Recibido, 28/04/2010 Aceptado, 23/07/2010

Automatización de procesos de categorización jerárquica documental en las organizaciones

[1]Rocío Rocha [a], Ángel Cobo [b]

RESUMEN

En un contexto global y caracterizado por el uso masivo de las tecnologías de la información y las comunicaciones, para cualquier organización resulta esencial la optimización de los procesos de búsqueda y gestión de repositorios documentales. En este trabajo se realiza un análisis de modernas técnicas de gestión documental que convenientemente combinadas con estrategias computacionales y recursos lingüísticos especializados permiten abordar la categorización automática de documentos en las organizaciones. Como caso particular se describe un sistema de clasificación diseñado de acuerdo a la taxonomía JEL (Journal of Economic Literature) y que hace uso de glosarios multilingües para realizar clasificaciones jerárquicas de documentos científico-técnicos vinculados con las áreas funcionales de la empresa.

Palabras clave: Categorización; Taxonomías; Minería de texto; Gestión documental

ABSTRACT

In a global context characterized by the massive use of information technology and communications any organization needs to optimize the search and document management processes. In this paper an analysis of modern document management techniques and computational strategies with specialized language resources is presented and a model that can be used in automatic text categorization in the context of organizations is proposed. As a particular case we describe a classification system according to the taxonomy JEL (Journal of Economic Literature) and that makes use of multilingual glossaries for hierarchical classifications of scientific and technical documents related to the business functional areas.

Key words: Categorization; Taxonomies; Text mining; Document management

[1] Dpto. de Administración de Empresas. Universidad de Cantabria, Santander, España.
[a] Licenciada en Informática, [b] Licenciado en Matemáticas

INTRODUCCIÓN

En un ambiente caracterizado por el uso intensivo de las TIC y las enormes facilidades para el intercambio de información en entornos globales, una adecuada gestión de la información exige la utilización de modernos métodos y soluciones tecnológicas para el almacenamiento, clasificación y administración de colecciones documentales, así como para la extracción de conocimiento a partir de ellas. Uno de los principales problemas a los que se enfrenta la sociedad de la información es la gestión óptima y productiva de la documentación disponible. Las organizaciones necesitan establecer técnicas que ayuden a localizar y organizar la información relevante y para que su recuperación sea lo más efectiva posible. Los sistemas de gestión documental resultan adecuados para la recuperación de los documentos o la información que éstos contienen o, en definitiva, hacer accesible el conocimiento.

La categorización es una técnica documental que pretende agrupar los documentos por su contenido con objeto de facilitar su situación y búsqueda. Implica, por tanto, asignar los documentos, lo más exactamente posible, a una rama del saber, de manera que queden agrupados con documentos similares. Para ello se utilizan sistemas de clasificación que dividen un dominio de la realidad en una serie ordenada de clases y subclases, organizadas jerárquicamente. Clasificar un documento implica encuadrarlo en un lugar exacto dentro del sistema de clasificación que se utilice. En este trabajo se pretende mostrar la enorme potencialidad del uso de técnicas de minería de texto combinadas con la utilización de recursos lingüísticos especializados para facilitar la gestión y organización de colecciones documentales y la generación de conocimiento.

RESULTADOS

Clasificación documental y categorización de texto

El proceso de clasificación se podría definir como *"...el acto de organizar el universo del conocimiento en algún orden sistemático. Ha sido considerada la actividad más fundamental de la mente humana. El acto de clasificar consiste en el dicotómico proceso de distinguir cosas u objetos que poseen cierta característica de aquellos que no la tienen y agrupar en una clase cosas u objetos que tienen la propiedad o característica en común (Chan1981)"*.

En el caso particular de los documentos éstos pueden ser clasificados de acuerdo a diferentes criterios, por supuesto, también en base a sus contenidos, buscando en este caso la asignación de cada documento a una o varías categorías temáticas. La clasificación temática conduce muchas veces a sistemas ambiguos en los que las categorías no están definidas de forma exacta. En estos casos, la clasificación tiene un importante componente subjetivo.

