Sistema neuronal difuso para la gestión de documentos estructurados

G. I. Sainz Palmero† J. M. Cano Izquierdo† Yannis A. Dimitriadis‡
J. López Coronado†

†Dpto. de Ingeniería de Sistemas y Automática
E.T.S. de Ingenieros Industriales
Paseo del Cauce s/n
47011 Valladolid
Tfno. Fax: 983-423358
E-mail: gresai@dali.eis.uva.es

‡Dpto. de Teoría de la Señal, Comunicaciones e Ingeniería Telemática
E.T.S. de Ingenieros de Telecomunicación
C/ Real de Burgos s/n
47011 Valladolid

Palabras Clave: Documento electrónico, Objeto lógico, Objeto físico, Papel electrónico, Redes Neuronales, ART-MAP, Fuzzy, ODA/ODIF, SGML-HTML.

Resumen

En la presente comunicación se presenta el problema de la gestión automatizada de documentos electrónicos desde un doble punto de vista: normalización y reconocimiento-clasificación. Se propone un sistema que aborda estos problemas mediante la utilización, por un lado, de la arquitectura de documentos de la norma ODA/ODIF (ISO8613) y, por otro, de redes neuronales, Fuzzy ARTMAP, para la clasificación del documento y sus componentes.

1 Introducción

El documento en su forma impresa sobre algún tipo de material, en especial sobre papel, ha sido la manera tradicional con la que el hombre ha transmitido información a través del tiempo. Pero este formato presenta grandes inconvenientes, entre los que destacan: almacenamiento, gestión, conservación, actualización, etc... todo lo cual se traduce en elevados costos económicos.

Hoy en día parte de los inconvenientes expuestos se han resuelto gracias a la utilización de los denominados documentos electrónicos sobre sistemas informáticos, ordenadores y redes. Un documento electrónico puede ser creado por diversos medios:

- Procesador de textos convencional, por muy elemental que éste sea: Wordperfect, MS Word, etc...
- Captura de documentos impresos a través de un scanner.

- Editores "*en línea*" que utilizan el concepto de papel electrónico [Higgins-91] y la escritura manuscrita [Dimitriadis-95].

  En estos sistemas los documentos se crean en tiempo real por medio de un interfaz "*natural*", papel y lapicero electrónicos, entre la máquina y el usuario. Estos sistemas son de gran interés tanto en ofimática, como en aquellas áreas donde se busca una mejora de la relación hombre-máquina.

Pero, por otro lado, con la utilización de estos documentos electrónicos han surgido nuevas exigencias derivadas del objetivo final que se desea alcanzar: una gestión y manipulación automatizada, objetivo que nos plantea:

1. Necesidad de reconocer y comprender el documento .
2. Capacidad para compartir e intercambiar dicho documento.

Por tanto, la gestión avanzada de documentos electrónicos se debe centrar en la resolución de dos tipos de problemas fuertemente interrelacionados:

- Identificación y Clasificación del documento y de su contenido.
- Normalización o Estandarización de la organización del documento que nos permita compartirlo o intercambiarlo entre diferentes usuarios o sistemas.

A partir de esto, el desarrollo de la presente comunicación será el siguiente: Una primera parte donde se presenta "*el documento*" como problema de clasificación y de normalización, a continuación de lo cual se comentarán los rasgos generales de la Teoría de Resonancia Adaptativa (ART) de redes neuronales, ampliamente utilizada para la resolucíon de problemas de reconocimiento y clasificación de patrones.

En la segunda parte se esboza nuestro sistema para la creación y gestión de documentos, más tarde se propone un método para la caracterización de los bloques físicos de un documento y se expone el trabajo experimental con los resultados obtenidos. Finalmente se presentan las conclusiones alcanzadas.

2 El documento como problema de normalización y clasificación

Un documento se puede considerar como una suma estructurada de texto (información perceptible por el ser humano) que puede ser intercambiado entre un emisor y un receptor [Horak-85b].

Por tanto, un documento va a ser un objeto estructurado, en un doble sentido:

- Físico, puesto que el documento se organiza en páginas, parrafos, líneas, etc...
- Lógico, ya que en un documento hay títulos, tablas, capítulos, resúmenes, etc... elementos con significado para el lector.

