**Rect@ Vol 10 Diciembre 2009. Pp 105-124**

**Recibido**

***30/10/2009***

**Revisado**

**23/11/2009**

**Aceptado**

**4/12/2009**

# Descubrimiento de conocimiento en repositorios documentales mediante técnicas de Minería de Texto y Swarm Intelligence

Cobo Ortega, Angel acobo@unican.es (*)
Rocha Blanco, Rocio rochar@unican.es (**)
Alonso Martínez, Margarita alonsom@unican.es (**)
*(*) Departamento de Matemática Aplicada y Ciencias de la Computación*
*(**) Departamento de Administración de Empresas*
*Universidad de Cantabria*

**RESUMEN**

El uso combinado de metodologías de minería de texto y técnicas de Inteligencia Artificial favorece los procesos de gestión documental y optimiza los mecanismos de categorización, extracción automática de conocimiento y agrupamiento de colecciones documentales. En el trabajo se propone un modelo de gestión documental integral para el proceso de información no estructurada. Se utilizan glosarios y tesauros especializados para establecer relaciones semánticas entre los términos, y técnicas de Swarm Intelligence para la extracción del conocimiento. El modelo ha sido implementado en una aplicación de uso intuitivo, multilingüe e integradora de técnicas de minería de texto

***Palabras claves*:** Minería de texto; gestión documental; swarm intelligence

## ABSTRACT

The combined use of text mining methodologies and Artificial Intelligence techniques articulate document management processes to optimize categorization mechanisms, automatic knowledge extraction and grouping document collections. The article proposed an integral document management model to process unstructured information. In this context, semantic relations in document collections are implemented by

specialized thesaurus and glossaries, and knowledge feature extraction are facilitated by Swarm Intelligence techniques. The model has been implemented in an intuitive, integral and multilingual text mining application
***Keywords***: text mining; document management; swarm intelligence

# 1. INTRODUCCIÓN

Una de las características más importantes de la sociedad moderna es el papel clave que la información desempeña en los procesos de toma de decisiones. En un mundo globalizado como es en el que desarrollan sus actividades las organizaciones en la actualidad, el disponer de la información adecuada y en el momento adecuado puede suponer una clara ventaja competitiva para éstas. La información permite conocer lo que está ocurriendo y lo que va a ocurrir, y es un elemento esencial para la planificación, operación y control efectivo de las actividades de cualquier organización. Por ello, las organizaciones necesitan cada vez más del uso inteligente de la información y de las tecnologías que permiten una gestión eficiente de la misma.

Hoy en día se genera más información de la que físicamente somos capaces de almacenar y ésta fluye con rapidez a través de las redes de comunicación como Internet. Estudios realizados por consultoras como IDC-EMC (Gantz et al., 2008) ponen de manifiesto el crecimiento exponencial en la generación de nueva información en los últimos años. Especialmente destacado es el volumen de nueva información que se genera en los servicios de la Web 2.0 (foros, redes sociales, wikis, blogs,…). Aunque es difícil de cuantificar, a principios de este año se estimaba que el número de servidores Web activos en Internet superaba los 215 millones y se había alcanzado la cifra de 100 millones de blogs en Internet, con una tasa de crecimiento de varios miles de nuevos blogs al día.

Esa enorme explosión de información trae consigo dificultades para el acceso a la información realmente relevante y para una adecuada gestión de la misma. Hoy en día los buscadores de Internet, por ejemplo, disponen de mecanismos de búsqueda cada vez más sofisticados, y los servicios de sindicación de contenidos (RSS) permiten que la información llegue al usuario y no que éste sea el que realice activamente el proceso de búsqueda. Las organizaciones están suscritas a servicios en línea de bases de datos y recursos de información, recibiendo de manera totalmente automatizada información que se plasma en documentos electrónicos. De hecho, la mayor parte de la información que manejan las organizaciones corresponde a información no estructurada, información que no puede ser almacenada en un formato estructurado como el que establecen las bases de datos. Ejemplos de información no estructurada serían los contratos, presupuestos, correos electrónicos, informes, ofertas, órdenes de compra, facturas y recibos, contenidos en Internet, materiales de marketing, artículos de prensa,