La clasificación más antigua que se conoce es la del *templo de Horus* en Egipto. Pero el primer sistema de clasificación bibliográfica como tal es el de *Conrad Gessner* en el siglo XVI. Es en el siglo XIX cuando empezaron a aparecer los grandes sistemas de clasificación bibliográfica. Algunos de estos sistemas se deben a destacados bibliotecarios del siglo XIX como Cutter o Dewey. El desarrollo de Internet en las últimas décadas ha contribuido al desarrollo de nuevos sistemas de clasificación que faciliten la organización y recuperación de sus recursos electrónicos. Actualmente, existen diferentes vías para la localización de estos recursos, por un lado, se utilizan motores de búsqueda basados en palabras clave, cuyo principal inconveniente está en la tendencia a recuperar información irrelevante en muchos casos. La alternativa es la creación de directorios organizados temáticamente que faciliten al usuario la navegación por la Web. Para ello es preciso adoptar esquemas de clasificación que proporcionen la necesaria estructura semántica.

Tipologías de sistemas de clasificación

Se puede entender un sistema de clasificación como la segmentación y estructuración arbitraria del conocimiento humano, con el objeto de crear categorías y clases de temas, organizando de forma temática el conocimiento en un gran esquema que abarque y comprenda las distintas facetas del mismo. Los sistemas de clasificación deben caracterizarse por una serie de propiedades como su estructuración jerárquica y coherente, amplitud del área de conocimiento, ser explícitos y concisos, y fácilmente adaptables.

Los sistemas de clasificación pueden clasificarse en base a varios criterios. Así, según el ámbito temático se distinguen los sistemas generales y especializados. Los primeros, también conocidos como universales o enciclopédicos, tratan de cubrir todos los ámbitos del conocimiento; su principal problema es precisamente la búsqueda de una validez universal, entrando en conflicto con el carácter dinámico de la actividad científica en una sociedad moderna como la actual. Ejemplos de sistemas de clasificación bibliográfica de carácter

universal son el *sistema Dewey Decimal Classification* (DDC), el *Universal Decimal Classification* (UDC) o el desarrollado por la Biblioteca del Congreso de Washington el *Library of Congress* (LCC). Los sistemas de clasificación especializados se limitan, en cambio, a un campo de conocimiento específico, buscando una mejor adaptación a dicha disciplina. Un ejemplo de sistema de clasificación específico es el *sistema JEL* para la clasificación de documentos de carácter económico, sistema que se ha utilizado en este trabajo de investigación para automatizar los procesos de categorización de documentos económicos.

Según su estructura, los sistemas de clasificación se dividen en jerárquicos y facetados. El modelo de clasificación jerárquico utiliza un método de organización muy simple inspirado en la lógica clásica. Se comienza por un campo del conocimiento especializado o universal al que se le va aplicando un criterio de división. En este campo se establecerán distintas clases y en sus epígrafes se enumerarán todos los conceptos y todas las materias que se pueden clasificar con ese sistema. En las distintas clases se aplicarán nuevas subdivisiones según nuevos criterios de división. Las relaciones semánticas que se establecen en este sistema son de tipo vertical o jerárquico. Las principales ventajas de este modelo de clasificación son la simplicidad de los esquemas, facilidad de uso y una gran claridad y difusión en el ámbito internacional. Por otro lado, también presenta algunos inconvenientes como por ejemplo la rigidez del sistema, dónde solo están presentes relaciones semánticas de tipo jerárquico en la que no se permite la relación de conceptos de distintas clases. Otro de los inconvenientes es la difícil incorporación de conceptos o nuevos temas.

