Modelización automática de textos en castellano

Daniela López De Luise y Mariana Soffer.

***Abstract*— Este trabajo se centra en aspectos de la manipulación automática de textos en castellano y estudia el uso de una métrica de ponderación denominada p_o, parte de un prototipo funcional llamado WIB. Se mostrará que las ponderaciones, generadas con criterios morfosintácticos, permiten una agrupación en categorías de palabras por conjuntos difusos y que pueden funcionar como modelización de los contenidos textuales. Se estudia el texto de una página Web seleccionada al azar y se le calcula p_o. Las palabras así procesadas se agrupan sobre la base de p_o usando distintas algorítmicas. También se estudia el significado de estas agrupaciones en términos del contexto y características morfosintácticas de las palabras.**

***Index Terms*— Fuzzy sets, Information retrieval, Multilayer perceptrons, Text processing.**

I. Introduccion

Básicamente, una lengua se expresa muy heterogéneamente conforme el nivel de cultura, temática, situación geográfica, etc. [1]. Es frecuente la necesidad de manipular textos en la WEB escritos en cierta lengua. Para ello resulta provechoso establecer una caracterización en la forma de redactar sentencias por parte de los distintos autores y la posibilidad de reflejar parte del contenido conciso y eficientemente de manera automática así como los niveles de significación a nivel de sentencias.

Los tratamientos con manejo de niveles de representatividad no son nuevos y fueron propuestos ya por varios autores [2], [12], [13]. Algunos modelos propuestos parten de un conjunto de documentos D= $\{d_1, d_2,..., d_m\}$, que constituyen la base de datos sobre la cual se desean realizar consultas. Se suelen asociar manualmente ciertas palabras (denominadas normalmente *labels*), seleccionadas de un conjunto predeterminado. A cada una de estas palabras se les establece manualmente niveles de representatividad en referencia al texto asociado. Cuando se realiza una consulta (denominada también *query*), se utilizan comparaciones contra estas etiquetas y sus niveles de representatividad a fin de filtrar los documentos que más probablemente sean relevantes. A tal fin se definen una serie de modelos matemáticos con tratamientos difusos alternativos [5].

D. López De Luise preside el centro de investigaciones IT-Lab y al grupo AIGroup, en la Universidad de Palermo. Mario Bravo 1050. C1175ABT. Buenos Aires. Argentina (phone: +54 11 – 5199 4520; e-mail: aigroup@palermo.edu).

Las propuestas con operadores difusos, han probado ser tan competitivas y eficientes en la etapa de recuperación como las algorítmicas clásicas [12]. Si bien los primeros trabajos comenzaron con la manipulación difusa (con lógica difusa) de los términos de una consulta [13] (para evaluar grados de necesidad de requerimientos durante la formulación de las consultas), se ha mostrado que el modelado con lingüística difusa ayuda a mejorar los resultados de las búsquedas [6] Algunas propuestas incluso, van más allá: en [12] se propone una combinación de palabras cuantificadas con sentencias ponderadas usando operadores de cuantificación semi-difusos. Estas ideas se aplicadan también a la fase de recuperación de información desde la Web (denominada IR, Information Retrieval) para mejorar el funcionamiento de un CBR (Case Based Reasoner) [3].

El prototipo WIB (Web Intelligent Browser) realiza una manipulación de textos dual: a nivel de palabras y sentencias. Para ello genera una ponderación denominada p_o [7]. La misma ayuda a calificar automáticamente las palabras castellanas dentro de un texto [8] y es usable para perfilar el tipo de escritura. El sistema considera al valor p_o, derivado por WIB, no sólo como apoyo al tratamiento sino también como criterio de elaboración de ciertas estructuras dentro de la base de datos del prototipo. Todo esto es útil para soportar actividades de navegación o búsqueda. A continuación se detallan las características de esta métrica y se profundiza su utilidad potencial.

