risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido / Recibido: 14/10/2011
Aceitação / Aceptación: 09/12/2011

Clasificación de textos en lenguaje natural usando la Wikipedia

Jose María Quinteiro-González [1,2], Ernestina Martel-Jordán [1,2], Pablo Hernández-Morera [1,2], Juan A. Ligero-Fleitas [1], Aaron López-Rodriguez [1]

{jquinteiro,emartel,pablo,jligero,alopez}@iuma.ulpgc.es

[1] IUMA Sistemas de Información y Comunicaciones. División Tecnología de la Información, Universidad de Las Palmas de Gran Canaria, Campus Universitario de Tafira, 35017 - Las Palmas de Gran Canaria, España.

[2] Departamento de Ingeniería Telemática, Universidad de Las Palmas de Gran Canaria, Campus Universitario de Tafira, 35017 - Las Palmas de Gran Canaria, España.

Resumen: La clasificación de textos, en entornos en los que el volumen de datos a clasificar es tan elevado que resulta muy costosa la realización de esta tarea por parte de humanos, requiere la utilización de clasificadores de textos en lenguaje natural automáticos. El clasificador propuesto en el presente estudio toma como base la *Wikipedia* para la creación del *corpus* que define una categoría mediante técnicas de *Procesado de Lenguaje Natural* (PLN) que analizan sintácticamente los textos a clasificar. El resultado final del sistema propuesto presenta un alto porcentaje de acierto, incluso cuando se compara con los resultados obtenidos con técnicas alternativas de *Aprendizaje Automático*.

Palabras clave: Categorización de textos; Wikipedia; tf-idf; Aprendizaje Automático; Procesado de Lenguaje Natural.

Abstract: Automatic Text Classifiers are needed in environments where the amount of data to handle is so high that human classification would be ineffective. In our study, the proposed classifier takes advantage of the Wikipedia to generate the corpus defining each category. The text is then analyzed syntactically using Natural Language Processing software. The proposed classifier is highly accurate and outperforms Machine Learning trained classifiers.

Keywords: Text Categorization; Wikipedia; tf-idf; Machine Learning; Natural Language Processing.

1. Introducción

Hoy en día la información se ha convertido en un recurso estratégico de primer orden para las organizaciones, que obtienen y almacenan grandes volúmenes de datos de diversas fuentes y de manera automática. En estos casos, los sistemas de clasificación

pueden ayudar a administrar, consultar y extraer información de grandes sistemas de documentos. La *categorización o clasificación de textos* (Sebastiani, 2005) consiste en asignar textos a una o varias categorías.

El proceso de clasificación de textos comienza con el indexado del documento, consistente en mapear el documento a una representación compacta de su contenido. Los métodos de indexación normalmente utilizados en la categorización de textos utilizan una representación del documento mediante el *Modelo del Espacio Vectorial* (Salton, Wong & Buckley, 1975), donde un documento en lenguaje natural se representa mediante vectores de términos.

Cada término del documento ocupa una posición en el vector de términos, originando un problema de dimensionalidad en documentos extensos. Por esta razón existe una fase previa que intenta reducir el tamaño del documento. Esta reducción hace el problema más manejable durante el aprendizaje, y repercute positivamente tanto en el tiempo de proceso como en el espacio ocupado. La reducción de la dimensionalidad del vector de términos de un documento normalmente se consigue mediante las siguientes acciones: tokenización, eliminación de *stopwords* y palabras nulas (que no aportan valor dentro del contexto del documento) y lematización.

Una vez reducida la dimensionalidad del documento, se calcula la relevancia de cada término del vector de términos. Para este fin, uno de los métodos más extendidos es la obtención del *tf-idf* (Salton & Buckley, 1988). El *tf-idf* comprende la frecuencia del término en el documento (*tf)* y la inversa de la frecuencia de documentos que poseen el término (*idf)*. El *tf* representa la importancia local que el término posee en el documento, es decir, cuanto más aparezca un término en un documento, más relevante será ese término para ese documento. El *idf* representa la importancia global de un término en relación inversa, es decir, cuantos más documentos incluyan el término, este término será menos relevante.

Para clasificar un documento se calcula la similitud entre el vector de términos característicos de la categoría y el vector de términos del documento. El ángulo que forman los dos vectores se usa como medida de divergencia y el coseno del ángulo se utiliza como valor de similitud: si el coseno vale 1, los vectores son idénticos; si vale 0 se trata de vectores ortogonales y no hay coincidencia entre ambos (el coseno del ángulo no puede tomar valores negativos por cuanto la medida de tf-idf siempre tiene valores mayores o iguales a cero).

