

# INTEGRACIÓN DE INFORMACIÓN CONCEPTUAL DE WORDNET EN CATEGORIZACIÓN AUTOMÁTICA DE TEXTOS

José Carlos Cortizo Pérez<sup>(1)</sup>, Miguel Jaime Ruiz Leyva<sup>(1)</sup> y José María Gómez Hidalgo<sup>(2)</sup>  
Universidad Europea de Madrid

<sup>(1)</sup>Estudiantes de Ingeniería Superior en Informática

<sup>(2)</sup>Departamento de Inteligencia Artificial

[joseca@bravusoft.zzn.com](mailto:joseca@bravusoft.zzn.com), [ethan113@hotmail.com](mailto:ethan113@hotmail.com), [jmgomez@dinar.esi.uem.es](mailto:jmgomez@dinar.esi.uem.es)

## Resumen

Este documento pretende explicar los beneficios o las desventajas que puede aportar el uso de información conceptual proveniente de WordNet, en las tareas de Categorización Automática de textos, en comparación con la Categorización usando representaciones basadas en el uso de las palabras en sus múltiples vertientes.

## 1 Introducción

La Categorización Automática de textos utiliza, típicamente, el Modelo del Espacio Vectorial para la representación de los documentos tomando las palabras sin importar su orden y teniendo en cuenta las frecuencias de aparición de éstas en los documentos. Una vez se tienen estas representaciones, se pueden aplicar infinidad de algoritmos para establecer la/s categoría/s del nuevo documento. Desde luego, la efectividad de los algoritmos en cuanto a la clasificación es algo que no se puede discutir, pues es inherente al algoritmo en sí, pero, a priori, una representación de los documentos exclusivamente basada en palabras y sin tener, si quiera, en cuenta el orden, implica una gran pérdida de conocimiento.

Para evitar esta pérdida de conocimiento, una de las posibles opciones es el tomar los conceptos que se usan en los documentos, no las palabras. Se pueden dar, al menos, dos razones obvias para usar conceptos en lugar de palabras. La primera es que, a pesar que se siga sin tener en cuenta el orden de aparición, los conceptos tienen información “per se” del contexto en el que se encuentran, mientras que las palabras, a menudo, no proporcionan este conocimiento. La segunda razón, muy ligada a la primera, es la desambiguación; las palabras pueden reflejar, lo cual ocurre a menudo, distintos conceptos que pueden ser totalmente dispares (p. e. banco puede ser un lugar donde se almacena dinero, un objeto en el que las personas se sientan o, incluso, un conjunto de peces). Si se representa un documento tomando los conceptos adecuados al uso de la palabra en ese contexto, se gana (o al menos se debería ganar) mucha información pues estamos eliminando, en parte, las suposiciones de independencia entre los términos

y estamos agrupando aquellas palabras que representan un mismo concepto en una única unidad de información y, a la vez, dividiendo otras palabras en N unidades de información dependiendo de los conceptos asociados.

De todas formas la información conceptual a usar no es solo aquella ligada directamente con los conceptos pues se pueden usar las categorías de los conceptos relacionados con las palabras. Wordnet ofrece información acerca de los archivos lexicográficos en los que se encuentran las palabras. Cada archivo lexicográfico (hay un total de 44 archivos) contiene las palabras que tienen conceptos asociados con una determinada categoría, por ejemplo los verbos relacionados con la higiene personal, los sustantivos que reflejan nombres propios, etc.

Este último enfoque supone la semilla para la investigación que, en este documento, pretende plasmar su recorrido y conclusiones. Para poder comprobar si lo anteriormente planteado surte buenos resultados, se establecen dos etapas, primero una etapa totalmente comparativa en la cuál se intentará evaluar la efectividad de la clasificación usando conceptos frente a la categorización mediante palabras usando una colección de documentos determinada, SEMCOR. Esta colección de documentos nos ofrece, para cada término en cada documento, la referencia al synset en WordNet [Beckwith *et al.*, 91] sin ambigüedad alguna ya que ha sido procesado manualmente para establecer estos conceptos. Al tener una colección de textos que nos proporciona ambos enfoques, solo bastará probarlos bajo mismas condiciones para obtener los resultados y poder evaluarlos.