A partir de esto podemos representar un documento como dos estructuras jerárquicas, una física y otra lógica, en las cuales se disponen los diferentes componentes u objetos del documento. Pero el documento tiene un único contenido, luego ambas estructuras se van

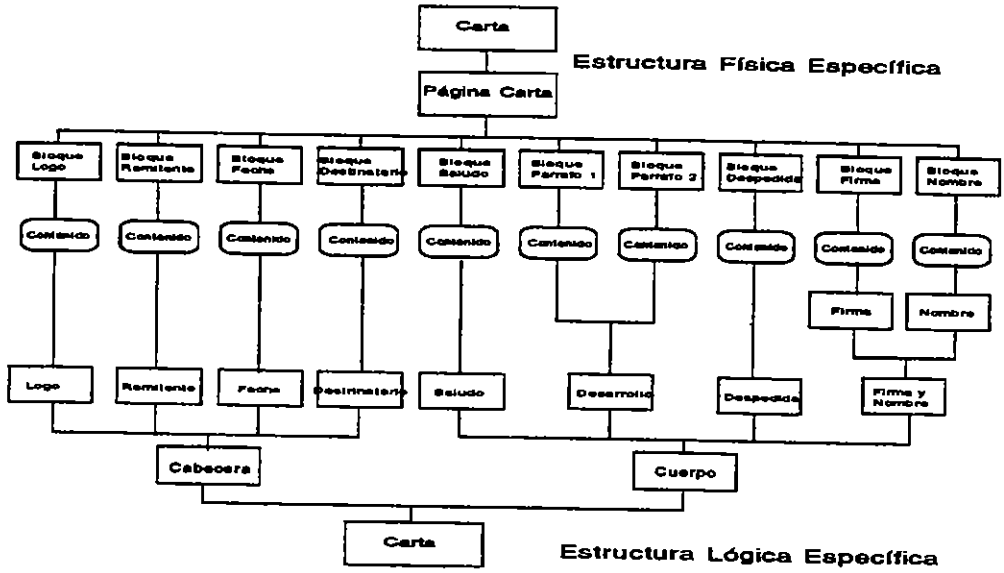

Figura 1: Estructura de un documento

a relacionar a través del contenido de la carta. Un ejemplo de esta relación se muestra en la figura 1 para el caso de un documento tipo carta.

En este momento ya surgen los primeros problemas de organización: ¿Cómo se deben organizar estas estructuras? ¿Qué tipo de contenido puede tener el documento? ¿Qué objetos lógicos y/o físicos puede tener el documento?, etc... Estamos, por tanto, ante el problema de normalización o estandarización en la organización interna o arquitectura de los documentos.

A este problema se añade uno similar si queremos intercambiar nuestro documento, ¿Cómo debemos realizar el intercambio?, ¿En qué formato debemos expresar el documento para poder almacenarlo o intercambiarlo de manera que tenga la mayor difusión posible?. En este punto hay que tener en cuenta que una mala elección, tanto en la arquitectura, como sobre todo en el formato de intercambio nos puede llevar a la situación de encontrarnos con un documento "*inútil*".

Así se llega a la conclusión de que si se desea conseguir una gestión automatizada de documentos es imprescindible definir las arquitecturas y las formas de almacenamiento e intercambio de los documentos electrónicos, con lo cual surge el problema de la elección de estándares. Entre las arquitecturas y formatos propuestos para resolver este problema están: Office Document Architecture/Interchange Format (ODA/ODIF), Standard Graphics Markup Language (SGML-HTML), Document Computer Architecture (DCA), DDIF, etç... [Appelt-91] [Chabrier-90] [Horak-85b]

Una vez elegida la forma de cómo organizar internamente e intercambiar el documento debemos obtener o crear el documento electrónico, es decir, "*rellenar*" las estructuras lógica y física del mismo.

El punto de partida de este proceso va a ser la información física obtenida por medio de un scanner, una tableta digitalizadora, etc.... A partir de ésta, mediante procesos de segmentación, podemos obtener la estructura física del documento, es decir, la disposición

geométrica, las dimensiones, etc... de los bloques de información (físicos) del documento. Pero esto es insuficiente para reconocer o "*comprender*" el documento, necesitamos saber qué significa cada bloque: un título, una dirección, un autor, etc... pues es con esta información con la que podemos realizar una gestión automática de alto nivel.

Por tanto se nos plantea el problema de reconocer y clasificar los distintos objetos lógicos que componen un documento a partir de la estructura física del mismo, este proceso se denomina frecuentemente etiquetado lógico [Dengel-94] [Kreich-93] [Marovac-90] [Andre-89].

Los problemas generales de clasificación y "*clustering*" (agrupamiento) están relacionados con procesos de aprendizaje, con la adquisición de conceptos a partir de los ejemplos que nos vamos encontrando y la posterior generalización de esos conceptos para que puedan englobar a elementos nuevos. Durante este proceso, a partir de una serie de características que percibimos de los objetos mediante los sentidos y con una posible ayuda de un maestro, alguien que es capaz de darnos las etiquetas que debemos asociar a cada objeto que percibimos, creamos ideas abstractas a partir de objetos concretos.

Esto aplicado a nuestro problema de reconocimiento-clasificación en un documento corresponde a una serie de objetos o bloques físicos y una serie de etiquetas, lógicas (título, autor, etc...), que queremos asociar a ellos. En la fase de aprendizaje, si existe la presencia de un maestro exterior, sabemos fehacientemente qué etiqueta (objeto) lógica se corresponde con cada uno de los bloques físicos. Después del aprendizaje deberemos ser capaces de asociar correctamente, no sólo los objetos físicos que hemos aprendido con sus etiquetas, sino aquellos que nunca hemos visto pero que pertenecen a las mismas categorías que hemos aprendido. Podemos considerar esto como un sistema que, durante la fase de aprendizaje se alimenta con vectores de características de los objetos físicos y con las etiquetas o bloques lógicos asociadas a dichos objetos físicos. Después del aprendizaje el sistema es capaz de dar como resultado la etiqueta asociada al bloque físico que le estemos dando. La presencia del maestro exterior que proporciona las etiquetas hace que este proceso se conozca como aprendizaje supervisado.