El principal problema de modelar un documento mediante el *Modelo de Espacio Vectorial* es que los documentos se convierten en bolsas de palabras (*bag of words*). Esta aproximación tiene 3 inconvenientes cuando se aplica a la clasificación de textos (Wang & Domeniconi, 2008):

1. Si se separan palabras que reflejan un único concepto, se pierde el significado original.
2. Las palabras sinónimas se representan como diferentes dentro del *Modelo de Espacio Vectorial*, afectando a la frecuencia de aparición de las palabras. Dentro de este modelo, los términos *computador*, *ordenador* y *pc* se consideran términos sin relación.

3. Con palabras polisémicas se considera un único significado, cuando una misma palabra puede estar utilizándose con varios significados dentro del documento.

Existen diversas aproximaciones para enriquecer la bolsa de palabras, tratando así de solventar uno o varios de los problemas anteriormente citados. En este trabajo, se ha optado por usar la herramienta de PLN *FreeLing* (Atserias, Casas, Comelles, González, Padró & Padró, 2006) para extraer la información sintáctica de un texto, solventado en parte la separación de términos relacionados y el problema de la polisemia de las palabras.

En este artículo se presenta un proceso de clasificación de textos basado en la construcción de un vector de términos de cada categoría apoyándose en la Wikipedia.

A continuación, se indica el trabajo relacionado. En los apartados 3, 4 y 5 se presenta en detalle el proceso utilizado para la categorización de textos, los experimentos diseñados y los resultados obtenidos. Finalmente en los apartados 6 y 7 se presentan las conclusiones y el trabajo futuro.

2. Trabajo relacionado

El *Modelo de Espacio Vectorial*, siendo una aproximación válida a la clasificación de textos, presenta ciertos inconvenientes que se han intentado subsanar. Estos intentos han ido encaminados a enriquecer con conocimiento externo la bolsa de palabras, añadiéndole nuevos elementos.

En los últimos años, por el tamaño y notoriedad que ha alcanzado, ese conocimiento extra se ha buscado en la Wikipedia (Strube & Ponzetto, 2006), (Gabrilovich & Markovitch, 2006), (Chang, Ratinov, Roth & Srikumar, 2008), (Wang, Hu, Zeng & Chen, 2009). Pero el uso de la Wikipedia para la construcción de ontologías o de relaciones semánticas tiene una serie de obstáculos importantes. En general se considera que un artículo de la Wikipedia es un concepto *per se*, aunque esto no sea cierto siempre. Los artículos entre ellos tienen una estructura de enlaces entrantes y salientes, que permite construir una estructura de relaciones entre conceptos, ya sean estas relaciones jerárquicas o semánticas.

En (Strube & Ponzetto, 2006) se centran en medir la relación entre dos conceptos usando la Wikipedia. Aunque no se centra en la clasificación de texto, este trabajo si es interesante para resaltar que los artículos son descripciones de los conceptos a los que se refiere.

En (Gabrilovich & Markovitch, 2006) se relacionan los términos de los documentos a clasificar con conceptos de la Wikipedia. Para extraer los conceptos de la Wikipedia, se siguen una serie de pasos: se eliminan los artículos que no son conceptos, se eliminan las palabras raras (presentes en menos de tres artículos), *stop words*, y se realiza el stemming sobre el resto. Del texto a clasificar se toman sucesivamente diferentes elementos y con ellos se van buscando los conceptos relevantes. Posteriormente, la bolsa de palabras se ve enriquecida con estos nuevos conceptos y finalmente se decide la categoría a la que pertenece.

En (Gabrilovich & Markovitch, 2007) se introduce el *Explicit Semantic Analysis* (ESA). En este enfoque se emplean técnicas de clasificación de textos para representar explícitamente el significado de un texto en lenguaje natural en términos de un espacio multidimensional de conceptos Wikipedia. Básicamente elaboran un índice invertido de la Wikipedia, en el que cada palabra tiene asociada una lista de conceptos, también de la Wikipedia. Para clasificar un texto, se extraen sus términos y se construye un vector con pesos de conceptos relacionados. Se puede determinar la relación existente entre dos textos comparando los vectores de conceptos relacionados con sus términos.