En situaciones reales, los problemas de clasificación de textos se enfrentan a la clasificación de textos compuestos por palabras, pues es así como se escriben, y no contamos con los conceptos adecuados a cada palabra en su entorno. Además, SEMCOR contiene documentos que pertenecen a una categoría de forma excluyente, lo cual facilita un tanto la tarea de clasificación. Por todo esto conviene, también, evaluar la clasificación tanto con palabras como con conceptos con una colección de textos que podemos denominar “real”, es decir, con unas dimensiones parecidas a las dimensiones en las situaciones típicas, con la posibilidad de asignar a un documento más de una

categoría y, sobre todo, con los textos sin desambiguar. La colección aquí usada es Reuters 21578 [Lewis, 99], tomando la partición de Lewis [Lewis, 92] para seleccionar los conjuntos de entrenamiento y prueba, pues es tanto la colección como la partición están muy extendida en otros trabajos con lo que se podrán obtener resultados realmente comparables con resultados de trabajos previos en el tema

## **2 El Modelo del Espacio Vectorial con palabras para Categorización de Textos**

Para realizar la investigación que en este documento nos atañe, se pretenden realizar diversos experimentos probando distintos algoritmos de clasificación, pero la entrada para todos ellos ha de ser la misma, ya que requerirá de un procesado previo de los documentos. La representación elegida como entrada para los distintos algoritmos es la representación de los documentos según el modelo del espacio vectorial, es decir, se va a recorrer la colección entera de documentos evaluando los términos que aparecen en cada uno de ellos y calculando sus valores  $tf$  e  $idf$  [Salton *et al.*, 1975] correspondientes de forma tal que un documento se representará como un conjunto de números que representan los valores de  $tf*idf$  de un determinado término en el documento (el término  $N$  aparecerá en la posición  $N$  dentro del conjunto de números). Cabe reseñar que los atributos no serán las palabras que aparecen en los documentos, tal cuál, pues resulta muy útil descartar palabras que aparecen muy frecuentemente en el uso del lenguaje tales como conjunciones, preposiciones y demás palabras sin información realmente conceptualmente que son utilizadas para la composición de estructuras del lenguaje tales como frases y diversos sintagmas y también se aplicará otro proceso que reduce el número de atributos y mantiene la información conceptual como es la extracción de raíces que permite obtener la raíz de una palabra de forma que se agrupan todas las derivaciones de un término en un solo atributo.

La información de  $tf*idf$  puede resultar muy interesante pues aporta un indicador de la relevancia de un determinado término en un documento, pero también conviene probar algún otro enfoque en el que simplemente se proporcione la información de la presencia o no de un determinado término en un documento, para ello se binarizarán los resultados obtenidos en el procesamiento previo cambiando los valores de los atributos que sean distintos de 0 por 1.

## **3 El Modelo del Espacio Vectorial usando conceptos para Categorización de Textos**

Cuando, en los documentos, en lugar de palabras, se tiene información conceptual acerca del tipo de categoría

sintáctica y del archivo lexicográfico en el que se encuentra la palabra, el proceso es realmente similar salvo que no hacen falta la extracción de raíces, ya que los conceptos son únicos y representan de por sí el conjunto de derivaciones de una palabra, no siendo necesario eliminar elementos de una lista de conceptos que no aportan información pues las conjunciones y demás elementos típicamente presentes en una lista de parada de palabras no tienen concepto asociado.

Lo único necesario para poder preprocesar los documentos representados como información conceptual según el modelo del espacio vectorial será el tener los documentos representados únicamente mediante conceptos. Para ello se necesita un preprocesado de los documentos que se encuentran representados como palabras para obtener un conjunto de valores de  $lexsn$  que representen las categorías conceptuales asociadas a las palabras. He aquí el principal problema o punto de interés, ya que bien es sabido que una palabra puede tener diversos conceptos asociados y también ocurre que un conjunto de palabras tengan un significado asociado como conjunto que nada o poco tenga que ver con el significado de las palabras (p. e. “red de computadoras” tiene un concepto muy específico referente a un conjunto de ordenadores independientes interconectados entre si, mientras que red tiene diversos conceptos que suelen relacionarse con mayas o similares) y como los archivos lexicográficos agrupan palabras según sus conceptos asociados, también presentan el mismo problema. A partir de aquí nos referiremos al enfoque usando conceptos como el enfoque que integra la información de los archivos lexicográficos de las palabras.

## **4 Utilizando SEMCOR para la comparación de los enfoques**

Como ya se ha comentado anteriormente, la investigación consta de dos partes; una primera que sirve como comparación objetiva entre la categorización usando palabras y la categorización usando información conceptual y otra parte que se sumerge en la manera de conseguir los conceptos a partir de las palabras de forma que no se pierda eficacia. Para poder realizar la primera parte, se necesita una colección de documentos que esté tanto en formato estándar, es decir, con palabras, como representada por información conceptual con la salvedad que esta información debe representar a las palabras que se encuentran en los documentos representados de esta forma. Básicamente se necesita tener los textos y por otra parte la información conceptual correctamente desambiguada.