etc. Esta información no estructurada en las empresas representa un alto valor estratégico, es imprescindible en la toma de decisiones, sirve como base de muchos procesos de negocio y favorece el trabajo colaborativo. Para una adecuación gestión de esta información no estructurada existen soluciones tecnológicas como los sistemas de gestión documental, que permiten crear repositorios de documentos, controlar el acceso a los mismos por los usuarios de la organización, realizar seguimiento de versiones, etc. Sin embargo, la extracción de verdadero conocimiento a partir de grandes volúmenes de documentos exige integrar en estos sistemas modernas metodologías como la minería de texto o la aplicación de técnicas de inteligencia artificial.

## 2. GENERACIÓN DE CONOCIMIENTO Y MINERÍA DE TEXTO

La información en sí misma es útil para las organizaciones cuando se contextualiza, la información asociada a un contexto y a una experiencia se convierte en conocimiento convirtiéndose un recurso intangible que aporta verdadero valor a la organización. Se podría decir que los datos están localizados en el mundo y el conocimiento en agentes (personas, organizaciones), mientras que la información adopta un papel mediador entre ambos. El conocimiento asociado a una organización y a una serie de capacidades organizativas se convierte en su *Capital Intelectual*.

La generación de capital intelectual a partir de información que llega a las organizaciones de forma masiva y automatizada no es una tarea sencilla, ante estos enormes volúmenes de información no estructurada que almacenan los repositorios documentales se necesitan sistemas automatizados que permitan extraer conocimiento a partir de ella. En particular, las técnicas de minería de texto permiten explorar y extraer conocimiento de colecciones de documentos textuales, (Baeza y Ribeiro, 1999). Los tres problemas básicos que pueden abordarse con técnicas de minería de texto son:

- Recuperación de información relevante, es decir, extraer de manera automática aquellos documentos que puedan resultar interesantes para el usuario a partir de una consulta realizada por éste. Esta labor es la que realizan, por ejemplo, los buscadores de Internet.
- Categorización de documentos, consiste en asignar a cada documento una o varias categorías temáticas de entre un conjunto de categorías preestablecido.

- Clustering, consiste en la generación automática de grupos de documentos relacionados, por ejemplo, documentos que traten un mismo tema o asunto. A diferencia de lo que ocurre en la categorización, en los procesos de clustering no existe un conjunto de categorías preestablecido, sino que el propio algoritmo a utilizar debe generar automáticamente esas categorías, contribuyendo de esta forma a generar un nuevo conocimiento.

Tanto la categorización como el clustering pueden verse como un proceso de clasificación, en el primer caso se habla de clasificación supervisada mientras que en el segundo se utiliza el concepto de clasificación no supervisada (Brücher, et al. 2002). El objetivo principal de la clasificación de documentos, como concepto global, es reducir la diversidad de datos y la sobrecarga de información mediante la agrupación de documentos similares. Con respecto a la gestión del conocimiento, la clasificación de documentos puede ser vista como una herramienta que permite simplificar el acceso y procesamiento del conocimiento explícito, facilitando la recuperación, organización visualización, desarrollo e intercambio de conocimientos.

Un primer aspecto a salvar a la hora de afrontar la integración de técnicas de minería de texto en los sistemas de gestión documental es la necesidad de disponer de modelos de representación de documentos que permitan la aplicación de técnicas numéricas sobre ellos. El modelo vectorial propuesto por (Salton, 1971) permite representar los documentos a partir de un vector de pesos asociados a una serie de rasgos seleccionados del documento. Habitualmente estos rasgos se obtienen a partir de las palabras presentes en el texto tras realizar diferentes operaciones de filtrado, eliminación de palabras con poco valor discriminante y transformaciones morfológicas, reduciendo el tamaño del diccionario total de palabras utilizado. En el modelo vectorial cada rasgo, por tanto, representa una dimensión del espacio. En el caso de colecciones de documentos escritos en diferentes idiomas resulta interesante la utilización de diferentes recursos lingüísticos (glosarios, tesauros, ontologías,…) para representar los documentos mediante rasgos independientes del idioma (Steinberger et al, 2005) y de esta forma poder aplicar las técnicas de minería de texto con independencia del idioma en el que se encuentren escritos los documentos.