A. Principios para Generar p_o

Los textos capturados por el prototipo WIB son convertidos a un conjunto de palabras y estas son asociadas a ciertos campos que describen sus características morfosintácticas, denominadas *EBH* (Elemento Básico Homogenizado). Los *EBH* correspondientes a una misma sentencia son estructurados en una E_{ci} (Estructura de Composición Interna). Las E_{ci} de un mismo texto son resumidas a unas E_{ce} (Estructura de Composición Externa), muy compactas, que las representa. La función encargada de la promoción de términos desde una E_{ci} a una E_{ce}, trabaja según unos principios básicos para obtener los valores de ponderación p_o correspondientes a las E_{ci} y E_{ce}. En términos generales, esos principios son:

-Los valores de p_o deben caracterizar situaciones morfosintácticas sencillas a nivel palabra (ej. longitud, cantidad de vocales fuertes, está resaltada, etc).
-Los valores deben tener dominio [-2.0 ; +2.0].
-La mayoría de los valores serán 0.0

-Las ponderaciones de términos deben ser cercanas a 0.0 cuando la palabra asociada no modifica radicalmente la palabra / sentencia involucrada.
-Las ponderaciones de términos deben ser lejanas a 0.0 cuando la palabra asociada modifica radicalmente la palabra / sentencia involucrada.
-La ponderación de una sentencia se traduce en una ponderación p_o^{Eci} al nivel de E_{ci}, donde se combinan los p_o. La forma específica es esta combinación variará según el comportamiento del sistema, definido por el estado del mismo. Siguiendo lo presentado en [7], actualmente se usa la fórmula:

$$p_o = \frac{p(w_0)}{2^n} + \sum_{i=1}^{n} \frac{p(w_i)}{2^{n+1-i}} \quad (1)$$

Donde se considera *n* como la cantidad de palabras dentro de la sentencia actual del texto, *w* es el *EBH* de cierta palabra, y *p(w)* refiere siembre a la heurística de ponderación aplicada a la *EBH* denominada aquí con *w*.
-La ponderación de un párrafo se traduce en una ponderación p_o^{Ece} a nivel E_{ce}, resultante de todas las p_o^{Eci} dentro de la misma E_{ce}. Se elige la p_o^{Eci} más *optimista*[1].

B. Características Distintivas de p_o

El manejo p_o en WIB tiene como objetivo el procesamiento de palabras con agrupaciones difusas, pero existe una marcada diferencia con el procedimiento tradicional que sustenta la generación de los conjuntos difusos:
-No se predefine un conjunto específico de palabras sino que cualquiera de las que figuran en un texto son candidatas a ser *EBH* relevantes. Las mismas palabras del texto (y no etiquetas predefinidas) son detectadas y derivadas en elementos homogeneizados procesados y sopesados automáticamente.
-La ponderación de un *EBH* se proyecta a toda la estructura E_{ci}, que representa la sentencia que engloba a la palabra representada por ese elemento homogeneizado, y se combina con la ponderación del resto de las palabras de la frase.
-Las ponderaciones se derivan de la posición y conformación de las palabras (morfosintaxis).
-Las ponderaciones no se restringen al rango [0..1]. Pueden abarcar [-2..+2].

El estudio detallado del modelo difuso y su implementación excede el alcance de este trabajo, pero se estudian algunas características básicas de los conjuntos difusos que se conformarían con WIB.

C. Comportamiento estadístico de p_o

En esta sección se transcriben las características estadísticas del tratamiento de ponderaciones ya expuesto en [8]. Hay dos clasificaciones fundamentales de los textos a trabajar:

1) Tipos de Texto: Los textos escritos tienen una estructura típica que todo autor debiera seguir para cubrir las expectativas de su interlocutor. En este sentido, se propone diferenciar al menos los siguientes estilos en la Web: literario, técnico y mensajes.

2) Perfiles de Narración: Los textos escritos estimulan al autor de los mismos en cierto sentido según el objetivo perseguido. En este sentido, se propone diferenciar al menos los siguientes estilos en la Web: foro, índice de Web (página que sólo sirve a los efectos de indexar una serie de otras páginas), documento y blog (recopilación cronológica de artículos).

Según [8], la métrica p_o es una nueva métrica *invariante* al tamaño de un documento y tipo de texto, que permitiría diferenciar el perfil del narrador, y que mediría razonablemente bien la relevancia relativa de las frases dentro de un documento.

D. El Modelo Difuso en WIB

El esquema aquí propuesto interpreta que, sobre la base de lo expuesto en [8] durante la redacción existen perfiles de narración y tipos de sentencias (más o menos representativas y de menor o mayor calidad). Una eventual búsqueda o consulta, podrá usar lógica difusa sobre las ponderaciones y determinar si la frase merece ser posicionada mejor en un listado de respuestas candidatas debido a la calidad probable de su redacción, obtenida en función del perfil y del tipo de representatividad del texto en cuestión.