Para todo esto, una colección de textos que, siendo utilizada en el ámbito de la categorización, se ajusta a nuestras necesidades, es SEMCOR [Rada, 98] que, por tanto, será la colección a utilizar. SEMCOR es una selección de documentos del “Corpus Brown”, cada uno de ellos perteneciente a una de las 15 categorías posibles (es importante reseñar que cada documentos solo puede estar

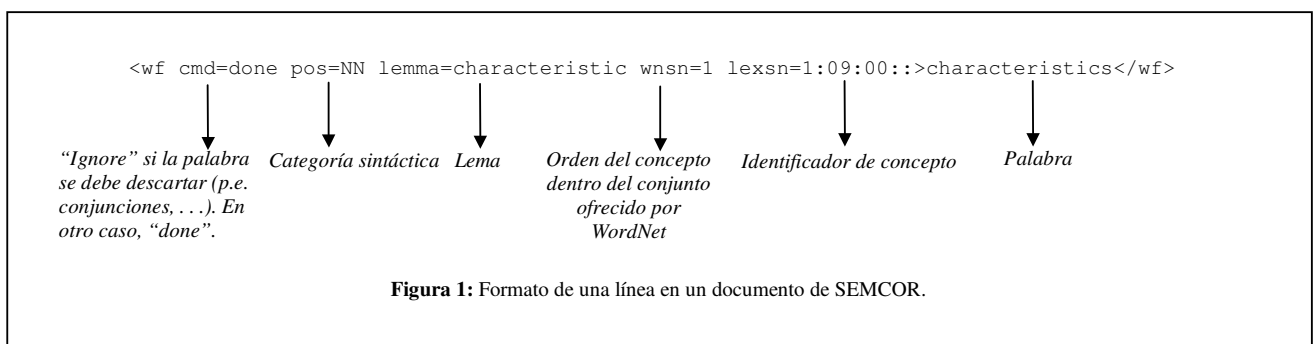
en 1 categoría; además se han seleccionado 186 documentos de los 352 totales que tiene SEMCOR pues estos 186 seleccionados están etiquetados totalmente mientras que el resto solo etiqueta los verbos). Los documentos están representados en formato SGML apareciendo, en cada línea, un conjunto de etiquetas y términos (ver Figura 1). Cada línea simplemente representa un término del documento, dando la palabra en si, el lexsns asociado y más información. Con todo esto se puede obtener el concepto asociado pero también se puede usar simplemente el lexsns que ofrece información acerca de la categoría sintáctica y acerca del archivo lexicográfico.

Una vez procesados los documentos y obtenidos los dos conjuntos necesarios es el momento de realizar una serie de experimentos para comparar ambos enfoques. Típicamente, en la representación mediante palabras, se seleccionan los atributos por InfoGain, cogiendo el 10% superior en valor de esta medida para quedarnos con los términos que aportan más información sobre la clasificación y descartar aquellos que lo único que provocarán es sobrecargar el sistema a la hora de clasificar. Así pues uno de los experimentos comparativos será tomar el 10% de los atributos seleccionándolos por InfoGain, tanto en la representación conceptual como por palabras. Otro de los experimentos será realizar lo mismo pero binarizando los valores de tf\*idf (explicado en el punto 2) y finalmente otro experimento será el no descartar atributos por InfoGain para poder estudiar si el descarte de atributos es realmente aceptable en representación conceptual, ya que estudios anteriores han demostrado que sobre palabras la selección del 10% es un valor muy útil.

Con estos 3 enfoques sobre cada conjunto de documentos, se aplicarán una serie de algoritmos, mediante la herramienta Weka [Witten, 99] (pues da soporte para manejar múltiples algoritmos y entre ellos los que aquí se presentan): Ibk ( con valores de K para 1, 2, 5, 10 y 16), J48 (C4.5), NaiveBayes, SMO, y AdaBoost, que es un metaclasificador que mejora otros algoritmos, sobre J48 (con 10, 20 y 30 iteraciones) y también sobre Ibk con valor 16 para k. Al aplicar todos estos algoritmos sobre las pruebas se obtienen los valores de F1 por micro y Macromedia que son los valores usados para comparar enfoques, algoritmos y variantes.