La ponderación de los rasgos seleccionados de cada documento se suele realizar con diferentes estrategias, la más habitual es la utilización del denominado esquema *tf*-

*idf* en el que el peso de un rasgo se obtiene como producto de dos factores; el primero de ellos, conocido como *factor tf*, mide la frecuencia de aparición del rasgo en el documento, mientras que el *factor idf*, conocido usualmente como frecuencia inversa del documento, permite rebajar significativamente el valor de los pesos correspondientes a rasgos con poco valor discriminante por aparecer en muchos documentos de la colección.

La representación vectorial de documentos facilita además la comparación de éstos utilizando métricas sobre el espacio vectorial en el que se han representado. Aunque varias métricas son admisibles (Edgge, 2002), una de las más utilizadas en minería de texto es la conocida como similitud coseno o separación angular que calcula la similitud entre dos documentos a partir del coseno del ángulo formado por sus vectores.

### 2.1. Soluciones lingüísticas para la representación de documentos

Con objeto de facilitar las labores de extracción de conocimiento en repositorios documentales se han venido confeccionando diferentes elementos de naturaleza lingüística que tratan de representar el conocimiento compartido y común sobre áreas temáticas específicas. Entre ellos destacan los glosarios especializados, taxonomías, tesauros y ontologías.

Se podría definir un glosario como un repertorio de términos pertenecientes a un área de conocimiento o disciplina, añadiendo por lo general definiciones o explicaciones necesarias para su comprensión. La utilización de glosarios específicos posibilita la representación de los documentos sobre un espacio de dimensión más reducida que la que se necesitaría si se utiliza el conjunto completo de palabras presentes en la colección. Es de destacar además la existencia en todas las disciplinas científicas de glosarios más o menos completos de carácter multilingüe, en los que los términos se presentan en diferentes idiomas. Como ejemplos vinculados al campo de la economía se podrían citar el glosario multilingüe elaborado por el Fondo Monetario Internacional (FMI) (http://www.imf.org), con 11.624 registros de palabras, frases, y títulos institucionales en diferentes idiomas (español, inglés, alemán, francés y portugués) comúnmente encontrados en documentos del FMI. Se trata de un glosario muy completo para su utilización en procesos de indización de documentación técnica

en el campo de la economía. Otra referencia interesante es el *Glosario de Inter-Active Terminology for Europe* (http://iate.europa.eu/iatediff), glosario que cuenta con 1.400.000 elementos de la terminología específica de la Unión Europea, cubriendo 24 idiomas y una gama de áreas temáticas bastante amplia.

La extracción de conocimiento a partir de colecciones documentales puede facilitarse mediante la utilización de otro tipo de vocabularios controlados con una estructura más compleja y que no se limiten a una recopilación más o menos amplia de términos o expresiones. En este contexto se sitúan las taxonomías, tesauros y ontologías, que por ese orden, amplían cada vez más las prestaciones de los glosarios.

Las taxonomías constituyen una interesante alternativa para la gestión del conocimiento, y muy en particular para su utilización en procesos de clasificación. De hecho, la taxonomía, entendida como ciencia, se ocupa de los principios, métodos y fines de la clasificación. Desde el punto de vista de la lingüística computacional, se puede ver una taxonomía como una lista estructurada en árbol, organizada jerárquicamente desde los términos más generales hasta los términos más específicos. Las aristas del árbol, por tanto, definen las conexiones entre los términos. De una manera un tanto simple se podría entender también una taxonomía como cualquier conjunto de términos que comparten algún principio de organización. Toda taxonomía es diseñada para facilitar la recuperación de información de una manera flexible y se caracteriza por una estructura mucho más simple que la presente en otras alternativas como son los tesauros o las ontologías.

Un ejemplo de taxonomía utilizada para la clasificación de documentación científica en el campo de la economía o las ciencias empresariales es la taxonomía del sistema JEL, diseñada por la *Journal of Economic Literature* (http://www.aea-web.org). Esta taxonomía surgió para facilitar la clasificación de los artículos y trabajos científicos publicados en dicha revista, pero con el tiempo se ha convertido en un estándar de clasificación en el campo de la economía. La taxonomía JEL está estructurada jerárquicamente en 3 niveles con 20 categorías principales, que se subdividen a su vez en subcategorías y subsubcategorías.

