje (porque se trataba con entradas de alta dimensionalidad). Los mejores comportamientos se obtuvieron utilizando el producto escalar y normalizando las entradas. Sin embargo, dado el conjunto de datos utilizados, resulta difícil cuantificar la bondad del sistema, pues a pesar de que los clusters parecen coherentes, también se encontraron otros en los que los sitios webs agrupados no parecían guardar mucha relación. Por lo que estos resultados dan pie para seguir investigando en la mejora y evaluación más profunda del sistema propuesto de catalogación de sitios webs, tal vez afinando el mapa autoorganizado.

VI. CONCLUSIONES

El presente trabajo sienta las bases de lo que en un futuro podrán ser los agentes webs inteligentes aplicados al campo de la documentación. Las pruebas experimentales se realizaron tomando en cuenta un determinado sitio web, y los resultados obtenidos en la catalogación son prometedores, todo ello nos indica la factibilidad del uso de un agente neuronal auto-organizado para la catalogación respectiva del sitio web. Las pruebas realizadas proporcionan los lineamientos a seguir para la mejora de los métodos de representación de los documentos en un sitio web, en especial cuando se trabaja en un espacio *n-dimensional*.

ACKNOWLEDGMENT

We would like to thank Departament of Computer Science of University of Alcalá for its support and for providing this research oportunity.

Work done under project: VOA3Rr: Virtual Open Access Agriculture & Aquaculture Repository: Sharing Scientific and Scholarly Research related to Agriculture, Food, and Environment (CIP-ICT-PSP.2009.2.4: 250525 - VOA3R)

REFERENCES

[1] R. Baeza-Yates, B. Ribeiro-Neto *"Modern Information Retrieval"*. ACM Press Addison-Wesley, 1999. ISBN:0-201-39829-X .

[2] Berry Michael, Malu Castellano Editors: *"Survey of Text Mining II "*, Springer 2008. ISBN: 978-1-84800-045-2

[3] José Luis Castillo S, José R. Fernández del Castillo, León González S., *"Feature reduction for document clustering with NZIPF method"*, Proceedings of IADIS International Conference e-Society 2009. Press- ISBN: 978-972-8924-78-2.

[4] Cios Krzysztof, Pedrycz Witold, Swiniarski Roman, Lukasz Kurgan, *"Data Mining:A Knowledge Discovery Approach"*,Springer, 2007,

[5] J.A. Freeman, Skapura, "*Redes Neuronales. Algoritmos, aplicaciones y técnicas de propagación*". México: Addison – Wesley. 1993.

[6] José Ramón Hilera González.; Martínez Hernando. "*Redes neuronales artificiales: fundamentos, modelos y aplicaciones*". Madrid: RAMA.1995.

[7] G, Kowalsky Gerald "*Information Retrieval Systems*" Kluwer Academic Publishers 1997. ISBN:0-7923-9899-8.

[8] T. Larose Daniel *"Data Mining Methods and Models"* A John Wiley & Sons, Inc. Publication, 2006. ISBN-13: 978-0-471-66656-1.

[9] D. Olson *"Advanced Data Mining Techniques"*, Springer 2008 ISBN:978-3-540-76916-3.

[10] Witold Pedrycz "Computational Intelligence" CRC Press 1998, ISBN:0-8493-2643-5.

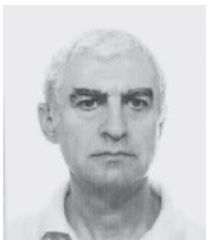
[11] M.F. Porter *"An Algorithm for Suffix Stripping, Program"*, vol. 14, no. 3, 130- 137, 1980.

[12] Wget:WWW Mirroring & Retrieval Software: http://www.gnu.org/s/wget/ GNU.

[13] Man L. Wong , Kwong S. *"Data Mining using grammar based Genetic Programming and Applications*" Kluwer Academic 2002 Editor John Koza. ebook ISBN:0-306-47012-8.

[14] N. Ye , *"The Handbook of Data Mining"* Lawrence Erlbaum Associates, Publishers, London, ISBN: 0-8058-4081-8, 2003.

BIOGRAFÍA



León Atilano González Sotos, Licenciado en Matemáticas por la Universidad de Valencia (1978); Doctor en Ciencias Matemáticas por la Universidad de Alcalá (1987). Assistant Professor Universidad es de Sevilla y Pública de Navarra. Full Professor de Ciencias de la Computación de la Universidad de Alcalá. Research Fields: Uncertainty Logic; Artificial Intelligence; Information Systems.

José Luis Castillo Sequera, Licenciado en Computación por la Universidad Nacional Mayor de San Marcos (Perú); Doctor en Informática por la Universidad de Alcalá (España). Full Professor de Ciencias de la Computación de la Universidad de Alcalá. Research Fields: Information retrieval, Knowledge extraction based on evolutionary learning and Neural Network, Data Mining, Documentation, Learning and teaching innovation with Web 2.0.

José Raúl Fernández del Castillo Díez, Licenciado en Físicas por la Universidad Autónoma de Madrid en 1994 y doctorado en Ciencias Físicas en 1998 por la misma universidad. Desde 1998 imparte clases en la Universidad de Alcalá, en la que es Profesor Titular desde el año 2000 en el departamento de Ciencias de la Computación. Research Fields: Uncertainty Logic; Artificial Intelligence; Information Systems Information retrieval, Knowledge extraction, Data Mining, modeling systems behavior and Learning and teaching innovation.