En (Wang & Domeniconi, 2008) intentan resolver los inconvenientes que supone la representación de los documentos como meras *bolsas de palabras* mediante la inclusión de conocimiento procedente de la Wikipedia en un núcleo semántico que se utiliza para mejorar la representación de los documentos. Como principal aportación logran mantener como un único concepto los conceptos compuestos de más de un término, capturan la semántica de los sinónimos, y eliminan la ambigüedad de los términos polisémicos. Dado un texto a clasificar, primero elaboran una lista de *conceptos candidatos*, es decir, conceptos que están presentes en el texto y que se pueden mapear a conceptos de la Wikipedia. Con la lista de conceptos candidatos, se usa un tesauro elaborado con la Wikipedia para seleccionar conceptos relacionados semánticamente, aplicándose una medida de esta relación semántica según cada caso. Con una matriz de proximidad elaboran el *Modelo de Espacio Vectorial* extendido para el documento en cuestión.

En (Cui, Lu, Li & Chen, 2009) proponen un método para extraer conceptos de la Wikipedia sin proponer ningún clasificador. Simplemente se centran en encontrar los conceptos y separarlos de las instancias, cosa que no se hace en (Strube & Ponzetto, 2006) y (Gabrilovich & Markovitch, 2006). Pongamos por ejemplo que el concepto *empresa* tendría como instancia *Microsoft*. Lo interesante de este trabajo para nosotros es que usaron el Stanford POS Tagger, un software de PLN, para identificar los conceptos, analizando las frases que contienen elementos del tipo "*is a*","*type of*","*name of*", "*a kind of*" y "*one of*" para reconocer conceptos.

En este trabajo no se opta por la creación automática de ontologías o tesauros basados en la Wikipedia, ya que se desea dar al usuario la decisión sobre cómo construir sus categorías. Para ello, el trabajo está inspirado en (Chang, Ratinov, Roth & Srikumar, 2008) que usa las etiquetas como elemento para definir una categoría. Éstos, a su vez, se basan en (Gabrilovich & Markovitch, 2007) para incrementar el contenido semántico de un fragmento de texto. Básicamente, utilizan las etiquetas de las categorías como base para la clasificación, extrayendo, mediante ESA (*Explicit Semantic Analysis*), conceptos de la Wikipedia relacionados con dicho texto. Estos conceptos son utilizados posteriormente para determinar la pertenencia o no de un documento a una categoría, comparando vectores de conceptos.

Aún considerando que la descripción textual que se hace de una categoría es importante para la clasificación, en este artículo se opta por aprovechar la función sintáctica de las palabras dentro de un texto. Dada una o varias palabras, éstas se emplean para encontrar artículos de la Wikipedia donde dichas palabras son relevantes en función de su categoría gramatical (sustantivo, verbo y adjetivo).

3. Proceso de clasificación

Para el proceso de clasificación propuesto es necesario realizar el trabajo previo de preparación del contenido de la Wikipedia. Éste se encuentra disponible para su descarga vía Web. Debido a que no toda la información existente resulta útil para nuestro propósito, es necesario realizar un filtrado con objeto de identificar los artículos y desechar el resto de información.

Los artículos identificados se almacenan en una tabla, con objeto de utilizar las funcionalidades de indexación y búsqueda de texto completo de MySQL. MySQL mantiene un índice numérico para cada término (índice *Full-Text*) que constituye una variante del *tf-idf* (Oracle), y cuyo cálculo viene dado por la expresión:

$$W = \frac{\log(dtf)+1}{sumdtf} * \frac{U}{1+0.0115*U} * \log(\frac{N-nf}{nf}) \qquad (1)$$

donde *dtf* es el número de veces que un término aparece en el artículo, *sumdtf* es la suma de [log*(dft)*+1] para todos los términos del mismo artículo, *U* es el número de términos únicos en el artículo, *N* es el número total de artículos de la tabla y *nf* es el número de artículos que contienen el término.

A partir de este punto, el proceso de clasificación se puede dividir en dos etapas: una centrada en la creación de las categorías y otra dedicada a la clasificación de los textos en las categorías creadas. El proceso de clasificación global se muestra en la figura 1.