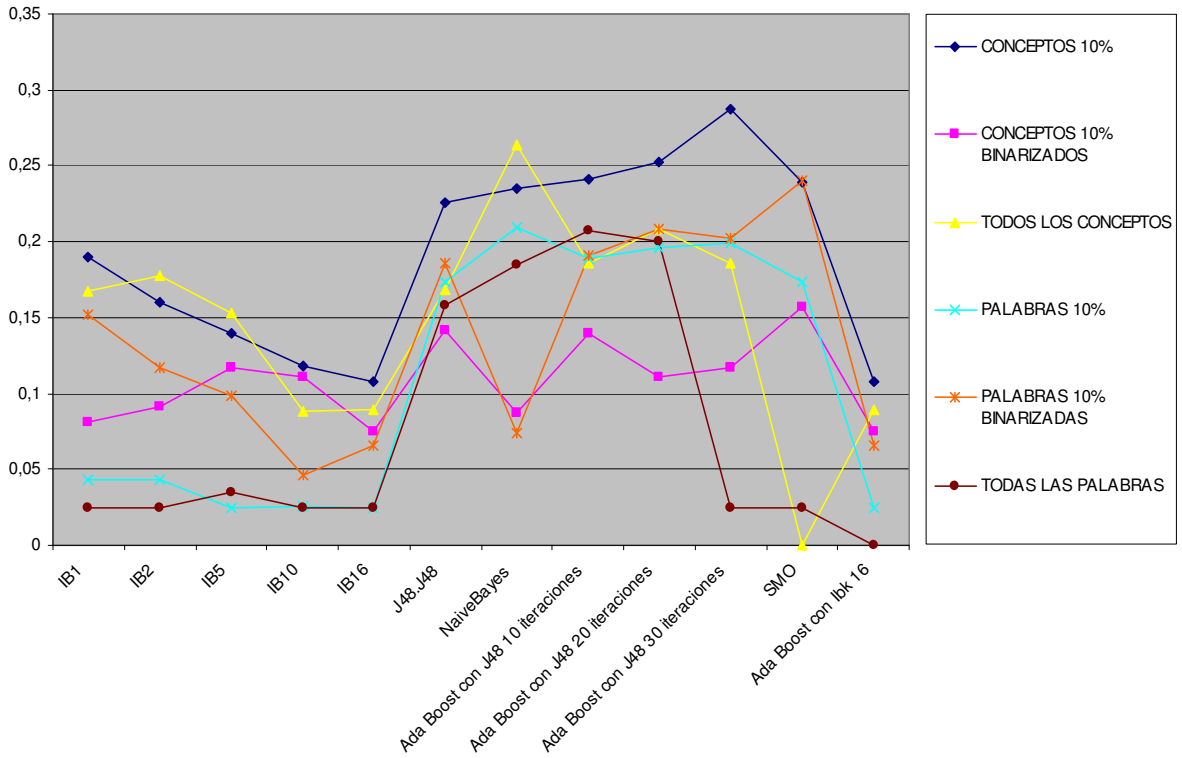
Los resultados (ver Tablas 1 y 2, los resultados se obtienen por validación cruzada con 10 carpetas) son muy parecidos en cuanto a los mejores resultados para el enfoque de representación con palabras y el enfoque con categorías conceptuales ( para palabras el mejor resultado es el algoritmo SMO para el 10% de las palabras binarizadas y para categorías conceptuales el mejor resultado es el AdaBoost con 30 iteraciones sobre el J48, existiendo una diferencia de unas 5 milésimas en F1 por micromedia a favor del enfoque mediante palabras, peor por Macromedia el enfoque ganador es por conceptos al superar en 5 centésimas). Pero esto no quiere decir que ambos enfoques sean similares, ya que si se analizan ambas representaciones se puede extraer un dato muy interesante: para la representación mediante palabras se usan 15.328 atributos en total mientras que para la representación de categorías conceptuales solo son 2.207, lo cual representa un 14% de los atributos en la representación por palabras. Así pues, el enfoque por categorías conceptuales será más eficiente (el tener tan pocos atributos, en relación al otro enfoque, hace que los algoritmos tarden mucho menos en categorizar), o mas eficaz si lo que se trata de equiparar eficiencia, ya que la representación del 10% sobre categorías conceptuales tiene del orden de 220 atributos mientras que la representación por palabras presenta 1.533 atributos; si reducimos esta última a 220, los resultados obtenidos son realmente malos como se puede apreciar (ver Tabla 3).

Además de todo esto, se realizan otra serie de experimentos para evaluar la adecuación del valor de 10% para la selección de atributos por InfoGain sobre las categorías conceptuales ya que es un valor que funciona muy bien con las palabras (conviene ver la gran bajada de los valores de F1 en el enfoque por palabras cuando el valor del porcentaje disminuye drásticamente) pero que no está contrastado para el enfoque con información conceptual. Cogiendo el algoritmo AdaBoost sobre J48 con 30 iteraciones (que es el que mejor funciona para categorías conceptuales al 10% de InfoGain) se realizan una serie de pruebas con los siguientes valores para el InfoGain: 15%, 10%, 7%, 5%, 3% y un último enfoque que es seleccionar únicamente los atributos cuyo valor de InfoGain sea distinto de 0 (21 atributos). Los resultados son bastante