Los sistemas de clasificación facetados, o también denominados *multijerárquicos*, están ligados a la insuficiencia del método jerárquico simple y se basan en el principio de que cualquier tema se puede descomponer en conceptos simples que compartan un atributo o propiedad común (Gödert1991). Cada uno de estos grupos simples se denomina faceta. Las facetas son contenedores mutuamente excluyentes de categorías gracias a las características de división, es decir que una categoría no puede pertenecer a dos facetas diferentes; además deben cumplir la característica de homogeneidad. En cada una de las facetas se puede hacer subdivisiones aplicando nuevos criterios lógicos que dan origen a las subfacetas. Dentro de cada grupo, los conceptos se pueden organizar jerárquicamente y alfabéticamente o aplicando cualquier criterio clasificatorio. Este tipo de sistemas tiene varias ventajas: permite una clasificación precisa de los conceptos, tiene una estructura clara y sencilla, se adapta a la renovación de las ciencias y a la incorporación de nuevos conceptos y permite una amplia explotación en el campo de la recuperación de información. Por otro lado, también tiene algún inconveniente, como una escasa difusión a nivel práctico; en cambio en el aspecto teórico ha sido objeto de un gran desarrollo y formalización. Entre los sistemas facetados más importantes se pueden citar la *clasificación colonada* del bibliotecario hindú S. R. Ranganathan y la *clasificación de Bliss*.

A partir de estos sistemas se ha creado muchos sistemas de clasificación especializados. La Figura 1 trata de ilustrar la diferencia entre un sistema de clasificación jerárquico simple y un sistema facetado.

Otra alternativa consiste en combinar las dos estructuras de clasificación anteriores, para generar sistemas de clasificación híbridos. Como resultado no se consigue la integración total de los dos modelos, sólo la incorporación de facetas a los sistemas jerárquicos predominando esta misma estructura. Este enfoque híbrido es el que siguen sistemas de clasificación ampliamente difundidos como el *Dewey Decimal Classification* (DDC).

Clasificación documental automática

La labor de clasificación de documentos mediante un sistema de clasificación documental no es una tarea sencilla, hasta el punto de intervenir muchas veces una componente subjetiva, ya que, por ejemplo, un mismo documento podría ser clasificado de diferentes maneras por dos expertos diferentes en catalogación de fondos bibliográficos. Por otro lado, cuando el número de documentos a clasificar es elevado o se requiere una respuesta rápida en el proceso de clasificación, la tarea se hace aún más compleja. Es por ello que disponer de sistemas de clasificación automática es una necesidad para muchas organizaciones.

La clasificación automática de documentos, también conocida como categorización de textos o *topic spotting*, es la tarea de asignar automáticamente un conjunto de documentos a una o más categorías preexistentes a partir de un conjunto de documentos clasificados por expertos sobre los que el sistema lleva a cabo un proceso de aprendizaje supervisado (Sánchez, 2007).

La clasificación manual de un documento implica

tres etapas claramente diferenciadas (Gil 1996): examinar un documento para tener una idea clara de sus contenidos, sintetizar dichos contenidos en un tema principal, y, finalmente, contrastar este tema con un lenguaje clasificatorio para determinar la categoría más próxima, representando además el documento mediante la notación propia de la clasificación. A la hora de desarrollar un sistema de clasificación automática, una primera dificultad a salvar tiene relación con la primera de las etapas anteriores. Examinar el documento para obtener una idea clara de sus contenidos. Puede verse en este caso como el proceso de identificar rasgos relevantes del documento (palabras, términos específicos,...) y realizar una ponderación de los mismos para reflejar su importancia. En este caso, las técnicas de minería de texto permitirán obtener una representación vectorial de los documentos que facilite la comparación y análisis de similitud con otros documentos (Baeza y Ribeiro 1999). La segunda etapa del proceso de clasificación manual es la que claramente se trata de un sistema artificial que no puede realizar algo como un experto humano. Finalmente, la identificación de la categoría a asignar el documento; en el caso de un experto humano consistiría en comparar la materia del documento con las materias del sistema de clasificación hasta encontrar una correspondencia. De nuevo, el procedimiento a seguir por un clasificador automático debe ser diferente.

El comportamiento inteligente de un clasificador automático no debe verse como un intento de emular el comportamiento del ser humano, sino de aprovecharse de él para llegar a las mismas conclusiones pero por caminos diferentes. Este aprovechamiento se plasma en la utilización de conjuntos de documentos de entrenamiento que hayan sido preclasificados por expertos. El sistema deberá utilizar esta información para ser capaz de deducir clasificaciones para nuevos documentos. En definitiva, para poder plantear la construcción de sistemas de clasificación documental automática se requiere de una base de conocimiento, y una serie de técnicas de procesamiento, extracción de rasgos, representación de los documentos y algoritmos de aprendizaje que permitan automatizar las tareas propias de dicha labor.