Con un mayor nivel de complejidad surgen los tesauros, entendidos de alguna manera como "taxonomías con extras". Los tesauros facilitan un almacenamiento adecuado de la información, así como implementan un nexo de comunicación, identidad conceptual o interfase entre el lenguaje natural y el lenguaje en que se hayan escritos los documentos contenidos en un sistema de gestión documental. Básicamente los tesauros

son listas estructuradas que pretenden representar de forma unívoca el contenido conceptual de los documentos asociados a un área temática determinada y que pueden ser fácilmente integrados en los sistemas de gestión documental. Los tesauros suelen ser polijerárquicos, con diferentes tipos de relaciones y pueden contener notas de alcance para indicar el significado de algunos términos. Los conceptos asociados a los documentos pueden ser expresados a través de los descriptores del tesauro, facilitando las operaciones posteriores de búsqueda y clasificación. Además de la presencia de relaciones más complejas entre los términos, otra de las diferencias de los tesauros con respecto a las taxonomías es que los tesauros suelen estar más orientados a facilitar las operaciones de búsqueda o recuperación, mientras que las taxonomías son diseñadas con el objeto de agilizar la clasificación.

Existen tesauros que cubren ampliamente diversas áreas, desde disciplinas científicas como la astronomía, física, informática o medicina; hasta áreas vinculadas con las ciencias sociales. Gran parte de los tesauros presentes en Internet han sido construidos por entidades públicas o privadas cuyo trabajo está vinculado directa o indirectamente con el tratamiento de información y los lenguajes documentales. Mención destacada merece el tesauro multilingüe *Eurovoc* (http://eurovoc.europa.eu), creado por la Comisión Europea, que cubre la totalidad de idiomas oficiales de la Unión Europea. Este tesauro se viene utilizando con buenos resultados en proyectos de investigación sobre recuperación de información, clustering y clasificación documental, Steinberger (2005). Otro ejemplo es el tesauro de la *Organización Internacional del Trabajo* (http://www.ilo.org/thesaurus), elaborado como mecanismo de indexación del catálogo automatizado de su biblioteca central y otros servicios de información. Recopila más de 4.000 términos relativos al entorno laboral y al desarrollo económico y social, todos ellos expresados en tres lenguas: inglés, francés y español.

Finalmente, las ontologías, al igual que los tesauros, tratan de organizar de manera sistemática el conocimiento a partir de un conjunto de términos, conceptos y relaciones entre ellos. Las ontologías definen relaciones complejas, e incorporan reglas y axiomas que no tienen los tesauros, obteniendo una representación formal de los conceptos y las relaciones existentes entre ellos. Una de las definiciones más acertadas, y adoptada ya como un estándar, del concepto de ontología es la aportada por (Gruber, 1995): “especificación explícita y formal de una conceptualización compartida”.

Existen ontologías específicas y ontologías de carácter general (proporcionan terminologías útiles para varios campos). Por ejemplo, WordNet (http://wordnet.princeton.edu) es una ontología lingüística de carácter general que organiza los nombres, verbos y adjetivos del idioma inglés en grupos de sinónimos.

El uso de glosarios, tesauros y ontologías facilita la representación vectorial de los documentos, extrayendo como rasgos, por ejemplo, los términos identificados de un glosario o los descriptores de un tesauro.

## 3. SWARM INTELLIGENCE

La observación y el análisis del comportamiento de grupos de seres vivos o de determinados fenómenos naturales han inspirado diferentes modelos matemáticos para la resolución de una gran variedad de problemas prácticos en todo tipo de disciplinas. El término *Swarm Intelligence* (Bonabeau, 1999), que puede ser traducido como inteligencia de enjambre o inteligencia colectiva, hace referencia a diversas técnicas utilizadas en el ámbito de la *Inteligencia Artificial* y que se basan en la idea de que grupos de agentes extremadamente sencillos y poco o nada organizados pueden exhibir un comportamiento complejo, incluso inteligente, utilizando reglas y mecanismos de comunicación local simples. Gracias a la cooperación, en situaciones donde no se tiene un conocimiento global del entorno, los grupos o *swarms* pueden alcanzar objetivos globales intercambiando información disponible localmente, y llegar a resolver en grupo problemas difícilmente resolubles de manera individual.