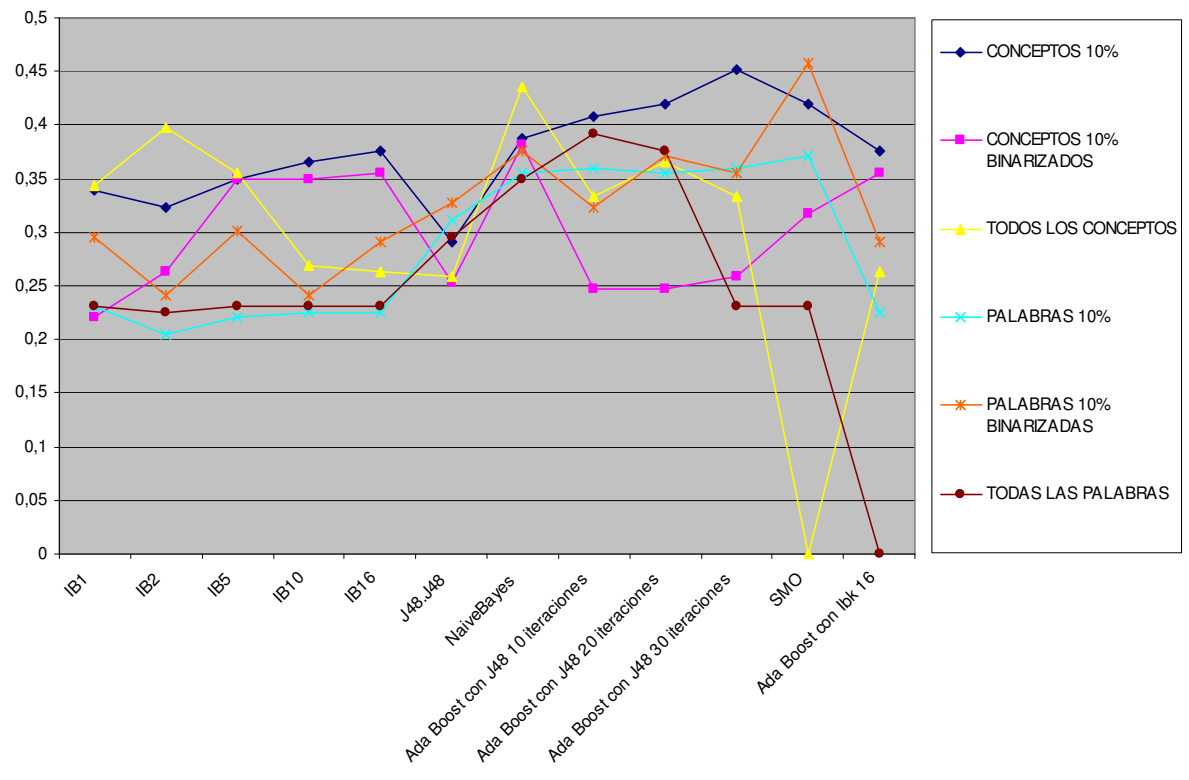


**Figura 1:** Formato de una línea en un documento de SEMCOR.

**TABLA 1: F1 por Macromedia**



**TABLA 2: F1 por micromedia**



sorprendentes (ver tabla 4) ya que se consiguen valores cercanos al 0.49 para F1 por micromedia e incluso superiores al 0.5 usando el algoritmo AdaBoost sobre j48 pero con 40 iteraciones; esto implica que ya no solo el enfoque por categorías conceptuales es tan bueno como el enfoque por palabras y más eficiente si no que es bastante mejor al superar en mas del 0.05 de F1 por micromedia al mejor valor para el enfoque por palabras si no que lo consigue con una cantidad ridícula de atributos lo cuál demuestra una gran eficiencia (21 atributos frente a 15 categorías lo cuál es prácticamente asociar una categoría por atributo). Otro dato comparativo muy importante es que tal y como se puede apreciar en la tabla 3, con 220 atributos el enfoque por palabras solo consigue un 0,38 a lo sumo mientras que con 21 atributos (menos del 10%) por categorías conceptuales se consigue un 0,4892 (y un 0,5054 con 40 iteraciones), lo cuál es una muestra del poder de las categorías conceptuales. Así pues, el enfoque por categorías conceptuales será más eficiente (el tener tan pocos atributos, en relación al otro enfoque, hace que los algoritmos tarden mucho menos en categorizar), o mas eficaz si lo que se trata de equiparar eficiencia, ya que la representación del 10% sobre categorías conceptuales tiene del orden de 220 atributos mientras que la representación por palabras presenta 1.533 atributos; si reducimos esta última a 220, los resultados obtenidos son realmente malos como se puede apreciar (ver Tabla 3).