Caso Práctico: desarrollo de un sistema de clasificación de literatura económica de acuerdo a la taxonomía JEL

Como ejemplo práctico de diseño de un clasificador documental automático que pueda ser útil para optimizar la gestión documental en las organizaciones, se presenta a continuación un completo sistema para clasificar documentos de carácter económico de acuerdo a la taxonomía JEL. Esta taxonomía fue elaborada por la *Journal of Economic Literature* (JEL), publicada desde 1969 bajo el auspicio de la *American Economic Association* (http://www.aea-web.org). Para facilitar la clasificación de los artículos y trabajos científicos publicados en ella, se desarrolló un sistema de clasificación que, con el tiempo, se ha convertido en un estándar de clasificación en el campo de la Economía: *el sistema de clasificación JEL*. Este sistema está estructurado actualmente en 20 categorías principales, que se subdividen a su vez en 131 subcategorías y 774 subsubcategorías. De esta forma el sistema consta de tres niveles de clasificación, que van de lo más general a lo más específico. Cada categoría está identificada por un código alfanumérico. En el caso de las categorías de primer nivel (las que identifican los campos generales), este código está formado por una única letra. Para las subcategorías, el código alfanumérico se forma a partir del código de su categoría raíz, incluyendo un dígito numérico.

La taxonomía JEL no debe verse como un sistema de categorías excluyentes, sino que lo habitual es la existencia de áreas de solapamiento entre las diferentes categorías. Todo sistema de clasificación automática que pretenda realizar categorizaciones de documentos económicos de acuerdo a este sistema debe tener presente esta circunstancia. En este trabajo, de hecho, el algoritmo de clasificación asigna a cada documento varios códigos de categoría, con una ponderación del grado de afinidad del documento clasificado con la categoría en cuestión.

En este trabajo de investigación se ha utilizado diferentes estrategias de representación y técnicas computacionales para asignar de manera automática a cada documento una o varias categorías o códigos JEL a partir del análisis de sus rasgos y la comparación de la similitud con documentos cuya clasificación se conoce de antemano (*corpus de entrenamiento*). Para realizar este análisis de similitud, el documento a clasificar debe ser procesado y representado conforme a los mismos criterios aplicados sobre los documentos de entrenamiento.

Representación vectorial de documentos y análisis de similitudes

Con el objeto de poder aplicar técnicas computacionales, es preciso disponer de un sistema

de representación numérica de los documentos. El modelo vectorial, propuesto por (Salton, 1971) es un estándar en minería de texto. Consiste en representar todo documento por un vector cuyas componentes miden el peso o grado de relevancia de cada rasgo identificado en el mismo. Los rasgos pueden ser el conjunto completo de palabras, pero habitualmente se reduce la dimensión utilizando procesos de lematización o reducción de las palabras a sus raíces gramaticales. Otra estrategia puede ser seleccionar rasgos a partir de recursos lingüísticos especializados como tesauros o glosarios. Para el cálculo de los pesos también es posible utilizar diferentes estrategias, una de las más comunes es el *esquema de ponderación tf-idf*, por el cuál el peso se calcula como producto de dos factores, uno de ellos mide la importancia de ese rasgo en el propio documento y el otro la importancia de ese rasgo en el resto de documentos de la colección. Este esquema es definido de la siguiente manera:

$$w_{ij} = f_{ij} \times idf_i = \frac{freq_{i,j}}{\max_p freq_{p,j}} \log \frac{N}{n_i} \qquad (1)$$

donde $freq_{i,j}$ es el número de veces que el rasgo k_i aparece en el documento d_j, N es el número total de documentos y n_i el número de documentos en los que el rasgo k_i aparece. En (Baeza y Ribeiro 1999) puede encontrarse información más detallada de estos procesos.