Es importante destacar que la problemática de recuperación de textos con conjuntos difusos en este trabajo es ligeramente diferente a otras aplicaciones donde cada punto del conjunto es un documento, en cambio aquí cada punto representa una palabra de una cierto documento (sea una página Web o no). Dada la manipulación que se realiza, la unidad más importante resulta ser la sentencia y no el documento.

Dos son los aspectos principales a considerar:

1) Manejo *Difuso en la Base de Documentos:* A fin de realizar el manejo difuso en la base de documentos, los siguientes pasos son realizados por el prototipo:
-Los documentos son almacenados.
-Se definen las estructuras E_{ci} y E_{ce}.
-Se deriva el grado de significación de las sentencias de cada documento, asignando un valor de p_o correspondiente a E_{ci} y E_{ce} (palabras y sentencias respectivamente).

Los problemas que surgen en el manejo difuso suelen relacionarse con [13]:
-La descripción matemática de un cuantificador adecuado.
-La adecuación de cierto valor numérico como resultado de la aplicación de los operadores difusos, ante cierta búsqueda concreta.

En este trabajo, precisamente se evaluará si realmente la definición de p_o es suficiente para obtener los conjuntos difusos sobre los que deben aplicarse las operaciones correspondientes. El tratamiento deberá completarse implementando el uso adecuado de esta ponderación para filtrar los resultados ya sea en función de una consulta o de una navegación.

La segunda problemática será controlada por una serie de métricas de calidad, evaluadas por el sistema probablemente basadas en lo presentado en [7].

[1] El concepto de optimista variará según el comportamiento del sistema, definido por el estado del sistema.

2) Manejo Difuso en las Consultas: El uso de lógica difusa introduce el manejo de la ambigüedad propia de la lengua natural cuando se usa como parte de la *query* o consulta. La definición y tratamiento al respecto, excede el objetivo de este trabajo, pero para completitud del tratado, a continuación se detallan algunas apreciaciones que orientan acerca del tema.

En los tratamientos difusos de las consultas, se suele requerir al usuario una apreciación numérica que represente sus prioridades para cada término dentro de la búsqueda. Pero muchos usuarios no están habilitados para dar una valoración numérica de sus necesidades de información. Por eso se categorizan lingüísticamente (con términos tales como muy importante, importante, etc.) [5]. El estudio y manejo de términos sopesados en la consulta, manifestó la necesidad de definir un nuevo operador de agregación eficiente llamado LOWA (Linguistic Ordered Weighted Averaging). Su uso se extiende también a manejos de ambigüedades similares a los de toma de decisión [4]. En estos problemas es más complicado pues coexisten varias opiniones en forma de relaciones lingüísticas de preferencias. Por ello debe proveerse un adecuado operador de agregación además de los tradicionales para manipular conjuntos difusos, y asì asegurar un resultado racional. De lo aquí expuesto, los pasos a seguir por parte del sistema WIB seràn:

-Definir una secuencia corta de *EBH* (7 ó 9) que califican lingüísticamente el documento. Obsérvese que en otras propuestas se utilizan etiquetas prefijadas que reflejan la importancia de las palabras expresadas. Por Ej.{null, muy_bajo, bajo, medio, alto, muy_alto}.
-Definir las funciones de pertenencia de cada etiqueta. En este paso es donde se realiza el agrupamiento difuso.
-Definir los operadores: operador negación, comparación y agregación.
-Reducir las consultas a un conjunto de términos a buscar.

El resto de este trabajo se organiza como sigue: En la sección II se presentan los datos para un caso de estudio, se evalúan las características de los datos a través de sus estadísticas descriptivas, agrupación de las palabras del texto en conjuntos no difusos y difusos para mostrar la importancia de su uso y finalmente en la sección III se describen las conclusiones y trabajo a futuro.

II. Estudio de Caso: Uso de p_o

Se toma como base del presente estudio el texto original de una página WEB en castellano dedicada al tema de las orquídeas (url: orquidea.blogia.com temas-que-es-una-orquidea-.php). El mismo fue convertido en un conjunto de palabras y símbolos. Algunas palabras fueron eliminadas por carecer de contenido significativo (por ejemplo los artículos). Luego, el sistema genera un *EBH* (Elemento Básico Homogenizado) [9] para cada una de las palabras, los cuales se ordenaron en un grafo orientado (o E_{ci}) construido según criterios morfo-sintácticos [11]. Cada *EBH* consta de una serie de *descriptores* o campos que describen morfo-sintácticamente la palabra y la sentencia en la que se encuentra. Con criterios por el estilo WIB detecta y clasifica como importantes algunas de éstas (son entonces denominadas *indicadora* del contenido del texto).