Diversos estudios sobre animales e insectos sociales han inspirado numerosos modelos computacionales de *Swarm Intelligence* (Garnier, 2007). Las hormigas, constituyen el ejemplo más clásico de este tipo de comportamiento, y el modo en el que utilizan compuestos químicos (feromonas) para transmitir información ha sido fuente de inspiración de diversas líneas de investigación. Entre las técnicas más conocidas, e inspiradas en los comportamientos observados en las colonias de hormigas, se encuentran las técnicas de optimización de colonias de hormigas, ACO (Ant Colony Optimization) (Dorigo, 1992). Se trata de algoritmos metaheurísticos de naturaleza estocástica aplicados a problemas de búsqueda de soluciones óptimas sobre un grafo y que representan una interesante alternativa para problemas de naturaleza combinatoria difíciles de resolver con técnicas clásicas.

Existen otras técnicas también enmarcadas dentro de *Swarm Intelligence* e inspiradas igualmente en comportamientos de colectividades biológicas, ejemplos de colectividades que han inspirado modelos computacionales son las termitas, determinadas especies de abejas, arañas, bancos de peces, bandadas de pájaros, etc. En todos los casos, los comportamientos complejos observados en las colectividades son el resultado de la interacción a lo largo del tiempo entre los individuos que las componen. El comportamiento del grupo viene influenciado por los comportamientos individuales de sus integrantes, pero, a su vez, el comportamiento del grupo influye también en las acciones individuales de los integrantes. Este tipo de comportamientos emergentes también puede observarse en sistemas no biológicos, como por ejemplo los mercados bursátiles, el tráfico en las grandes ciudades, o las estructuras espaciales de las galaxias.

La mayoría de las técnicas computacionales de *Swarm Intelligence* se orientan hacia la resolución de problemas de optimización. Junto con las técnicas ACO, anteriormente citadas, se pueden destacar las técnicas de optimización conocidas como PSO (*Particle Swarm Optimization*) iniciadas por (Kennedy y Eberhart, 1995), (Eberhart y Kennedy, 1995) y que se inspiran en el comportamiento de las bandadas de aves o bancos de peces. Las técnicas PSO tratan de modelar dos tipos de comportamientos: cada individuo trata de moverse hacia la posición de su mejor vecino y al mismo tiempo trata de dirigirse hacia su mejor posición previa. Especialmente en los últimos años, las técnicas PSO han comenzado a utilizarse para resolver con éxito un gran número de problemas en campos muy diversos (Poli, Kennedy y Blackwell, 2007). Un algoritmo PSO trabaja en todo momento con una población de "partículas" o vectores de posición en un espacio de búsqueda, cada uno representando una posible solución del problema. Además de la posición cada partícula lleva asociado un vector de velocidad que determina el desplazamiento de la posición para la siguiente iteración. Los vectores de velocidad, y por tanto también las posiciones, se van actualizando de acuerdo a la experiencia propia y a la de sus partículas vecinas. Una vez modificadas sus posiciones, las partículas evalúan la solución alcanzada y actualizan, en su caso, su memoria (mejores posiciones locales y globales). El proceso iterativo continúa hasta que se cumpla cierto criterio de parada.

Existen también dentro de este campo de la Inteligencia Artificial algoritmos específicos que se aplican a problemas de clustering y que se inspiran igualmente en comportamientos observados en colonias de hormigas. En este caso se reproducen los

mecanismos utilizados por las hormigas para ordenar sus nidos, mecanismos que aplican a tareas variadas, como el agrupamiento de las larvas, organización de cadáveres en cementerios, o la colocación de los alimentos en la colonia. En (Garnier, 2007) puede encontrarse una descripción de estos procesos observados en diferentes especies de hormigas. Básicamente, la tarea realizada por las hormigas en todos esos casos consiste en una sucesión de procesos de recogida y colocación de objetos en función de la densidad de objetos detectada en un entorno local. El trabajo pionero en el desarrollo de este tipo de técnicas matemáticas es (Deneubourg, 1990). Este tipo de algoritmos de clustering se denominan de manera general algoritmos de *ant clustering*.