En el caso concreto de los experimentos realizados en el contexto de este trabajo, se optó por identificar dos grupos de rasgos que fueron ponderados con el esquema tf-idf anterior:

- Palabras identificadas en el documento tras un proceso previo de filtrado (eliminación de *stopwords*), seleccionando únicamente los sustantivos, adjetivos y verbos. Además se utilizó la herramienta de análisis lingüístico *TreeTagger* para reducir cada palabra a un representante de su familia léxica, así por ejemplo, la herramienta reduce toda forma verbal a su infinitivo, o los sustantivos a su forma masculina y singular.
- Términos y expresiones localizadas en el texto y presentes en un glosario de términos económicos. Se utilizó el glosario económico del Fondo Monetario Internacional (http://www.imf.org/external/np/term/index.asp). Se trata de un recurso lingüístico interesante, de carácter multilingüe y que cuenta con más de 11.600 registros.

La ventaja de disponer de una representación vectorial de los documentos es la posibilidad de utilizar diferentes métricas o distancias en el espacio vectorial para calcular similitudes entre documentos. Una de las métricas más utilizada en minería de texto es la conocida como separación angular o medida del coseno. Esta métrica calcula la similitud de dos documentos como el producto escalar de sus respectivas representaciones vectoriales, siempre que dichos vectores hayan sido normalizados previamente. Existen otras medidas de similitud posibles, pero la medida del coseno ha demostrado mejor eficacia en diferentes problemas de minería de texto (Michel 2001; Egghe y Michel 2002). En el caso de este trabajo, al tener cada documento dos vectores asociados, se puede tomar como medida una combinación lineal convexa de las correspondientes medidas sobre cada grupo de rasgos.

Base de conocimiento: corpus de entrenamiento

Resulta esencial para el correcto funcionamiento de un clasificador contar con un conocimiento codificado de expertos de elevada calidad; lo que se traduce en una base de conocimiento con un buen número de documentos correctamente clasificados. En este caso, para la construcción del corpus de entrenamiento que actúa como base de conocimiento se recurrió al *servicio IDEAS* (http://ideas.repec.org) ofrecido por *RePEc* (Research Papers in Economics). Este servicio es una gran base de datos descentralizada de acceso libre y contiene artículos científicos, componentes de software, revisiones de libros,... Se trata de la mayor base de datos de bibliografía económica disponible libremente a través de Internet.

El corpus extraído de IDEAS para esta investigación está formado por una colección de artículos científicos de diversas áreas de la economía y las ciencias empresariales escritos en inglés y clasificados de acuerdo a la taxonomía JEL. Este corpus está formado por 445 documentos, y asociados a 50 categorías JEL diferentes previamente seleccionadas. Los diferentes documentos se encuentran clasificados por el propio sistema JEL, estando en principio asignados cada uno de ellos a una única categoría. No obstante, es de destacar el hecho de que muchos de ellos, en su propio contenido, incluyen asociaciones a más de una categoría realizadas por los propios autores.

Como complemento, y con el objeto de probar la efectividad del modelo de clasificación propuesto, se recopiló igualmente un pequeño corpus de prueba formado por 74 documentos asociados a las categorías presentes en el corpus de entrenamiento.

En este caso, se debe considerar como clasificación correcta la realizada por los propios autores de los documentos, al incluir junto con el resumen del trabajo la codificación JEL más apropiada a su juicio. Estos documentos son todos ellos artículos de revistas especializadas e incorporan diferentes códigos JEL cada uno, de hecho, como media los 74 documentos tienen asociadas 4,05 categorías JEL. Este hecho de pertenencia a varias categorías simultáneamente debe tenerse en cuenta al valorar la efectividad del sistema de clasificación automática.

Algoritmo de clasificación

El proceso de clasificación desarrollado se basa en realizar análisis de similitudes medias de los documentos a clasificar con aquellos que se encuentran en la base de conocimiento y que están asignados a una misma categoría. El algoritmo utiliza un parámetro de umbral mínimo que determina si un documento puede ser asignado o no a una categoría. Cuando la similitud media de los documentos de una categoría presentes en la base de conocimiento con respecto al documento a clasificar no supere dicho umbral la asignación será descartada. Cuando esa similitud media supere el umbral establecido se mostrará al usuario la asignación correspondiente, pudiendo éste aceptarla o rechazarla. En el momento de aceptar la clasificación de un documento éste pasará a incorporarse a la base de conocimiento para ser utilizado en futuros procesos de clasificación. El algoritmo permite además limitar el número máximo de categorías a las que se puede asociar un documento.