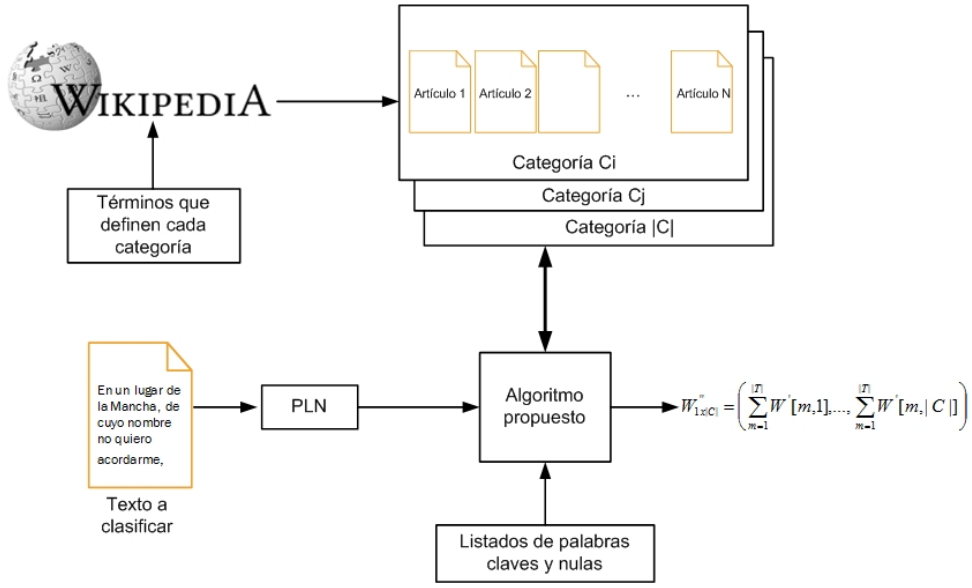

Figura 1 – Proceso global de clasificación

3.1. Creación de categorías

En este enfoque, el usuario aporta las categorías y algunos términos que considera relevantes para la definición de cada categoría. En general, lo importante no es que el usuario especifique un gran conjunto de palabras, sino un conjunto que permita desambiguar la categoría.

A partir de los términos de una categoría determinada, y apoyándose en la variante de *tf-idf* utilizada en (1), se extraen los *N* artículos de la Wikipedia en los que la relevancia de estos términos es elevada. La *relevancia* (*R*) de un término en un artículo se obtiene de la forma:

$$R = qf * w \tag{2}$$

donde *qf* representa el número de veces que el término se repite en la consulta, y *w* es el valor dado por la expresión (1).

A continuación se procesan los artículos extraídos mediante técnicas de PLN. Este procesado reduce la dimensionalidad en varias fases: se lematizan los términos encontrados reduciéndolos a una forma común, generalmente al masculino singular o al infinitivo. Posteriormente, se catalogan los elementos sintácticamente más relevantes: entidades o nombres propios, nombres comunes, verbos y adjetivos. Finalmente, se eliminan las palabras vacías y los signos de puntuación.

La capacidad de detección de nombres propios de la herramienta de PLN *FreeLing* se ha ampliado, considerando como un único nombre secuencias de caracteres como por ejemplo "*De la Rosa*", "*McLaren*", "*US Open*", "*ONU*".

Al final de este proceso, cada uno de los *N* artículos se reducen a un número términos con una representación homogénea. Los nombres propios se conservan como tales y el resto de términos relevantes comparten una representación similar en los distintos artículos. El conjunto de elementos identificados en cada uno de los *N* artículos conforman el conocimiento que el sistema tiene sobre esa categoría.

Cada categoría creada se almacena mediante una tabla, donde cada tupla almacena la información relativa a cada uno de los *N* artículos, esto es, la relevancia del artículo *n* en la categoría C_j, $R(n,C_j)$, y el conjunto de términos identificados para ese artículo. De este modo, el proceso de creación de categorías finaliza con tantas tablas como categorías se hayan creado.

3.2. Algoritmo de clasificación

La clasificación consiste en determinar a qué categoría, dentro de un conjunto predeterminado $\{C_1,\ldots, C_{|C|}\}$, pertenece un texto. Como paso previo al proceso de clasificación se debe procesar el texto con el fin de obtener los términos relevantes para el algoritmo propuesto. Este procesamiento transforma el texto en un vector de términos relevantes $\{t_1,\ldots,t_{|T|}\}$, mediante el mismo proceso y herramientas utilizadas con los artículos de la Wikipedia.

El peso de cada término t_i en una categoría C_j se obtiene mediante la siguiente expresión:

$$W(t_i, C_j) = \sum_{n=1}^{N} R(n, C_j) * P(t_i, n, C_j) \quad (3)$$

donde N es el número de artículos extraídos de la Wikipedia para crear cada una de las categorías, $R(n,C_j)$ es la relevancia del artículo n en la categoría C_j, y $P(t_i,n,C_j)$ representa el peso del término t_i dentro del artículo n, que viene dado por la expresión (1).