### 3.1. Swarm Intelligence en procesos de minería de texto

Las técnicas computacionales de Swarm Intelligence tienen una amplísima aplicabilidad, y pueden ser aplicadas a procesos de extracción de conocimiento en minería de texto. De hecho en este trabajo se presenta un modelo integrador en el que se aplican estrategias ACO, PSO y ant clustering para labores de clasificación documental. Este modelo ha sido implementado en una plataforma informática y será descrito en la próxima sección.

Un proceso de clasificación documental puede verse como un problema de optimización en el que se desea realizar una división del repositorio documental en grupos en los que la similitud de los documentos dentro de cada grupo sea máxima, o dicho de otra manera, la distancia media entre ellos con la métrica considerada sea mínima. Bajo esta consideración, las técnicas ACO y PSO pueden adaptarse para ser aplicadas en estos procesos. Por otro lado, las técnicas de ant clustering son algoritmos específicos de clustering que pueden ser aplicadas directamente sobre las representaciones vectoriales de los documentos.

En los últimos años puede encontrarse en la bibliografía especializada en minería de texto diferentes ejemplos de aplicación de estas técnicas de Swarm Intelligence. Muchas de las aplicaciones en clustering documental utilizan diferentes variantes de los algoritmos de ant clustering, (Handl y Meyer, 2002), (Azzag et al., 2004), (Vizine et al., 2005) son algunos ejemplos. También se encuentran modelos de clustering k-PSO en los que las posiciones de las partículas del enjambre determinan los centros de atracción de los grupos a crear (Wang, et al. 2007), (Cobo y Rocha, 2008),

(Rocha, et al., 2008), o como ejemplo de aplicación de técnicas ACO a problemas de agrupamiento de documentos puede verse (Cobo y Rocha, 2007).

## 4. IMPLEMENTACIÓN DE UN MODELO INTEGRADOR DE GESTIÓN DE INFORMACIÓN NO ESTRUCTURADA

Las secciones precedentes han tratado de poner de manifiesto la importancia de la correcta gestión de la información no estructurada para las organizaciones, y la existencia de metodologías que pueden ayudar a automatizar esos procesos. Con objeto de poder acercar a las organizaciones estas metodologías se ha diseñado e implementado en una solución tecnológica un completo modelo de gestión documental desde una perspectiva multidimensional y que se integra en una herramienta de gestión documental. El modelo propuesto tiene por objeto generar información y conocimiento a partir de información no estructurada obtenida de documentos textuales; para ello es preciso utilizar técnicas que estructuren la información para que ésta pueda ser almacenada en sistemas gestores de bases de datos. En esta fase las técnicas de minería de texto desempeñan un papel esencial. En este proceso de estructuración se utilizan los glosarios y tesauros multilingües como elementos de identificación de rasgos y de representación de los documentos con independencia del idioma en el que se encuentren escritos. Gracias a la estructuración de la información el modelo propone la utilización de técnicas de Swarm Intelligence para la extracción de conocimiento. Finalmente el conocimiento o la nueva información generada deben ser adecuadamente presentada a través de una interfaz intuitiva; teniendo en cuenta que el modelo se orienta hacia un entorno de trabajo multiusuario, se requieren protocolos de comunicación entre los servidores en los que se alojan los sistemas de gestión documental y los equipos de los clientes. En el modelo propuesto el sistema de gestión documental actúa como núcleo, convirtiéndose en el elemento integrador del resto de componentes. Además proporciona al modelo las prestaciones básicas de todo sistema documental. La Figura 1 muestra la estructura del modelo.