Cuando carecemos de un supervisor externo nos enfrentamos ante un problema con mínima información. Tan solo disponemos de los ejemplos de bloques físicos. Se puede plantear entonces un problema de clustering. Se trata de que el sistema sea capaz de descubrir por sí mismo la estructura de los bloques físicos, su distribución de probabilidad. Para ello el sistema deberá agrupar aquellos datos que estén interrelacionados y crear, por sí mismo una serie de "*clases*" de objetos lógicos que reflejen la estructura de los objetos físicos. Ante la entrada de cada nuevo dato el sistema nos dirá a qué clase pertenece, es decir con cuáles de los datos aprendidos está relacionado este nuevo dato. Se dice que estamos ante un problema de aprendizaje no supervisado.

Para realizar nuestra clasificación hemos utilizado un sistema neuronal-difuso [Kosko-92], un modelo supervisado tipo ARTMAP, basada en la Teoria de Resonancia Adaptativa (ART), cuyas líneas generales se exponen a continuación.

3 Redes Neuronales: Teoría de la Resonancia Adaptativa

La teoría de la resonancia adaptativa (ART) es un intento de modelar ciertas facetas del comportamiento del sistema nervioso y perceptivo humano [Grossberg-82] [Grossberg-80]. Esta teoría ha sido propuesta a partir de los trabajos de Stephen Grossberg en la Uni-

versidad de Boston. Partiendo de investigaciones en el campo de la psicología se busca encontrar las leyes que rigen el comportamiento humano. Una vez descritas cualitativamente estas leyes, se busca un modelo matemático que sea capaz de reflejarlas.

Durante el transcurso de su vida el hombre va incorporando conocimientos nuevos en cada momento. La incorporación de estos nuevos conocimientos no implica la pérdida de los ya aprendidos, un niño no olvida lo que es un perro cuando aprende lo que es un gato. Esta observación, que puede parecer trivial, es una propiedad que difícilmente cumplen los modelos artificiales de redes neuronales propuestos. Este dilema entre estabilidad (capacidad de mantener la información aprendida por el sistema) y plasticidad (capacidad de incorporar nueva información) es asumido como una de las principales características de la teoría ART. Otro dilema que asume la teoría resonante adaptativa es el de ruido-saturación que se plantea como la capacidad del sistema para ser sensible ante señales de baja intensidad sin amplificar el ruido de las señales ni saturarse ante entradas de alta intensidad.

Otro fundamento en esta teoría es la búsqueda de sistemas que sean dinámicos y que trabajen en tiempo real, entendiéndose por trabajo en tiempo real el hecho de que la fase de aprendizaje y de funcionamiento del sistema no están disociadas. El sistema debe ser capaz de ir incorporando información (aprender) al mismo tiempo que está funcionando. El asumir esta serie de premisas iniciales nos conduce a la búsqueda de sistemas que reflejen en su funcionamiento, con la mayor fidelidad posible, el comportamiento humano.

Dentro de esta teoría se han propuesto una serie de modelos que tratan los problemas de clasificación y clustering. Como modelos no supervisados podríamos mencionar ART 1, ART 2, ART 3 y Fuzzy ART. Los modelos supervisados se construyen a partir de estos modelos no supervisados utilizando la arquitectura ARTMAP.

3.1 Modelos no supervisados

Podemos considerar los modelos ART como algoritmos de clustering con dos distancias, que se corresponderían con las preguntas:

1. ¿Qué categoría es la más parecida?
2. ¿Son suficientemente parecidas la entrada y la categoría?

El modelo Fuzzy ART incorpora ideas y operaciones de la teoría de conjuntos difusos en las arquitecturas ART. Esta característica es altamente sinergética ya que la arquitectura resultante hereda características de ambos campos tales como: el caracter intuitivo de las representaciones mediante conjuntos difusos, cercanas al lenguaje natural y las propiedades de aprendizaje de las redes neuronales.

En la figura 2 podemos ver dos módulos de la arquitectura Fuzzy ART. Sus componentes principales son:

- Capa de unidades de entrada: de dimensión M donde se registran las entradas.
- Nivel F_0: de dimensión $2M$ donde se lleva a cabo el cálculo del código complementario.
- Nivel F_1: de dimensión $2M$ que registra la entrada normalizada.

- Nivel F_2: con N unidades donde se registran las categorías.

Las unidades de F_1 y F_2 están interconectadas mediante pesos adaptativos. La arquitectura se completa con un mecanismo de RESET que se encarga de la generación de nuevas categorías.