El valor obtenido $W(t_i,C_j)$ se modifica según el tipo de término t_i de que se trate, atendiendo a la siguiente clasificación: palabra especial, palabra nula, palabra normal o entidad. Las *palabras especiales* son aquellas específicas, que no son exclusivas, de cada categoría. Las *palabras nulas* son aquellas que el usuario considera que pueden introducir distorsión en los resultados finales. Las *entidades* se corresponden con nombres de personas y organizaciones, que tienen importancia para determinar la categoría a la que pertenece un texto. Las *palabras normales* son aquellas que no están en ninguno de los grupos anteriores. Así, el peso $W(t_i,C_j)$, se incrementa para las palabras especiales y las entidades en un factor 100 y 10 respectivamente, se anula para las palabras nulas, y se disminuye para las palabras normales en un factor 0,1, obteniendo el valor modificado $W'(t_i,C_j)$.

El cálculo anterior debe extenderse sobre todas las categorías existentes, y aplicarse a todos los términos relevantes del texto que se desea clasificar. De este modo, el resultado es la matriz:

$$W'_{|T|x|C|} = \begin{pmatrix} W'_{11} & \cdots & W'_{1|C1} \\ \vdots & \ddots & \vdots \\ W'_{|T|1} & \cdots & W'_{|T||C|} \end{pmatrix} \quad (4)$$

Finalmente la suma de los valores de cada columna de la matriz $W'_{|T|x|C|}$ representa la relevancia del texto en cada categoría:

$$W''_{1x|C|} = \left(\sum_{m=1}^{|T|} W'[m,1], \ldots, \sum_{m=1}^{|T|} W'[m,|C|] \right) \quad (5)$$

El sistema concluye que la categoría a la que pertenece el texto es la correspondiente al máximo valor de $W''_{1x|C|}$.

4. Resultados experimentales

Para realizar los experimentos se ha tomado como conjunto de prueba 206 titulares de noticias recogidas durante varios días consecutivos, de las categorías '*Fórmula 1*', '*Tenis*', '*Ciclismo*', '*Golf*' y '*Atletismo*' (ver tabla 1) del sitio de noticias deportivas www.marca.com/deporte/rss/index.html. En esta URL podemos encontrar las noticias y la categoría asociada.

Se ha considerado una categoría de control, '*Cultura*', para detectar comportamientos anómalos del clasificador.

Tabla 1 - Categorías y número de noticias por categoría

Categorías	# Noticias
Atletismo	40
Ciclismo	49
Motor	44
Golf	23
Tenis	50

En la tabla 2 se indican los términos indicados por el usuario para crear cada categoría, buscando en la Wikipedia los artículos en los que mayor peso tengan.

Las palabras que se han considerado clave para cada categoría son las que se indican en la tabla 2. Las palabras clave tienen un alto significado en sus categorías, pero no son exclusivas a éstas. Por ejemplo, *vuelta* puede aparecer en ciclismo o motor, pero se desea que se valore más en la categoría ciclismo.

Tabla 2 - Términos que definen cada categoría y palabras clave por categoría

Categoría	Términos	Palabras clave
Atletismo	Atletismo	Altura, longitud, metros, m, kilómetros, carrera, correr
Ciclismo	tour, ciclista	Tour, etapa, giro, crono, vuelta
Motor	F1, Fórmula 1	Pole, Fórmula 1
Golf	golf, golfista	Golf, PGA, green, golpes
Tenis	tenis, grand slam	Atp, tenis
Cultura	Cultura	

El listado de palabras nulas se indica en la tabla 3. Las palabras nulas son globales a todas las categorías. Estas palabras pueden introducir sesgos indeseables porque pueden repetirse más en ciertos artículos que en otros.

Tabla 3 - Lista de palabras nulas

Palabras nulas
clasificado, clasificación, líder, liderato, triunfo, primero, segundo, final, victoria, plaza, vuelta, tiempo, equipo, vencer, podio

4.1. Resultados de las pruebas

La evaluación del rendimiento de las pruebas se realizará mediante el *índice de Cohen*, corrección del porcentaje de aciertos que no computa aquellos que hayan podido ser fruto del azar (Japkowicz & Shah, 2011).

En todas las pruebas realizadas se han empleando 750 artículos de la Wikipedia para crear cada categoría. En las pruebas se ha aplicado el algoritmo propuesto, evaluando el impacto de la incorporación de las siguientes funcionalidades:

- La lista de palabras clave
- La lista de palabras nulas
- La modificación del peso del término en función de su tipo: entidad, palabra clave, palabra nula, otros

Experimento 1

Los resultados se obtuvieron aplicando el algoritmo propuesto a las noticias que forman nuestro grupo de prueba, junto con las funcionalidades indicadas.