Resultados experimentales

Tomando como referencia el corpus de prueba, a continuación se resumen los resultados experimentales más destacados de la clasificación automática de sus 74 documentos, así como algunas conclusiones a las que se puede llegar a la vista de los mismos. En primer lugar se decidió comparar los resultados de la clasificación al calcular la similitud sobre los 2 grupos de rasgos diferentes: palabras lematizadas y términos del glosario.

La clasificación de los 74 documentos de prueba se realizó a tres niveles, correspondientes a los tres niveles jerárquicos de categorías que define el sistema JEL. En todos los casos se consideró un umbral de similitud de $sim_{umbral}=0.01$ para aceptar una clasificación, pero fijando diferentes valores para el número máximo de categorías a asignar ($Nmax_{cat}$) en función del nivel jerárquico de las categorías. Los valores de este parámetro fueron 3, 4 y 6 para los diferentes niveles jerárquicos de clasificación.

DISCUSIÓN

A la hora de analizar la efectividad de la clasificación automática se consideró como error de clasificación los casos en los que ninguna de las categorías asignadas por el sistema correspondía con alguna de las asignadas directamente por los autores del trabajo. Como clasificaciones correctas se consideraron aquellas en la que la categoría con mayor promedio de similitud coincidía con alguna de las proporcionadas por los autores del documento, y se consideraron clasificaciones aceptables aquellas en las que aunque la primera categoría ofrecida por el sistema no coincidía, sí que lo hacían alguna de las restantes dadas por el sistema. En la Tabla 1 se resumen los resultados obtenidos. En el caso de la clasificación jerárquica de mayor nivel la tasa de error resultó menor cuando se trabaja con la representación vectorial de los documentos a partir del conjunto completo de sus palabras lematizadas; sin embargo, en clasificaciones de nivel jerárquico menor la representación a partir de los registros del glosario especializado iguala la efectividad de la clasificación a partir del conjunto completo de palabras. En el caso de la clasificación al nivel jerárquico más bajo del sistema JEL, es decir, el tercer nivel o el correspondiente a lo más específico, hay que tener en cuenta que el sistema JEL define un total de 774 categorías, muchas de ellas con un alto grado de afinidad, por lo que la tarea de clasificación resulta aún más dificultosa. A la vista de los resultados obtenidos se puede decir que se confirma de nuevo que ante una clasificación a un nivel más específico, como es este tercer nivel, la utilización del glosario de términos económicos ofrece claramente mejores resultados que la utilización del conjunto completo de palabras que aparecen en los documentos. Las tasas de errores en la clasificación es lógico que resulten más altas cuanto más bajo sea el nivel jerárquico, ya que se entra a analizar divisiones más finas entre las categorías y con unas afinidades mayores entre ellas. La Figura 2 muestra gráficamente la evolución en las tasas de error de clasificación al realizar las clasificaciones a distintos niveles jerárquicos, como parece claro, a niveles más específicos de categorización la dificultad de diferenciar determinadas categorías aumenta y por tanto las tasas de error también se incrementan. La figura también pone de manifiesto el hecho de los mejores resultados de clasificación de tercer nivel utilizando el glosario de términos económicos.

Además de las menores tasas de error al clasificar mediante el glosario, la utilización de exclusivamente términos económicos para representar los documentos ofrece dos importantes ventajas respecto a la opción de utilización del conjunto completo de palabras:

- El trabajar sobre un espacio vectorial de menor dimensión. A pesar de haber eliminado un buen número de stopwords, lematizadas las palabras y seleccionadas únicamente los sustantivos, adjetivos y verbos, el tamaño del diccionario de palabras asociado al corpus fue de 28.856 palabras, lo que significa que cada documento se representa por un vector de esa misma dimensión. En cambio el tamaño del glosario de términos económicos del FMI es de poco más de 11.600. Trabajar sobre un espacio vectorial de menor dimensión supone ahorro de espacio de almacenamiento y tiempo de cómputo.
- El carácter multilingüe del glosario económico utilizado aporta esa misma característica al clasificador, ya que la representación vectorial de los documentos será independiente del idioma y podrán ser comparados documentos aunque estén escritos en idiomas diferentes.