También determina, según sus características morfo-sintácticas, si la palabra tiene un peso relativo positivo o negativo en el significado de la frase que la contiene (ponderación p_o). Según [8], dicho peso permite determinar automáticamente si la palabra pertenece o no a la descripción del tema principal del texto.

Tomados los datos consistentes en:

ID: palabra originalmente extraída de la página.

poID: p_o asignado a la palabra reconocida como *ID*, cuando se construye la E_{ci}.

poECI: p_o derivado al E_{ci} que contiene a *ID*.

indicadora: determinación de si el *EBH* al que pertenece *ID* está en una porción de sentencia con información representativa del texto.

A. Estudio de Estadísticas Descriptivas de Datos

Se dedujo, para las 60 E_{ci} (oraciones) del texto html original, cuántas palabras componen la E_{ci} (campo *long*), el peso asignado a la E_{ci} (campo *poECI*) y la cantidad de palabras *indicadora* (campo *cantIndic*). Se Aplicó agrupamiento (o *clustering*) con el algoritmo *Expectation Maximization*[2][3] (EM). El mejor *likelihood*[4] es obtenido cuando no se considera *cantIndic* y es el peor cuando no se considera *poECI* (ver Tabla 1), por lo que se puede inferir que *poECI* es un buen discriminante de subconjunto de E_{ci} en estudio.

Estudiando las E_{ci} en función de su longitud se observa que:
-Hay una tendencia en cluster 4 a valores *long* pequeños.
-En los cluster 1 y 3 aparenta ser intermedio el valor *long* y en el cluster 2 es mayor.

Cuando se visualizan las E_{ci} en función de *poECI*:
-Los cluster 4 y 3 se mantienen con valores aproximados.
-El Cluster 1 tiene gran dispersión pero siempre por debajo del valor promedio de *poECI*.
-El Cluster 2, por el contrario tiene dispersión por encima del promedio.

TABLA 1.
LOG LIKELIHOOD SEGÚN LOS CAMPOS.

campos	likelihood	cantidad clusters	detalle clusters
long poECI cantIndic	0.4657	4	0 1 (5%) 2 16 (76%) 3 3 (14%) 4 1 (5%)
long cantIndic	-4.37129	2	0 9 (43%) 1 12 (57%)
poECI cantIndic	0.09554	2	0 19 (90%) 2 2 (10%)
long poECI	1.06369	4	1 3 (14%) 2 2 (10%) 3 1 (5%) 4 15 (71%)

[2] Algoritmo que estima parámetros de un modelo probabilístico por màximo likelihood. El modelo depende de variables latentes no observadas. EM alterna entre un paso de (E)xpectation (calcula la esperanza poblacional estimando las variables latentes como si fueran observadas), y un paso (M)aximization, que maximiza el likelihood hallado en el paso anterior.

[3] con el framework gratuito de data mining WEKA.

[4] Estadístico que indica cuán bien conformados están las agrupaciones. Su valor deber ser el máximo posible.

Esto podría indicar que *cantIndic* no representa algo en particular, pero las longitudes de las E_{ci} son una manera de clasificar el conjunto de palabras y *poECI* es la otra. Para ver si inciden mutuamente se realizó un àrbol J48[5] (Fig. 1), el cual ordena cerca de la raíz los campos decisores y elimina el tratamiento de aquellos que no hacen a la discriminación de cuál cluster debe establecerse para determinado dato. Se puede ver que en general *poECI* es factor de decisión cuando los valores de *long* es superior a 7 palabras. Ya en [10] se muestra que a nivel de *poECI* la métrica se comporta como un discriminante bastante confiable de palabras pertenecientes al tema principal. En cambio el nivel de *poID*, su valor y significado será proveer base al clustering difuso.

B. *Estudio de Agrupaciones no Difusas*

Se realiza primeramente una clasificación por *EM* (usando WEKA), con una cantidad máxima de 100 iteraciones, y dejando que el software decida la cantidad de clusters usando el mínimo error por validación cruzada y probando con qué campos es conveniente realizar la clasificación. La cantidad de clusters resultantes de aplicar *EM* es 5, con un *log likelihood* máximo cuando se consideran todas las variables (Tabla 2).