Los distintos elementos que conforman la arquitectura Fuzzy ART son:

- Vectores de entrada: Son vectores M-dimensionales $I = (a_1 ... a_M)$ donde cada componente a_i pertenece al intervalo $[0, 1]$

- Vector de pesos: Cada unidad j de F_2 lleva asociado un vector de pesos $W_j = (w_{j1} \ldots w_{j2M})$. El número de unidades N que podemos tener en la capa F_2 puede ser tan grande como queramos. Los pesos se inicializan:

$$w_{j1} = \ldots w_{j2M} = 1$$

  Veremos que estos pesos representarán a las categorías asociadas a las respectivas unidades, y permanecen con este valor inicial hasta que son elegidas en alguno de los ciclos de aprendizaje; a partir de esto los valores decrecen monótonamente.

- Parámetros: en el modelo tenemos tres parámetros:

  - factor de elección $\alpha > 0$
  - factor de aprendizaje $\beta \in [0, 1]$
  - parámetro de vigilancia $\rho \in [0, 1]$

  La elección de estos parámetros condiciona en gran parte el funcionamiento del sistema, en ella intervienen factores heurísticos y de conocimiento del problema con el que tratamos.

3.2 Aprendizaje supervisado: Fuzzy ARTMAP

Hasta ahora hemos visto modelos no supervisados. Los modelos de este tipo son capaces de hacer una clasificación no supervisada, es decir de forma autónoma, de las entradas que vamos suministrando. Ahora nos vamos a centrar en modelos de tipo supervisado. La arquitectura Fuzzy ARTMAP consta de dos módulos de tipo ART y de un módulo de interconexión denominado Mapa InterART [Carpenter-92]. En la figura 2 podemos ver como sería esta arquitectura utilizando módulos Fuzzy ART.

Su funcionamiento está basado en la idea de asociación de conceptos. Cuando introducimos en cada uno de los módulos Fuzzy ART dos entradas, que sabemos *a priori* están relacionadas, se nos activarán dos categorías, una en cada uno de los módulos Fuzzy ART. Los pesos de las unidades del mapa Inter ART se adaptarán reforzándose el peso que interconecta las categorías que se han activado simultaneamente y atenúandose el resto de los pesos que conectan esas dos categorías con el resto. Cuando en la fase de funcionamiento introduzcamos una sola entrada por uno de los módulos Fuzzy ART se nos activará una categoría, observando con qué categoría del otro módulo Fuzzy ART está relacionada obtendremos la salida.

El mecanismo de RESET Inter ART permite que el aprendizaje sea estable, haciendo que el tamaño de los clusters que se van generando en la fase de aprendizaje, sea lo suficientemente grande para asegurar la generalización y lo suficientemente ajustado para evitar el error de predicción.

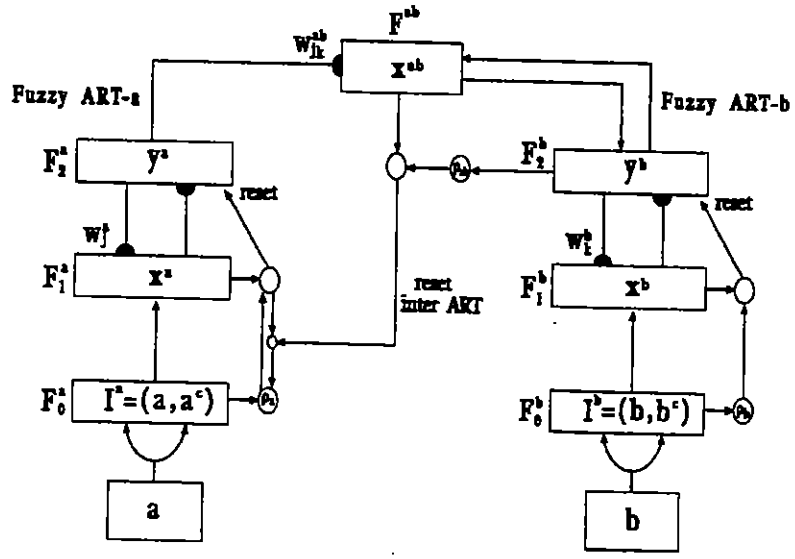

Figura 2: Arquitectura Fuzzy ARTMAP

4 Sistema para la creación y gestión de documentos electrónicos

Nuestro sistema intenta resolver el problema de clasificación de un documento por medio de redes neuronales, Fuzzy ARTMAP, y utiliza, básicamente, la arquitectura de documento propuesta en la norma ODA/ODIF. El diagrama de bloques del sistema se muestra en la figura 3, describiéndose su funcionamiento a continuación.