CONCLUSIONES

1. A la vista de los resultados experimentales, se podría llegar a algunas conclusiones interesantes. A la hora de clasificar documentos de áreas muy próximas interesa utilizar herramientas lingüísticas específicas del área. El esfuerzo que supone realizar un procesamiento completo del conjunto de palabras de los documentos, con la consiguiente eliminación de stopwords, lematización de palabras y selección de rasgos, supone un excesivo coste computacional y de almacenamiento que no se justifica al conseguirse resultados similares, o incluso superiores identificando como rasgos únicamente los que pertenezcan al glosario específico.
2. Evidentemente, la calidad del clasificador depende fuertemente de la calidad de la base de conocimiento. A ese respecto interesa contar con un conjunto elevado de documentos que se encuentren correctamente clasificados. El tamaño de la base de conocimiento utilizada en los experimentos es excesivamente pequeño y limitado a un subconjunto de categorías, por lo que convendría ampliar la base. A pesar de esas limitaciones puede considerarse que los resultados de clasificación obtenidos sobre el conjunto de documentos de prueba son satisfactorios.
3. En la fijación de un umbral mínimo para la asignación de documentos a categorías, las pruebas experimentales realizadas indican que convendría utilizar umbrales diferentes para las distintas categorías. Existen categorías en las cuales las similitudes entre los documentos pertenecientes a ellas son más altas que en otras con un carácter más generalista o incluso multidisciplinar, en tales casos convendria fijar umbrales de similitud más altos para proceder a la asignación. Incluso, fijar esos umbrales a partir de similitudes medias de documentos en la base de conocimiento.

REFERENCIAS BIBLIOGRÁFICAS

- Baeza, R., y Riveiro, B. 1999. *Modern Information Retrieval*. Addison Wesley.
- Chan, M. L. 1981. *Cataloging and classification: an introduction*. Mc Graw-Hill.
- Egghe, L., y Michel, C. 2002. Strong similarity measures for ordered sets of documents in information retrieval. *In Information Processing and Management*, p. 823-848.
- Gil, B. 1996. *Manual de lenguajes documentales*. NOESIS. p. 17-22
- Gödert, W. 1991. Facet classification in online retrieval. *International Classification*, 2(18):98-105.
- Hassan, Y., y J.Martín, F. 2003. Clasificaciones facetadas y metadatos: Conceptos básicos. No Solo Usabilidad (http://www.nosolousabilidad.com).
- Michel, C. 2001. Ordered similarity measures taking into account the rank of documents. *Information Processing Management*, 37(4):603-622.
- Salton, G. 1971. *The SMART Retrieval Sistem. Experiments in Automatic Document Processing*. Prentice Hall.
- Sánchez, R. 2007. La documentación en el proceso de evaluación de sistemas de clasificación automática. *Documentación de las Ciencias de la Información*, 30:25-44.

Correspondencia:
Rocío Rocha
Av. De los Castros s/n E-39005
Santader - Cantabria - España
rochar@unican.es