Además de todo esto, se realizan otra serie de experimentos para evaluar la adecuación del valor de 10% para la selección de atributos por InfoGain sobre las categorías conceptuales ya que es un valor que funciona muy bien con las palabras (conviene ver la gran bajada de los valores de F1 en el enfoque por palabras cuando el valor del porcentaje disminuye drásticamente) pero que no está contrastado para el enfoque conceptual. Cogiendo el algoritmo AdaBoost sobre J48 con 30 iteraciones (que es el que mejor funciona para categorías conceptuales al 10% de InfoGain) se realizan una serie de pruebas con los siguientes valores para el InfoGain: 15%, 10%, 7%, 5%, 3% y un último enfoque que es seleccionar únicamente los atributos cuyo valor de InfoGain sea distinto de 0 (21 atributos). Los resultados son bastante sorprendentes (ver tabla 4) ya que se consiguen valores cercanos al 0.49 para F1 por micromedia e incluso superiores al 0.5 usando el algoritmo AdaBoost sobre j48 pero con 40 iteraciones; esto implica que ya no solo el enfoque por categorías conceptuales es tan bueno como el enfoque por palabras y más eficiente si no que es bastante mejor al superar en mas del 0.05 de F1 por micromedia al mejor valor para el enfoque por palabras si no que lo consigue con una cantidad ridícula de atributos lo cuál demuestra una gran eficiencia (21 atributos frente a 15 categorías lo cuál es prácticamente asociar una categoría por atributo). Otro dato comparativo muy importante es que tal y como se puede apreciar en la tabla 3, con 220 atributos el enfoque por palabras solo consigue un 0,38 a lo sumo mientras que con

21 atributos (menos del 10%) por categorías conceptuales se consigue un 0,4892 (y un 0,5054 con 40 iteraciones), lo cuál es una muestra del poder de las categorías conceptuales (así como de la información conceptual en general).

Algoritmo	F1 Macromedia	F1 micromedia
IB1	0,18758	0,18817
IB5	0,08086	0,18817
IB10	0,08700	0,26344
IB16	0,04967	0,24731
J48.J48	0,26513	0,33871
NaiveBayes	0,21766	0,32258
Ada Boost-J48 30 it.	0,30837	0,38172

**Tabla 3:** Resultados para la selección de 220 atributos en la representación por palabras.

## 5 Utilizando Reuters como un ejemplo real de Categorización

El trabajar con Semcor proporciona una idea comparativa aproximada del funcionamiento de los enfoques de categorización tanto usando palabras para la representación como realizando esta por categorías conceptuales. A pesar de esto, SEMCOR presenta varias características que hacen que los resultados obtenidos no se puedan considerar reales en cuanto a su aplicabilidad:

- Es una colección realmente muy pequeña y con pocas categorías.
- Cada documento solo puede pertenecer a una categoría.
- Los documentos presentan los conceptos desambiguados de forma manual, lo que realmente interesa es que esto se haga de forma automática.

Enfoque	F1 micromedia
15% InfoGain; 30 iteraciones	0,4247
10% InfoGain; 30 iteraciones	0,4516
7% InfoGain; 30 iteraciones	0,4623
5% InfoGain; 30 iteraciones	0,4193
3% InfoGain; 30 iteraciones	0,4462
>0 InfoGain; 30 iteraciones	0,4892
>0 InfoGain; 40 iteraciones	0,5054

**Tabla 4:** Resultados de los experimentos variando el porcentaje de selección por InfoGain en el enfoque conceptual.

Con todo esto surge la necesidad de evaluar sobre otra colección que se acerque más a las necesidades típicas en el ámbito de la categorización automática de textos, como son

el tener una dimensión mediana-grande, el permitir que un documento esté presente en varias categorías y que no proporcione la desambiguación para los términos que se encuentran en los documentos. La elección, teniendo en cuenta el ámbito del experimento y las necesidades, ha sido la colección de documentos Reuters-21578 que presenta todas las características requeridas y, además, es ampliamente usada en el ámbito de la categorización automática lo cual añade a los experimentos el valor añadido de ser comparables de forma más directa con investigaciones previas.

Para poder realizar los experimentos sobre la representación mediante conceptos de la colección Reuters, primero hay que obtenerla. Para ello, se procesan los documentos originales de la colección obteniendo una serie de conceptos para cada término (para conseguir los conceptos asociados a un término, se usa WordNet y una interfaz de acceso desde java denominada JWNL que permite obtener todos los conceptos asociados gracias al método `lookupAllIndexWords(consulta)` de la clase Dictionary). Ahora es cuando surgen distintos enfoques acerca de cómo tratar estos conjuntos de conceptos:

- Para cada palabra usar todos los conceptos de todas las categorías sintácticas en la representación por conceptos.
- Usar solamente el primer concepto de la primera categoría sintáctica (sustantivo > verbo > adjetivo > adverbio).
- Realizar desambiguación teniendo en cuenta los conceptos posibles y las palabras del entorno. Este enfoque precisa bastante más carga computacional (no ha sido abordado en la presente versión de este documento).