La matriz de confusión obtenida es la de la Tabla 4. El valor del índice de Cohen es 0.9078, esto es, el clasificador puede alcanzar un 90,78% de acierto sin considerar aquellos que son fruto del azar. La categoría *cultura* no obtiene ningún resultado.

Tabla 4 - Resultados del experimento 1

		Clasificación del experimento 1							**Efectividad por categoría**		
		Atletismo	Ciclismo	Motor	Golf	Tenis	Cultura	Total	Precisión	Cobertura	F1-score
Categoría real	Atletismo	36	0	0	3	1	0	40	0,947	0,9	0,923
	Ciclismo	1	44	0	1	3	0	49	0,936	0,898	0,917
	Motor	0	3	40	1	0	0	44	1	0,909	0,952
	Golf	1	0	0	21	1	0	23	0,84	0,913	0,875
	Tenis	0	0	0	0	50	0	50	0,909	1	0,952
	Cultura	0	0	0	0	0	0	0	N/A	N/A	N/A
	Total	38	47	40	25	55	0	206			

Experimento 2

En esta prueba no se usa la lista de palabras clave ni la de palabras nulas. La matriz de confusión se muestra en la Tabla 5. El valor del índice de Cohen es 0.8522.

El resultado se degrada un 6,12%. Son 9 titulares más los que se clasifican mal frente al experimento 1. La categoría '*cultura*' no obtiene ningún resultado.

Tabla 5 - Resultados del experimento 2

		Clasificación del experimento 2							Efectividad por categoría		
		Atletismo	Ciclismo	Motor	Golf	Tenis	Cultura	Total	Precisión	Cobertura	F1-score
Categoría real	Atletismo	34	2	0	2	2	0	40	0,919	0,85	0,883
	Ciclismo	2	40	3	1	3	0	49	0,889	0,816	0,851
	Motor	0	2	40	1	1	0	44	0,93	0,909	0,92
	Golf	1	1	0	19	2	0	23	0,792	0,826	0,809
	Tenis	0	0	0	1	49	0	50	0,86	0,98	0,916
	Cultura	0	0	0	0	0	0	0	N/A	N/A	N/A
	Total	37	45	43	24	57	0	206			

Experimento 3

No se utilizan las listas de palabras claves y nulas, ni la modificación del peso del término en función de su tipo, de manera que todas las palabras serán consideradas iguales. La matriz de confusión que se obtiene se indica en la Tabla 6.

Tabla 6 - Resultados del experimento 3

		Clasificación del experimento 3							Efectividad por categoría		
		Atletismo	Ciclismo	Motor	Golf	Tenis	Cultura	Total	Precisión	Cobertura	F1-score
Categoría real	Atletismo	31	4	0	1	4	0	40	0,861	0,775	0,816
	Ciclismo	1	38	4	2	4	0	49	0,731	0,776	0,752
	Motor	2	3	34	4	1	0	44	0,895	0,773	0,829
	Golf	1	3	0	14	5	0	23	0,636	0,609	0,622
	Tenis	1	4	0	1	44	0	50	0,759	0,88	0,815
	Cultura	0	0	0	0	0	0	0	N/A	N/A	N/A
	Total	36	52	38	22	58	0	206			

El valor del índice de Cohen es 0.7222. Los resultados empeoran drásticamente: un 20,45% frente al experimento 1 y un 15,25% frente al experimento 2. Sigue sin haber titulares clasificados como pertenecientes a '*cultura*'.

Experimento 4

Se desea realizar la comparación del algoritmo propuesto con técnicas de *Aprendizaje Automático*. Como herramienta para esta tarea se ha usado el software Weka. El conjunto de entrenamiento se reduce a una propiedad por noticia estableciendo de esta forma las mismas condiciones que en el experimento 1, dado que es el más sencillo de los modelos y a la vez el que genera mejores resultados. Para la construcción del modelo se emplearon los algoritmos Bayes Multinomial y SVM (Máquinas de Vectores Soporte). El modelo fue validado mediante validación cruzada de las 206 noticias,

obteniéndose los resultados indicados en la tabla 7, donde no se ha realizado corrección del azar.

Tabla 7 - Resultados de las pruebas realizadas con validación cruzada

Prueba	% de acierto
Algoritmo propuesto (Experimento 1)	93,10%
Bayes Multinomial	91,26%
SVM	86,41%

Experimento 5

En los experimentos anteriores se han utilizado 750 artículos de la Wikipedia para crear las diferentes categorías. En esta prueba se evalúa la influencia del número de artículos en el resultado final. Para medir el efecto, se han mantenido las mismas condiciones que las utilizadas en el experimento 1, y se ha ido variando el número de artículos por categoría.