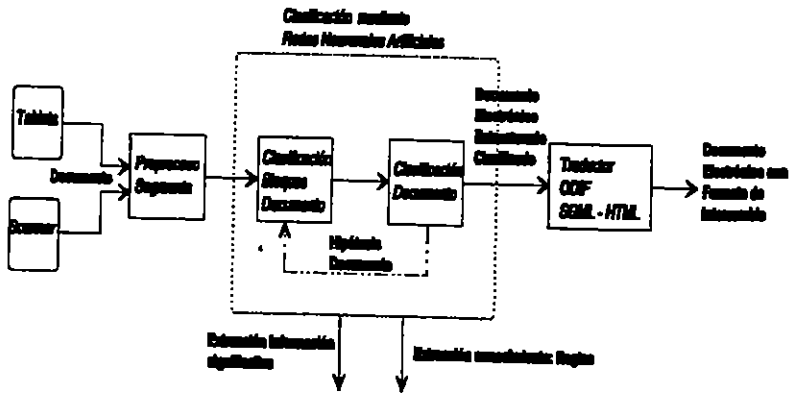

Figura 3: Diagrama de Bloques del Sistema

La entrada o captura del documento al sistema se puede realizar de dos formas:

- **Fuera de línea**, a través de un scanner se captura un documento impreso.
- **En línea**, a través de una tableta digitalizadora, utilizando el concepto de papel electrónico.

Después del preprocesamiento del documento de entrada, el primer objetivo del sistema es clasificar tanto los componentes lógicos que forman el documento: título, autor,

destinatario, etc..., como el documento considerado en conjunto: carta, artículo, etc..., para lo cual se utiliza un módulo Fuzzy ARTMAP, comentado anteriormente.

La utilización de redes neuronales va a permitir, en la fase inicial de aprendizaje, la incorporación de todo el conocimiento disponible sobre los tipos de documentos y sus componentes, y en una fase posterior de funcionamiento del sistema, donde podrá adquirir nueva información o modificar la inicial. Por otro lado, nos permite utilizar la lógica difusa aplicable en el campo del reconocimiento y clasificación de documentos debido a la gran variedad que presentan los mismos, incluso tratándose del mismo tipo/clase de documento, como consecuencia de los diferentes hábitos de cada autor o procesador de textos [Le-94] [Fujihara-93].

Durante este proceso, de reconocimiento y clasificación, se puede extraer:

1. Información parcial significativa sobre el documento tratado. Así por ejemplo, se puede obtener el título, el autor, la dirección, etc... que pueden ser incorporados a bases de datos, sistemas de gestión automática, etc...

2. Base de Conocimiento, por medio de reglas, sobre cada uno de los componentes del documento, como de este en su conjunto: ¿qué componentes lógicos tiene cada documento? ¿cuántos? ¿cómo son?, etc...

Una vez, reconocido, clasificado y estructurado el documento, de acuerdo a una arquitectura de documentos electrónica estándar (ODA), éste se debe poder almacenar e intercambiar de la manera más normalizada posible por lo cual se desarrollarán una serie de módulos cuya misión es servir de traductores, analizadores sintácticos, que permitan la codificación en formato ODA/ODIF y SGML-HTML del documento, y viceversa, partiendo de la codificación ODA/ODIF o SGML-HTML obtener la representación interna estructurada del documento.

5 Caracterización geométrica y cuantitativa difusa de un documento

Nuestro objetivo inicial es identificar los elementos representativos de cada tipo de documento: título, autor, etc... partiendo de la información obtenida tras el preprocesamiento del documento introducido por medio del scanner o de la tableta digitalizadora.

El primer paso en este proceso de reconocimiento y clasificación, es la elección del tipo de bloque/componente físico (líneas, palabras, caracteres, etc...) del documento y la caracterización que se va a presentar del mismo al módulo neuronal, el cual clasificará, o etiquetará, a dicho bloque físico como un determinado componente lógico, es decir, con significado para el autor o lector, de una determinada clase de documento.

Para esta elección, como bloques físicos proponemos la utilización de conjuntos de una o más líneas que sean característicos en la disposición física de un determinado tipo de documento y su caracterización mediante información básicamente geométrica [Dengel-88] y cuantitativa: posición, tamaño, número de líneas y palabras del bloque. Toda esta información es obtenida del preprocesamiento de la imagen del documento, sin la utilización, por tanto, de cualquier tipo de semántica sobre el contenido de dichos bloques.

Por otro lado, debido a la gran variabilidad que pueden presentar cada uno de los componentes en un documento, dependiendo del autor y/o de la clase de documentos, hemos

utilizado un proceso de "*difuminado*" (fuzzy) con el cual transformar los datos numéricos (posición, tamaño,número líneas, etc..), que nos sirven para caracterizar cada bloque físico, en la presencia o ausencia de unos determinados atributos que hacen referencia a la posición del bloque así como al número de palabras y líneas que contiene. Estos atributos son de tipo: pocas líneas, muchas líneas, pocas palabras, etc...

6 Trabajo experimental y resultados

En nuestro trabajo experimental nos hemos limitado a la utilización de un único, pero significativo, tipo de documento: la carta. Este tipo de documento es uno de los objetivos principales de nuestro sistema por ser, probablemente, el de mayor uso, y además servir para comparar nuestro sistema con otros que persiguen el mismo objetivo [Dengel-92] [Bleisinger-92].