Los conjuntos (o clusters) creados por *EM* tienen una distribución especial (ver Tabla 3): algunos parecen reflejar sentencias ponderadas siempre negativamente, otros lo contrario. Además, de ambos grupos parecen surgir subgrupos menores con algún rasgo distintivo (por ejemplo tener *indicadora* o palabras siempre ponderadas con valores no nulos). Realizada la verificación con la herramienta *sizing* que provee el software Nuclass[6], el mismo hace una estimación de la cantidad de clusters en función de la distancia promedio en los clusters (ver Fig. 2) usando *Mean Squared Error*[7]. Se confirmó la existencia de 5 agrupaciones.

Con una red neuronal Perceptrón multicapa[8](MLP) del software estadístico Nuclass, se clasificaron las palabras considerando los mismos campos utilizados para *EM*: *ID*, *poID*, *indicadora* y *poECI*. Para esta red se utilizó una arquitectura de 3 nodos de entrada y 3 capas ocultas de 3 nodos.

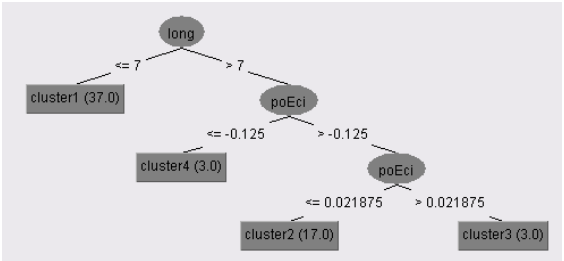

Fig. 1. **J48 entrenado con clusters *EM*.**

[5] Implementación del algoritmo de inducción C4.5 en WEKA[15].
[6] Software para clasificaciones no lineales de la Univ. de Arlington (Texas)
[7] Estadístico para medir cuánto difiere un estimador de los valores reales.
[8] Una Red neuronal artificial (ANN), frecuentemente denomiada Red Neuronal (NN), es un modelo matemático o computacional basado en las redes neuronales biológicas. Consta de un grupo de neuronas artificiales y procesa información con un método conexionista.

TABLA 2.
VALORES DE LOG LIKELIHOOD EN *EM*.

variables	log likelihood
todas	3.53
sin ID	3.17
sin indicadora	-1.76
sin poECI	-5.09
sin poID	-2.06
sin id ni poID	1.45
sin indicadora ni poID	-3.63
sin poECI ni id	0.40
sin poECI ni poID	-5.78

TABLA 3.
CONFORMACION DE CLUSTERS POR *EM*.

cluster	indicadora	poECI	poID
0 (18%)	nunca	poEci >0 mucha dispersion	poID ≅ 0
1 (2%)	algunas	poEci >=0	poID ≅0
2 (4%)	nunca	poEci <0	poID ≅0
3 (68%)	algunas	poEci ≅0	poID ≅ 0
4 (1%)	nunca	poEci<=0	poID<>0 mucha

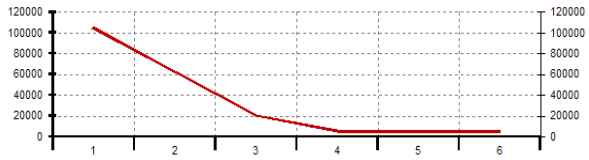

Fig. 2. MSE Vs cantidad de clusters.

Utilizando los valores propuestos por la herramienta se configuró con 1 capa interna de 6 nodos. La tasa de aprendizaje (learning rate) es de 0.3, y la inercia (momentum): de 0.2. Se repitió la medición con una estructura de 3 capas con 3 nodos. El sistema evolucionó mucho más lento (3000 ciclos o *epochs* contra 500) y el error obtenido por entrenar con el 67% y validar con el 33% de los datos fue de 0.0000502 (contra 0.0000034 del resultado anterior que tardó 500 ciclos). Se puede observar que el *Kappa statistics*[9] y las medidas de error son similares.

Respecto al nivel de importancia en la discriminación de los clusters, se usò nuevamente un árbol J48. En la Fig. 3 se puede apreciar que el hecho de ser *indicadora* o no es decisivo, seguido por el valor *poID* de la palabra. El valor *poECI* parece no incidir en la decisión de la asignación del cluster.

Recorriendo las agrupaciones, se puede decir que el cluster 1 tiene la mayoría de las palabras del texto, el cluster 2 tiene pocas palabras, pero ellas suelen tener con connotación negativa.