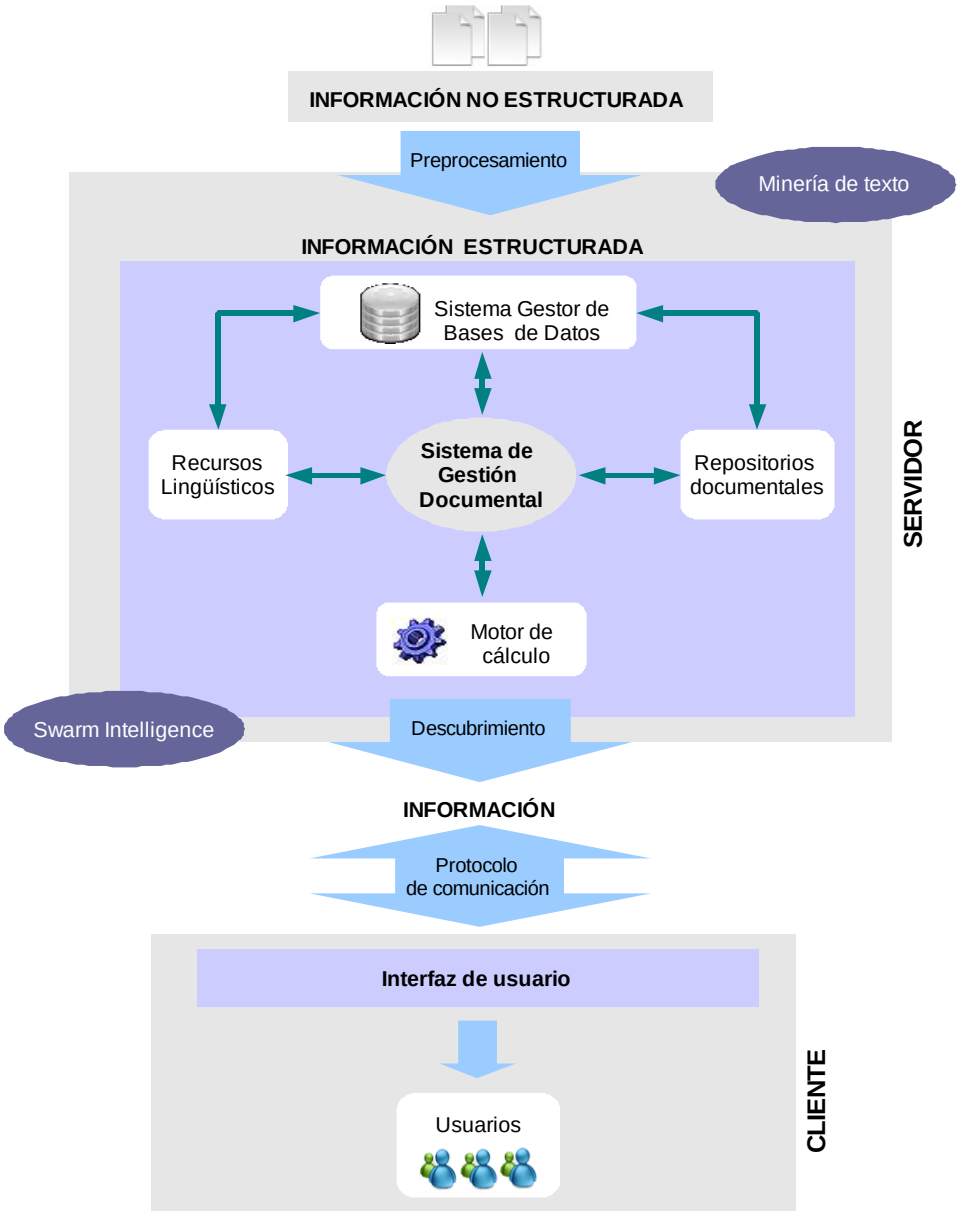


**Figura 1:** Estructura del modelo de gestión documental.

Para la implementación del modelo se ha optado por recurrir a tecnologías open source, tomando como núcleo la aplicación OWL que incorpora las funcionalidades básicas de un sistema de gestión documental. Sobre ese núcleo se han implementado procedimientos automáticos de extracción y ponderación de rasgos multilingües, utilizando como recursos lingüísticos el glosario del FMI y el tesauro Eurovoc. Además de extraer rasgos multilingües, la aplicación realiza análisis morfológicos del texto extraído para identificar sustantivos, adjetivos y verbos, procediendo a su lematización. La Figura 2 muestra parte de la interfaz de la aplicación con las nuevas funcionalidades implementadas, y a las que se accede con un conjunto de botones que se muestran en la