Los resultados se muestran en la figura 2. En el eje de abscisas se representa el número de artículos que se emplean. En el eje de ordenadas se representan dos magnitudes: número de aciertos y número de titulares asignados a '*cultura*'.

Los resultados muestran que 10 artículos son claramente insuficientes: 6 titulares son asignados a la categoría '*cultura*'. Al incrementar el número de artículos, va incrementándose progresivamente el número de aciertos, disminuyendo (hasta hacerse cero) el número de titulares asignados a '*cultura*'. Se observa un máximo sobre los 750 artículos. Incluir más artículos reduce el número de aciertos, pues se incrementa el vocabulario con palabras que no son realmente importantes en el corpus, y que se extraen de artículos cada vez menos relevantes.

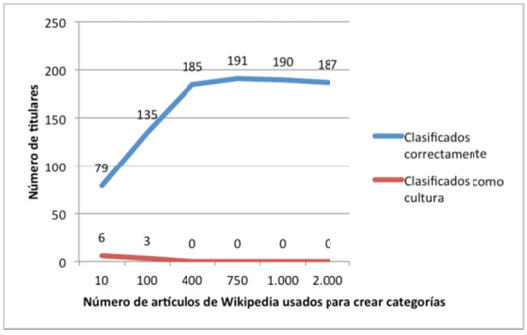

Figura 2 - Influencia del número de artículos en los resultados

5. Análisis de resultados

Del experimento 5 se observa un compromiso para escoger la cantidad de artículos para crear el corpus de las categorías. Un número pequeño de artículos no producen el vocabulario necesario para clasificar correctamente, y un número excesivo de artículos

introducen *ruido*, pues el vocabulario obtenido deja de ser plenamente representativo de la categoría.

Del resultado de los experimentos 1, 2 y 3, se observa que los mejores resultados se obtienen considerando las palabras clave y nulas, así como la modificación del peso del término en función de su tipo. En términos relativos, la modificación del peso del término en función de su tipo tiene un mayor impacto en los resultados, que la utilización de las listas de palabras especiales y nulas.

Los resultados obtenidos con las técnicas de *Aprendizaje Automático* son inferiores a los obtenidos con el algoritmo propuesto (91,26% de la prueba con Bayes Multinomial frente al 93,10% del algoritmo propuesto).

6. Conclusiones

Se ha creado un sistema de clasificación automático de textos que puede cumplir con las expectativas de efectividad que un usuario pudiera esperar en su uso. En el experimento 1 se comprueba que, dadas unas condiciones concretas, se alcanzan unas medidas del 93,10% de acierto en la clasificación, en una situación de partida desfavorable: se toman noticias de deportes, todas ellas compartiendo vocabulario.

El presente clasificador, muestra una forma más de salvar la visión que deja el *Modelo de Espacio Vectorial* de los documentos como meras *bolsas de palabras*, al valorar la función sintáctica que tiene una palabra dentro del texto a la hora de computar su peso.

Hemos de concluir, a la vista de los experimentos, que el rendimiento del clasificador está en relación con la configuración: dependiendo del número de artículos de la Wikipedia que se tomen o de los listados de palabras que se incluyan. El clasificador propuesto debe ser configurado según cada necesidad o la cercanía conceptual de los elementos a clasificar. El número de artículos seleccionados influye directamente en dos aspectos: el espacio que ocupan las categorías en las unidades de almacenamiento y el tiempo que se tarda en computar un texto en cada categoría.

Por otro lado, las listas de palabras clave y nulas tienen un impacto del 6,12% en el índice de aciertos entre el experimento 1 y 2, lo que indica que la capacidad de influencia de estas listas es limitada, aunque su aportación es deseable por cuanto incrementan el porcentaje de acierto. La mayor parte de los aciertos recae en el *Modelo de Espacio Vectorial* enriquecido con el análisis sintáctico propuesto.

Los resultados existentes en la literatura científica en su mayoría se basan en diferentes corpus en inglés, siendo más escasos los disponibles en español. En (Venegas, 2007) se utiliza una muestra de 222 artículos en español con unos resultados cercanos en fase de prueba al 76,74% con SVM y de 68,18% con Naive Bayesiano, aunque los resultados no son directamente comparables pues tanto el corpus de datos como las categorías elegidas (Química Industrial, Ingeniería en Construcción, Trabajo Social y Psicología) son diferentes.