A partir de los conceptos obtenidos y de la propia palabra es fácil encontrar el lexisn que contiene información sobre el archivo lexicográfico en el que se encuentra el concepto pues wordnet aporta un mapping directo mediante el archivo SENSES.IDX. Por lo tanto, para desambiguar las palabras solo nos centraremos en desambiguar los conceptos y a partir de esta desambiguación se obtendrá la información usada en la categorización.

Teniendo definidos los enfoques en cuanto a la problemática de los términos ambiguos, conviene examinar una cuestión derivada del uso típico del lenguaje que no es más que el uso de frases (conjuntos de palabras, generalmente entre 2 y 5 palabras) que tienen una serie de conceptos asociados de por sí que no tienen por qué tener relación con los conceptos que individualmente se asocian a las palabras sueltas. La resolución adoptada es el examinar los textos buscando conceptos asociados a grupos de N palabras y reduciendo ese N hasta 1 mientras no se encuentre un concepto asociado y continuando,

posteriormente, con el siguiente grupo de palabras (compuesto desde la primera palabra que no pertenezca al último grupo de palabras con concepto relacionado; esto ha sido detallado en el punto 3). Se han definido experimentos para los siguientes valores de N:

- N con valor 3 .
- N con valor 5 (experimento no abordado en la revisión actual del presente documento).

Llegados a este punto se tienen dos colecciones de documentos en las que cada documento tiene asociadas un conjunto de categorías a las que pertenece (existen 114 categorías y cada documento pertenece a N categorías, con  $0 \leq N \leq 114$ ). Esto plantea un pequeño problema ya que a la hora de clasificar lo que se hace es determinar para un documento si pertenece o no a una categoría dada. Por tanto no se puede entrenar con esta información tal cual pues no hay un algoritmo de clasificación que nos proporcione el conjunto de categorías a las que pertenece un documento con una cierta correctitud. Así pues, y teniendo en cuenta que Weka lo que realmente recibe es un archivo de índices representación del modelo del Espacio Vectorial sobre el conjunto de documentos, lo que se necesita es construir un archivo de índices por cada posible categoría representando todos los documentos y para cada documento todos sus atributos (palabras/conceptos) en formato `tf*idf` (y, al igual que en SEMCOR, en forma binarizada representando otro enfoque) y si pertenece o no a la categoría dada. Así, pues, se obtendrán 114 pares de archivos `arff` (Weka procesa archivos en formato `arff` [Garner, 95]), `trainI.arff` y `testI.arff` (que representan el conjunto de documentos de entrenamiento y el de prueba para el clasificador de la categoría I).

Una vez preprocesados los documentos según lo descrito anteriormente ya se tienen los dos tipos de representación de los documentos necesarios para la evaluación comparativa. En este punto, con los experimentos de la colección SEMCOR, se definieron los valores de InfoGain a usar para obtener un conjunto de atributos procesables con relativa facilidad. Si con SEMCOR el valor elegido fue del 10%, en los experimentos sobre Reuters, debido al gran número de atributos (básicamente palabras/términos), nos quedaremos con todos los atributos cuyo InfoGain sea mayor que 0 y, en el caso que estos representen más del 1%, nos quedaremos con el 1% mejor. Incluso a pesar que estemos pasando de una reducción del 10% a una reducción del 1%, hemos de contrastar con la comparativa entre 15.000 atributos, 15 categorías y 186 documentos en la representación por palabras en SEMCOR y los casi 30.000 términos, 11.000 documentos en entrenamiento y 114 categorías con los que abruma Reuters.

Teniendo ya en cuenta todas estas decisiones solo falta definir los algoritmos sobre los que se evalúan los distintos

enfoques. Para seleccionarlos simplemente se escogen aquellos que mejor resultado han dado en la colección SEMCOR, realizando una reducción a 5 algoritmos a lo sumo debido a la gran carga computacional que supone trabajar con el conjunto de 114 pares de documentos para cada aproximación. Los algoritmos seleccionados son: Ib2, Ib16, SMO, NaiveBayes y AdaBoost con J48 (C4.5) y 30 iteraciones.

De esta parte del trabajo todavía no se disponen resultados, por lo que queda como una línea de trabajo futuro a seguir.