En la tabla Tabla 4 se puede apreciar un resumen de lo expuesto y algunas de las palabras pertenecientes a cada cluster. Si se estudia la distribución de las clases (es decir los números de cluster) entre las palabras y sentencias puede observarse nuevamente un cierto cluster como base y cambios ocasionales.

[9] Coeficiente de Cohen que mide la confiabilidad de un predictor respecto a un clasificador azaroso.

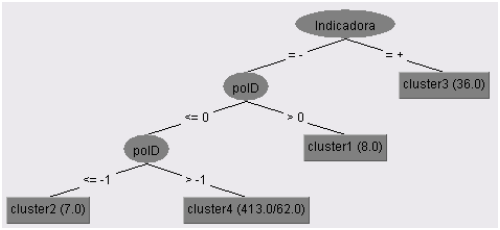

Fig. 3. **J48 entrenado con los clusters MLP.**

TABLA 4.
CARACTERÍSTICAS DE LOS CLUSTERS

cluster	característica	ejemplos
0	mucha cantidad de palabras	planta, crece, rocas
1	mínima cantidad palabras superfluas	abundantes
2	pocas palabras connotación negativa u opuesta a algo	no, distinto, distinguir, destacan
3	mucha cantidad de palabras son palabras con mucho significado o palabras ubicadas en frases importantes	orquídeas, estructuras, polen son: en la sentencia .."aunque SON mas abundantes en los trópicos también existen en climas templados.."
4	pocas palabras son calificativos	muy, muchas

En la Fig. 4 se representan 3 sentencias seleccionadas al azar (tres E_{ci}), se puede ver cada punto como una palabra (representada en un *EBH*). En general, los párrafos (representados como una secuencia ordenada de E_{ci}) muestran también ese tipo de asignación a sus E_{ci} existiendo escasas asignaciones a clusters alternativos (Fig. 5).

C. *Estudio de Agrupaciones Difusas*

En esta sección se usa la cantidad de agrupaciones hallada en la sección anterior para formar grupos difusos. Esto se realiza por dos razones:

-debido a que las mejores estadísticas resultan con 5 agrupaciones.

-para poder estudiar comparativamente el significado de los grupos formados por p_o, la incidencia de *poID*, y la posibilidad de que los grupos difusos representen las palabras con características específicas dentro de los documentos.

Realizando el estudio de agrupaciones difusas (o fuzzy sets) con los datos, surge la distribución en 5 grupos que se muestra en la Tabla 5. Se puede apreciar que la disposición de conjuntos realizada por *EM* y por una red neuronal multicapa (Multi Layer Perceptron, MLP) son distintas que para el agrupamiento difuso. Ahora las palabras concentradas en el cluster 1 parecen abrirse en 3 conjuntos. El conjunto 3 contribuye a 2 conjuntos difusos y el resto se concentra en el conjunto difuso 5. Esto parece indicar un criterio distinto de agrupación que no responde a los otros métodos. Las gráficas Fig. 6 y Fig. 7 indican ahora las secuencias de asignación de conjuntos en 3 E_{ci} y en 3 párrafos al azar respectivamente.

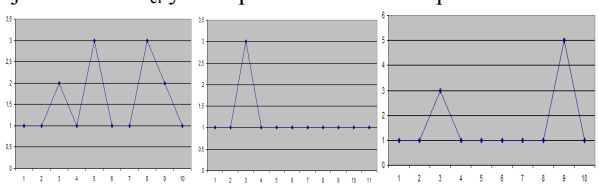

Fig. 4. **secuencias de clusters en las E_{ci}**

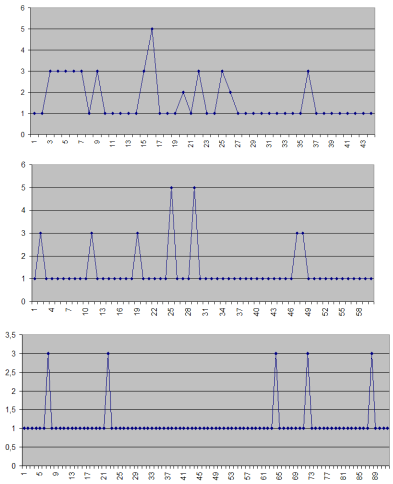

Fig. 5. **secuencias de clusters en porciones del texto**

TABLA 5.
DISTRIBUCIÓN COMPARADA DE CLUSTERS SEGÚN FUZZY SETS Y *EM*-MLP

	Fset (%)	EM-MLP(%)
cluster 1	8.8	89,01
cluster 2	8.83	1,51
cluster 3	71.3	7,76
cluster 4	5.25	1,29
cluster 5	4.74	0,43