ANEXO

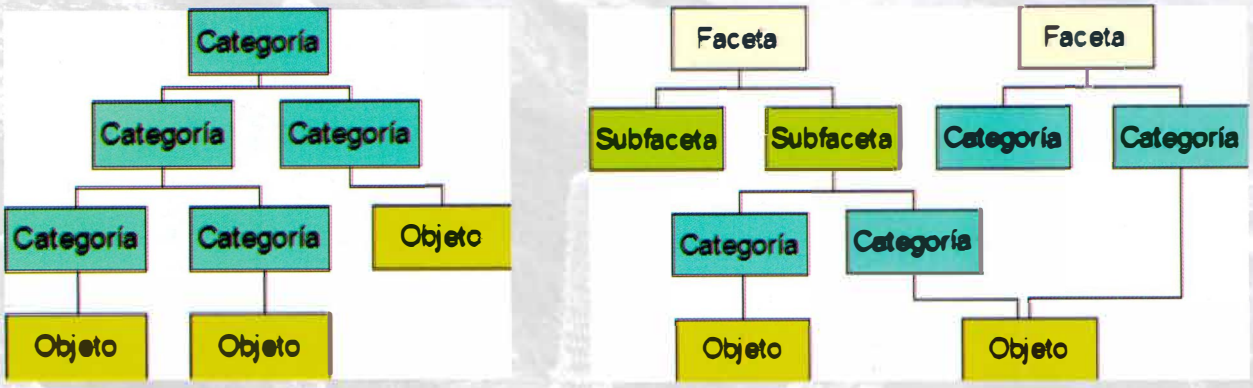

Figura 1. Clasificación jerárquica simple (izquierda) y facetada (derecha). Fuente: Hassan y Martín 2003.

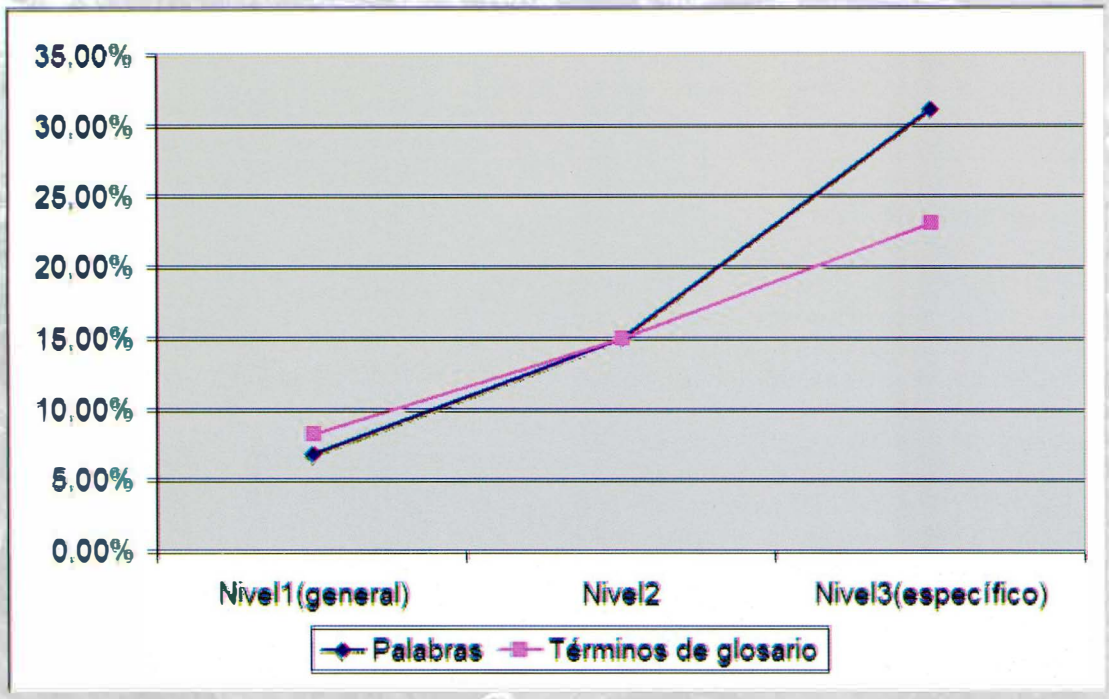

Figura 2. Evolución de la tasa de error de clasificación en los tres niveles jerárquicos del sistema JEL.

Tabla 1. Resultados de la clasificación a nivel de supercategorías, subcategorías y nivel específico (nivel 3).

	Clasificación de nivel 1		*Clasificación de nivel 2*		*Clasificación de nivel 3*	
Tipo de rasgos	palabras	glosario	palabras	glosario	palabras	glosario
% correctas	67.57%	71.62%	60.81%	59.46%	37.84%	43.24%
% aceptables	25.67%	20.27%	24.33%	25.68%	31.08%	33.79%
% erróneas	6.76%	8.11%	14.86%	14.86%	31.08%	22.97%