Igual que dos humanos pueden no estar de acuerdo al clasificar ciertos titulares, el clasificador automático, en determinados momentos, también hace *interpretaciones*, aunque estas sean de base matemática. Pensemos que el titular "*Jaime Alguersuari y Sergio García, amigos y solidarios en un torneo de golf benéfico*" estaba originalmente

clasificado como *motor*. Pero la palabra *golf* junto con el nombre *Sergio García*, tienen mayor peso que *Jaime Alguersuari*, por lo que el titular se clasifica en la categoría *golf*, contabilizándose como un error.

7. Líneas futuras

En esta propuesta, el corpus es la fuente de la que emana el conocimiento que se emplea en la clasificación. Por lo tanto, la principal línea de actuación sería trabajar en el estudio del corpus de Wikipedia, introduciendo la autoconfiguración del clasificador mediante desambiguación supervisada.

Hay que estudiar especialmente aquellas palabras compartidas por dos o más categorías. Dentro de estas palabras habrá que encontrar la manera de diferenciar cuándo esa palabra es realmente determinante en el corpus.

Por último, sería interesante incluir reconocimiento de patrones sintácticos que ayuden a reconocer expresiones lingüísticas recurrentes (Cruz, Troyano, Enriquez & Ortega, 2008). En la propuesta actual, únicamente se incluyen para el reconocimiento de entidades.

Agradecimientos

Este trabajo ha contado con la financiación del Fondo Europeo de Desarrollo Regional (FEDER) y el Ministerio de Industria, Turismo y Comercio (MITYC) a través del Plan Avanza I+D (TSI-020302-2008-115).

Referencias bibliográficas

Atserias, J., Casas, B., Comelles, E., González, M., Padró, L. & Padró, M. (2006). FreeLing 1.3: Syntactic and Semantic Services in an Open-source NLP Library. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, 48-55.

Chang, M.-W., Ratinov, L., Roth, D. & Srikumar, V. (2008). Importance of Semantic Representation: Dataless Classification. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, 830-835.

Cruz, F. L., Troyano, J. A., Enriquez, F. & Ortega, J. (2008). Experiments in Sentiment Classification of Movie Reviews in Spanish. In *Sociedad Española de Procesamiento del lenguaje Natural*, 41, 73-80.

Cui, G., Lu, Q., Li, W. & Chen, Y. (2009). Mining Concepts from Wikipedia for Ontology Construction. In *Proceedings of the 2009 IEEE/WIC/ACM international Joint Conference on Web Intelligence and Intelligent Agent Technology* - Vol. 03. Web Intelligence & Intelligent Agent. IEEE Computer Society, Washington, DC, 287-290.

Gabrilovich, E. & Markovitch, S. (2006). Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In

Proceedings of the 21st National Conference on Artificial Intelligence (AAAI), Boston, 1301–1306.

Gabrilovich, E. & Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, *IJCAI'07*, Hyderabad, India, 1606–1611.

Japkowicz, J. & Shah, M. (2011). Evaluating Learning Algorithms. A Classification Perspective. NY, USA: Editor Cambridge University Press. ISBN: 978-0-521-19600-0.

Oracle Corporation. MySQL Internals Algorithms - MySQL Forge Wiki. *MySQL Forge*. [En línea].

Quinlan, R. (1993). C4.5: Programs for Machine Learning. San Mateo, California: Editor Morgan Kauffman. ISBN: 978-1-55860-238-0.

Salton, G., Wong, A. & Buckley, C. (1975). A Vector Space Model for Automatic Indexing. *Communications of ACM, 8*(11), 613-620.

Salton, G. & Buckley, C. (1988). Term Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5), 513-523.

Sebastiani, F. (2005). Text Categorization. In Alessandro Zanasi (Ed.), *Text Mining and its Applications*. Southampton, UK: Editora WIT Press, 109-129.

Schapire, R. E., Freund, Y. & Schapire, R. (1996). Experiments with a new boosting algorithm. *International Conference on Machine Learning*, 148–156.

Strube, M. & Ponzetto, S. (2006). WikiRelate! Computing Semantic Relatedness Using Wikipedia. *Association for the Advancement of Artificial Intelligence*, 1419–1424.

Venegas, R. (2007). Clasificación de textos académicos en función de su contenido léxico-semántico. *Revista signos*, 40(63), 239-271. ISSN 0718-0934.

Wang, P. & Domeniconi, C. (2008). Building Semantic Kernels for Text Classification using Wikipedia. In Proceedings of the *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, PA, USA, 713-721.

Wang, P., Hu, J., Zeng, H.-J. & Chen, Z. (2009). Using Wikipedia Knowledge to Improve Text Classification. *Knowledge and Information Systems, 19*, 265-281.