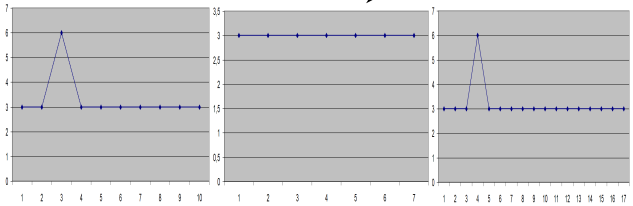

Fig. 6. **ejemplo de secuencias de agrupaciones difusas en las E_{ci}**

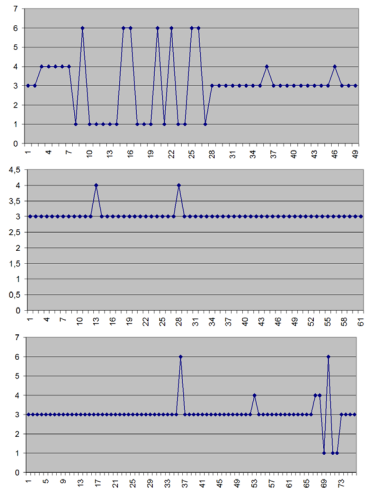

Fig. 7. **secuencias de conjuntos difusos en porciones del texto**

En términos generales las E_{ci} muestran menor variación de asignación usando conjuntos difusos. De hecho en una de las E_{ci} de la Fig. 6, el grupo es siempre el mismo. Para comprender esta nueva disposición se estudiaron nuevamente los contenidos. En la Tabla 6 se describen las características importantes de cada grupo o cluster.

TABLA 6.
POSIBLE SIGNIFICADO DE CLUSTERS DIFUSOS.

Cluster	Características
1	no indicadoras
	poECI<>0
	poID =0
	ej: semillas, esperma, fruto, géneros
	parecen indicar palabras que aportan al texto de manera esencial. Son generalmente la linea base del tema
2	no indicadoras
	poECI menores a cluster 1 en valores absolutos (pero positivos todos)
	poID =0
	ej: herbàceas, perennes, Orchidaceae, Liliopsida
	parecen indicar palabras que aportan un poco más a las bases del cluster 1
3	no indicadoras, son mayoritarias
	poECI menores a cluster 2en val absoluto (mayormente 0)
	poID =0
	ej: tres, sirve, atracción, modificado
	parecen indicar palabras que aportan al texto de manera general. Hacen al tema pero completándolo.
4	indicadoras
	poECI comparables con cluster 3 (mayormente 0).
	poID =0
	ej: orquídea, es, son, principales
	palabras que no contienen muy poca información por sí mismas sino por su ubicación en la sentencia (paper ....)
5	Si poID <>0 entonces son no indicadora
	Si poID =0 entonces son indicadora
	poECI <> 0 (valores comparables a cluster 1,2,3)
	ej: no, distintos, muy muchas, abundantes
	palabras importantes posicionalmente que inciden en el significado de la frase (atenuando o reforzando concepto), aunque carecen de significado por sí solas

Examinando los conjuntos de la Tabla 4 en comparación con la Tabla 6, se puede decir en términos generales que:

-usando los mismos campos, ahora se discriminan màs los que contribuyen posicionalmente. Ahora son tres categorías: con mayor significado (cluster 1), un aporte adicional (cluster 2) y aporte secundario (cluster 3). Antes sólo se distinguían las que contribuían en términos generales (cluster 3 de Tabla 6) y resultaba en un conjunto mucho más extenso.

-las palabras realmente modificadoras del significado de las sentencias son tratadas como un caso especial igual que antes. (cluster 5 de Tabla 6 y cluster 4 de la Tabla 4)

-las palabras con significación propia son discriminadas y recategorizadas en dos tipos según su mayor o menor aporte de significado (cluster 1 y 2)

-las palabras que sólo completan el texto para su conformación a las reglas de la gramática del lenguaje, son discriminadas (cluster 3).

Por lo antedicho, es razonable suponer que en el ámbito de sentencias se puede hallar un significado determinado en términos de conjuntos difusos procesables como lógica difusa (o fuzzy logic).