Las cartas utilizadas como entradas del sistema son mecanografiadas (ver figura 4), escritas libremente por múltiples usuarios, en diferentes partes del mundo, utilizando diversos tipos de máquinas de escribir, procesadores de texto e incluso habiendo alguna carta escrita a mano. Por tanto, estas cartas han sido escritas tal y como sus autores han considerado necesario para que cumplieran su misión.

Los elementos lógicos que se deseaban identificar, o etiquetar, eran:

- Remitente.
- Fecha.
- Destinatario.
- Presentación inicial.
- Despedida.
- Nombre y/o cargo del firmante.

No todas las cartas utilizadas poseen todos estos elementos u objetos lógicos, ni tampoco los presentan con el mismo formato o posición.

Los bloques físicos de partida han sido conjuntos de una o más lineas, a los cuales se les ha computado su posición en coordenadas X e Y, el número máximo de palabras entre todas las líneas del bloque y el número de líneas del bloque.

Para proceder con el proceso de "*difuminado*" se ha procedido de forma heurística a través de histogramas, donde se ha observado en qué zonas del area de la carta se situan habitualmente, el número de líneas, etc... de los componentes lógicos estudiados: fecha, remitente, etc...

Mediante este proceso se ha dividido la altura de la carta en 8 niveles o categorias, en principio no disjuntos, la anchura de la carta en 5, el número de palabras en 4 y el número de líneas del bloque en 3. Por tanto, se considerarán 20 atributos, la ausencia o presencia de estos en un bloque se reflejará por 0 o 1. En consecuencia trabajamos con vectores de características binarios, aunque ya se está probando con vectores analógicos que representan valores de pertenencia de los bloques a cada uno de los atributos o categorias.

Los vectores binarios de características de los bloques físicos son similares al siguiente, que representa a una **Fecha**:

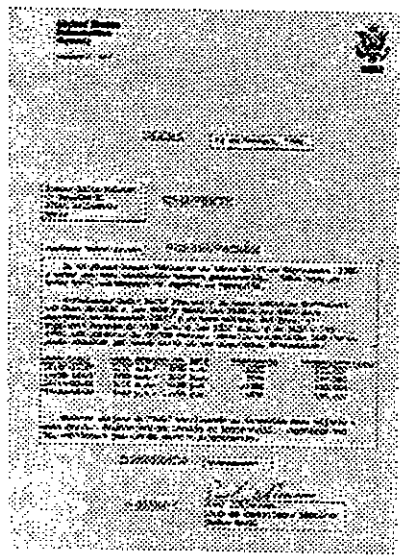

Figura 4: Ejemplo de carta utilizada

y0	y1	y2	y3	y4	y5	y6	y7	x0	x1
0	1	0	0	0	0	0	0	0	0
x2	x3	x4	p0	p2	p6	p9	l0	l2	l6
1	1	1	0	1	0	0	1	0	0

Tabla 1: Vector binario de características

Estos vectores se han introducido al módulo neuronal para su identificación y clasificación. La clasificación de un determinado bloque no va a depender de las clasificaciones anteriores, es decir, no está condicionada por los resultados obtenidos con los bloques previos. Por tanto, se puede dar el caso de que dos vectores de características sean asociados con el mismo componente lógico, por ejemplo **Fecha**, esto puede ser una inconsistencia, puesto que normalmente una carta sólo tiene un elemento lógico de este tipo. Elementos de este tipo se usarán en una fase posterior para la caracterización completa y consistente del docuemnto.

Los parametros de funcionamiento de la red neuronal son:

PARÁMETRO	Fuzzy ART^a	Fuzzy ART^b	InterART^{ab}
β	1	1	
α	0.0001	0.0001	
ρ	0.44	0.98	0.7

Tabla 2: Parámetros del módulo neuronal

El desarrollo de la actividad en el módulo neuronal tiene dos fases:

1. **Fase de Aprendizaje**, en la que se presentan al módulo 16 cartas, con 73 bloques físicos, y por tanto al ser supervisada esta fase, 73 etiquetas u objetos lógicos.

2. **Fase de Funcionamiento**, en la que se clasifican 6 cartas con 28 bloques físicos.

Una vez finalizada la *fase de aprendizaje*, se comprobó que se habían creado las diversas categorias de objetos lógicos que se muestran en la tabla 3.

Tipo Componente Lógico	Número de nodos
Fecha	4
Remitente	1
Destinatario	2
Presentación	4
Despedida	3
Nombre	6

Tabla 3. Resultados de la fase de aprendizaje

De acuerdo a esta tabla, el sistema distingue diferentes clases para cada uno de los tipos de componentes lógicos que intentamos obtener, así por ejemplo, diferencia cuatro clases de *fecha*, esto se debe a los distintos formatos que los usuarios dan a este componente, como por ejemplo:

Formato 1	*Valladolid, 10 de marzo de 1995*
Formato 2	*10 de marzo de 1995*

Tabla 4. Formatos del componente **Fecha**

Además se debe tener en cuenta su posición geométrica dentro del documento, tanto en coordenada X, como en coordenada Y, es decir, más arriba o más abajo, o más a la derecha o a la izquierda del área de la carta.