## 6 Trabajos relacionados

Existen diversos trabajos previos realizados con objetivo similar al del presente documento, es decir, evaluar el uso de conceptos en la clasificación automática de textos. La principal diferencia con todos ellos es que en nuestra propuesta se integra un tipo de información conceptual que no es directamente el concepto asociado si no es una categoría del concepto asociado lo cuál reduce la cantidad de información y generaliza más.

En [Petridis, 01] se estudia el uso de la red neuronal  $\sigma$ -FNLMAP en una comparativa en la que se pretende ver la superioridad de esta red frente a otros algoritmos simples como KNN o NaiveBayes y, a la vez, comparar el uso de palabras enfrentado al uso de conceptos como representaciones de los documentos. Los resultados obtenidos son muy similares a los explicados en la parte de SEMCOR en el presente documento, a pesar que existen varias diferencias en cuanto a la toma de documentos de SEMCOR ya que en [Petridis, 01] se toman todos los documentos de SEMCOR mientras que aquí solo se toman aquellos documentos que tienen todas las palabras etiquetadas, lo cual reduce el número medio de documentos por clase de 20 a 10 lo que hace que se tenga mucha menos información de cada categoría. En [Petridis, 01] se concluye diciendo que el enfoque por conceptos no aporta suficientes mejoras como para tenerlo en cuenta, pero no se comparan las mejoras en cuanto a eficiencia que pueden ser interesantes así como el uso de los conceptos en un entorno en el que el trabajo para la desambiguación esté por realizar.

[Buenaga, 97] si aborda, en parte, el problema de la desambiguación, pero desde un enfoque un tanto más pobre ya que no se trata de desambiguar los términos que se encuentran en los documentos si no se expanden las categorías para posteriormente crear un perfil por cada categoría con los sinónimos obtenidos. A partir de aquí se aplica el algoritmo de Rocchio de forma que con estos perfiles se consigue incrementar la información sobre cada categoría. El enfoque que en el presente documento se propone es totalmente distinto ya que se trata de procesar los documentos tratando sus conceptos y no el expandir la información si no tomar los conceptos adecuados para cada

palabra lo cual podrá, incluso, reducir la cantidad de información utilizada.

## 7 Conclusiones

Realmente el enfoque integrando información conceptual de agrupación de conceptos es un enfoque realmente efectivo y eficiente como se puede ver en los resultados aportados en la evaluación sobre SEMCOR. De todas formas faltaría evaluar la aplicación de todo esto sobre Reuters para evaluar la factibilidad de acoplar la categorización integrando esta información con la desambiguación de las palabras en una colección de textos en las que estas no se encuentran desambiguadas.

La integración de información conceptual de diversos tipos es una línea a seguir pues como bien se puede apreciar en este artículo ofrece unos resultados en cuanto eficiencia y eficacia realmente elevados y sobre todo mejores que el enfoque usando únicamente palabras.

Convendría evaluar también el uso de conceptos únicamente en lugar de usar información acerca de las categorías de palabras pues puede aportar mejores resultados en cuanto a eficacia, aunque seguramente no en cuanto a eficiencia pues la información sobre categorías de palabras permite agrupar muchos conceptos en una sola categoría reduciendo la cantidad de información a usar en el sistema.

## 8 Referencias

- [Beckwith *et al.*, 91] R. Beckwith, C. Fellbaum, D. Gross, G. Miller. "WordNet: A lexical database organized on psycholinguistic principles". En "Lexical acquisition: Exploiting On-Line resources to build a lexicon".
- [Buenaga, 97] M. Buenaga Rodríguez, J. M. Gómez Hidalgo, B. Díaz Agudo. "Using WordNet to complement training information in text categorization".
- [Garner, 95] Stephen R. Garner. "The Waikato Environment for Knowledge Analysis". University of Waikato, New Zealand.
- [Lewis, 92] D. D. Lewis. "Representation and Learning in information retrieval". Unpublished PhD thesis, University of Massachusetts.

- [Lewis, 99] D. D. Lewis. "Reuters-21578 text categorization test collection distribution 1.0".
- [Miller, 90] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. "References Introduction to WordNet: An on-line Lexical Database", *International Journal of Lexicography*, vol. 3.
- [Petridis, 01] V. Petridis, V. G. Kaburlasos, P. Frangkou, A. Kehagias. "Text classification using the  $\sigma$ -FLNMAP Neural Network".
- [Rada, 98] Rada Mihalcea. "SEMCOR, semantically tagged corpus".
- [Salton *et al.*, 01] G. Salton, A. Wong, and C. S. Yang. "A vector space model for automatic indexing". *Communications of the ACM*.
- [Witten, 99] Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham: "Weka: Practical Machine Learning Tools and Techniques with Java Implementations". University of Waikato, New Zealand.