III. CONCLUSIONES Y TRABAJO A FUTURO

Se mostró que p_o genera conjuntos (en el caso ejemplo fueron 5), que representan agrupaciones consistentes de las palabras. Los campos *poID* e *indicadora* son más importantes para el procesamiento a nivel de palabras, coadyuvando a una mejor discriminación en clusters difusos. A nivel de párrafos del texto en general, en cambio, pesan màs los campos *poECI* y la longitud *long* de las E_{ci}.

El tratamiento deberá completarse implementando el uso adecuado de esta ponderación para filtrar los resultados ya sea en función de una consulta o de una navegación y la definición de operadores difusos (o fuzzy operators) adecuados.

REFERENCES

[1] Bargalló M., Forgas E., Garriga C., Rubio A.”Las lenguas de especialidad y su didáctica”. J. Schnitzer Eds. Universitat Rovira i Virgili. Tarragona, cap. 1 (P. Schifko, Wirtschaftsuniversität Wien), pp. 21-29. 2001.

[2] Delgado M., Sánchez D., Serrano J.M., Vila M.A.”A Survey of methods to evaluate quantified sentences”. Mathware and Soft Computing. vol. 7 (2 - 3): pp. 149 - 158. 2000.

[3] Jackzynski M., Trousse B.”Fuzzy Logic for the retrieval step of a Case-Based Reasoner”. Proc. of the Second European Conference on Case-Based Reasoning, pp. 313-322. 1994.

[4] Herrera F., Herrera Viedma E., Verdegay J. L.”Aggregating linguistic preferences: properties of lowa operator. Proc. of VI IFSA World Congress, Sao Paulo. Brazil. Vol. II. pp. 153 - 157. 1995.

[5] Herrera Viedma E., López Herrera A. G., Luque M., Porcel C. “A Fuzzy Linguistic IRS Model Based on a 2-Tuple Fuzzy Linguistic Approach. V Congreso ISKO. pp. 148 - 157. España. España. 2001.

[6] Herrera Viedma E., Pasi G.”Approaches to access information on the Web: recent developments and research trends”. Fuzzy. Proc. International Conference on Fuzzy Logic and Technology (EUSFLAT 2003), pp. 25–31, Zittau (Germany). 2003.

[7] López De Luise M. D., Agüero M. J.”Aplicabilidad de métricas categóricas en sistemas difusos”. IEEE Latin America Magazine. Vol. 5. Issue 1. Editor Jefe José Antonio Jardini. 2007.

[8] López De Luise M. D.”A Metric for Automatical Word Categorization”. In Advances in Systems, Computing Sciences and Software Engineering. Proc. Of SCSS 2007. Springer. Tarek Sobh & Khaled Elleithy Editors. Aceptado para publicación. 2007.

[9] López De Luise M. D. “A Morphosyntactical Complementary Structure for Searching and Browsing”. In Advances in Systems, Computing Sciences and Software Engineering. Proc. Of SCSS 2007. Springer. Tarek Sobh & Khaled Elleithy Editors. Pp. 283 – 290. 2005.

[10] López De Luise M. D. “Induction Trees for Automatic Word Classification”. Anales XIII Congreso Argentino de Ciencias de la Computación (CACIC07). Corrientes. Argentina. Pp. 1702. 2007.

[11] López De Luise M. D.. “Una representación alternativa para textos”. Ciencia y Tecnología. Colección C&T. ISSN 1850 0870. 2007-4. Buenos Aires, Argentina. Pp. 119 – 130. 2007.

[12] Losada D. E., Barro Ameneiro S., Bugarín Diz A. J., Díaz Hermida F.”Experiments on using fuzzy quantified sentences in adhoc retrieval”. ACM Symposium on Applied Computing. vol. 0, pp. 1059 - 1066. 2004.

[13] Losada D. E., Díaz Hermida F., Bugarín A. “Semi-fuzzy quantifiers for information retrieval”. International Journal of approximate reasoning. vol. 34, pp. 49-88. 2003.

[14] Morillas Raya A.”Introducción al análisis de datos difusos”. Depto. de Estadística y Econometría. Univ. de Málaga. España. www.eumed.net/libros/2006b/amr/. ISBN: 84-689-9208-2. 2006.

[15] Witten I. H., Frank E. “DataMining - Practical Machine Learning Tools and Techniques”. 2nd ed. Morgan Kaufmann Publishers. 2005