Todo esto significa que para una clase de objeto lógico, perteneciente a un tipo de documento, el sistema es capaz de distinguir subclases originadas por el estilo del autor o del procesador de texto utilizado. Esto nos va a servir para poder clasificar el documento, entendido como una unidad, como perteneciente a una determinada clase y subclase, por ejemplo: Clase: carta, Subclase: Winword o bien Clase: carta, subclase: Universidad de Valladolid.

En la *fase de funcionamiento* o prueba, se obtuvieron los siguientes resultados:

Componente Lógico	Clasificación Correcta
Fecha	100%
Remitente	100%
Destinatario	100%
Presentación	100%
Despedida	80%
Nombre	40%

Tabla 5. Resultados de la fase de funcionamiento

En estos resultados se observa una apreciable diferencia del exito en el reconocimiento alcanzado de unos componentes respecto a otros. Una primera causa de estos es el intento de reconocer componentes tipo **Nombre**, cuyo formato más habitual suele contener varias líneas, como se puede observar en el Formato 1 de la tabla 6.

Formato 1	Formato 2
G. Ismael Sainz Director Técnico Intergat, S.L	G. Ismael Sainz

Tabla 6. Formatos del componente **Nombre**

Pero entre las cartas utilizadas existian varias de ellas con el componente tipo *Nombre* con una única línea, como el mostrado en el formato 2 de la tabla 6.

Este formato hace que, con la caracterización binaria utilizada hasta ahora, obtengamos un vector de características muy similar al obtenido para ciertos ejemplares de una determinada clase de **Despedida**, componente que suele presentar un formato similar, pues ocupa una sola línea, tiene pocas palabras y su posición geométrica es similar. Esto hace que la red neuronal, durante su fase de aprendizaje, cree dos nodos casi idénticos cuyos vectores de pesos se diferencian en una única posición, lo que hace que cuando se presenta a la red uno de estos vectores de características lo clasifique con el primero de estos nodos "*similares*" con el que se le intente asociar.

Un segundo aspecto a tener en cuenta con estos resultados, es la existencia de dos grupos de componentes lógicos en el documento tipo carta:

1. Componentes lógicos que no dependen del cuerpo de la carta, como es el caso de: Fecha, Remitente, Destinatario, Presentación. Estos no ven afectada, principalmente, su posición por la amplitud del cuerpo (desarrollo) de la carta, y por tanto, su identificacón y clasificación a partir de información geométrica no se ve afectada por un factor tan voluble como es la amplitud del desarrollo de la carta.

2. Componentes lógicos que dependen del cuerpo de la carta, como el caso de: Despedida y Nombre. En este caso, la identificación y clasificación se ve complicada por la amplitud del cuerpo de la carta lo que distorsiona de forma importante la información geométrica.

   Así por ejemplo, para el mismo autor y estilo de carta la posición del objeto lógico **Nombre** puede tener grandes variaciones, lo que implica que deberiamos incluir este tipo de carta en la fase de aprendizaje, pero esto nos llevaría a un problema de casuística, el cual, en parte, se intenta resolver mediante el proceso de "*difuminado*".

Algunos de estos problemas se están corrigiendo mediante el uso de vectores de características analógicos, que van a representar el grado de pertenencia de cada objeto físico a una determinada zona difusa del documento, así el vector binario dado en la tabla 1 se transforma en:

y0	y1	y2	y3	y4	y5	y6	y7	x0	x1
0	0.9	0.099	0	0	0	0	0	0	0
x2	x3	x4	p0	p2	p6	p9	l0	l2	l6
0.2	0.8	0	0	1	0	0	1	0	0

Tabla 7: Vector analógico de características

Aunque, para su total resolución se deberá utilizar conocimiento sobre la disposición o jerarquía de los componentes y de la carta en su conjunto. Dicho en otras palabras, será necesario un reconocimiento condicionado.

Cabe destacar que entre las cartas utilizadas en esta última fase se encontraban cartas del mismo autor y estilo que alguna de las utilizadas en la fase de aprendizaje del módulo neuronal, y el resultado ha sido óptimo, pues no sólo los componentes se han clasificado correctamente, sino que han sido clasificados de acuerdo a su estilo/autor.

7 Conclusiones

El proceso de reconocimiento de un documento, es decir, la obtención de los componentes lógicos más significativos de un documento, a partir de la información física del mismo (etiquetado lógico), es un proceso necesario para la gestión automatizada de dichos documentos, su intercambio en sistemas abiertos y su almacenamiento de acuerdo a normas que permitan una fácil gestión y manipulación. Además, para alcanzar estos objetivos de intercambio, el documento debe tener una organización interna y un formato de intercambio lo más estandarizado posible, pues en caso contrario, el documento puede ser irrecuperable.

El método presentado se basa por un lado, en la caracterización de los bloques físicos de información pertenecientes a un documento por medio de aspectos geométricos y cuantitivos: líneas y palabras, sometidos a un proceso de "*difuminado*",

Y por otro lado, en la utilización de un sistema de redes neuronales artificiales basadas en ART, es un método que permite obtener buenos resultados en el etiquetado lógico de los bloques de información que conforman un documento cualquiera. La utilización de esta red neuronal va a permitir un continuo "*aprendizaje*" de nuevos tipos de componentes y documentos, es decir, es capaz de aumentar su conocimiento, a diferencia de otros métodos simbólicos-heurísticos posibles para solucionar el mismo problema.

El reconocimiento de los componentes de un documento, es decir, de aquello que le va a definir o caracterizar, va a permitir:

- Utilizar esta información parcial de forma independiente del resto del documento, para procesos de gestión como: registros de bases de datos, automatización de procesos de distribución, etc...
- Clasificar el documento en conjunto como perteneciente a una determinada clase e incluso dentro de estas a alguna subclase si hubiera lugar.

La utilización tanto de lógica difusa como de redes neuronales, se adapta perfectamente al problema de reconocimiento, debida a la gran variabilidad que presentan los documentos dependiendo del autor, el estilo, etc... y a la necesidad de un continuo aprendizaje con respecto a nuevas clases de documentos y de componentes de los mismos.

8 Agradecimientos

Deseamos agradecer al Grupo de Reconocimiento y Procesamiento de Documentos de la E.T.S.I.T y al Grupo de Redes Neuronales de la E.T.S.I.I. de la Universidad de Valladolid su ayuda y apoyo en la elaboración de esta comunicación.

Referencias

[Andre-89] André J., Furuta R., and Quint V. *Structured documents.* Cambridge University, 1989.

[Appelt-91] Appelt W. *Document Architecture in Open Systems: The ODA Standard.* Springer-Verlag, Heidelberg, Germany, 1991.

[Bleisinger-92] Bleisinger R., Dengel A., and Hoch R. Oda - based modeling for document analysis. Technical report, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Kaiserslautern, Deutschland, November 1991.

[Carpenter-92] Carpenter G., Grossberg S., Markuzon N., and Reynolds J. Fuzzy artmap: A neural network architecture for incremental supervised learning of analog multi-dimensional maps. *IEEE Transactions on Neural Networks*, 3:698–713, 1992.

[Chabrier-90] Chabrier J. and Guillemin J.M. Oda: une plataforme pour la documentation technique des enterprises. *Genie Logiciel & Systemes Experts. Les outils logistiques: gestion de configurations, documentation & gestion des projets*, (21):54–61, Decembre 1990.

[Dengel-88] Dengel A. and Barth G. High level document analysis guided by geometric aspects. *Int. Journal of PR and AI*, 2(4):641–655, 1988.

[Dengel-92] Dengel A., Bleisinger R., and Hoch R. *Iloda*: The paper interface to oda. Technical report, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Kaiserslautern, Deutschland, February 1992.

[Dengel-94] Dengel A. About the logical partitioning of documents. In *Proceedings of International Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, USA, April 1994.

[Dimitriadis-95] Dimitriadis Y.A. and López Coronado J. Towards an art-based mathematical editor that uses on-line handwritten symbol recognition. *Pattern Recognition*, 28, June 1995.

[Fujihara-93] Fujihara H., Babiker E., and Simmons D.B. Fuzzy approach to document recognition. In *Procedings of the Second IEEE International Conference on Fuzzy Systems*, pp: 980–985, San Francisco, USA, March-April 1993.

[Grossberg-80] Grossberg S. How does a brain build a cognitive code? *Psychological Review*, 87, 1980.

[Grossberg-82] Grossberg S. *Studies of Mind and Brain: Neural principles of learning, perception, development, cognition and motor control.* Reidel Press, Boston, USA, 1982.

[Higgins-91] Higgins C. A. and Ford D. M. Stylus driven interface - the electronic paper concept. In *Proceedings of ICDAR' 91*, Saint Malo, France, 1991.

[Horak-85b] Horak W. Office document architecture and office document interchange formats: Current status of international standadization. *Computer*, (18):50 – 59, October 1985.

[Kosko-92] Kosko B. *Neural Networks and Fuzzy Systems.* Prentice Hall, 1992.

[Kreich-93] Kreich J. Robust recognition of documents. In *Proceedings of ICDAR' 93*, Tksuba, Japan, 1993.

[Le-94] Le D. X. and Thoma G. R. Document classification using connectionist models. In *Procedings of ICNN94*, pp: 3009–3014, Orlando, USA, June 1994.

[Marovac-90] Marovac N. Document structures and document recognition. *Bigre*, (69), Mai 1990.