parte superior. En la parte inferior de la página principal de la aplicación, se dispone de un segundo bloque de opciones destinado a la gestión y operación sobre el repositorio documental, permitiendo cargar y descargar documentos, organizarlos por carpetas, realizar seguimiento de versiones, búsqueda de documentos relevantes y todas las funciones propias de cualquier sistema gestor documental.

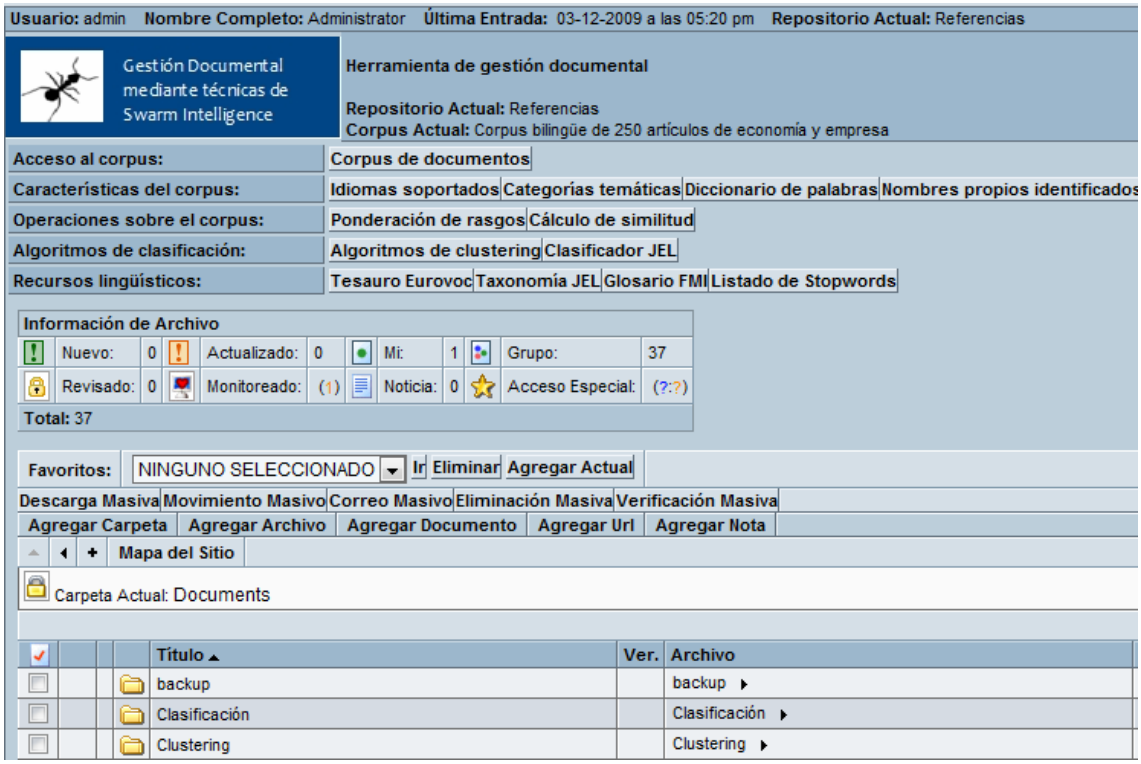


**Figura 2.** Núcleo del sistema a partir de la herramienta OWL, con funcionalidades adicionales para el procesamiento del texto y la aplicación de algoritmos de clasificación.

Toda la información sobre los rasgos extraídos de cada documento y sus respectivos pesos es almacenada en la base de datos de la aplicación. La Figura 3 muestra a modo de ejemplo la información identificada en un documento concreto (nombres propios, palabras lematizadas, términos del glosario y microtesauros de Eurovoc) junto a su frecuencia absoluta de aparición dentro del documento y su peso de acuerdo al esquema de ponderación *tf-idf*.

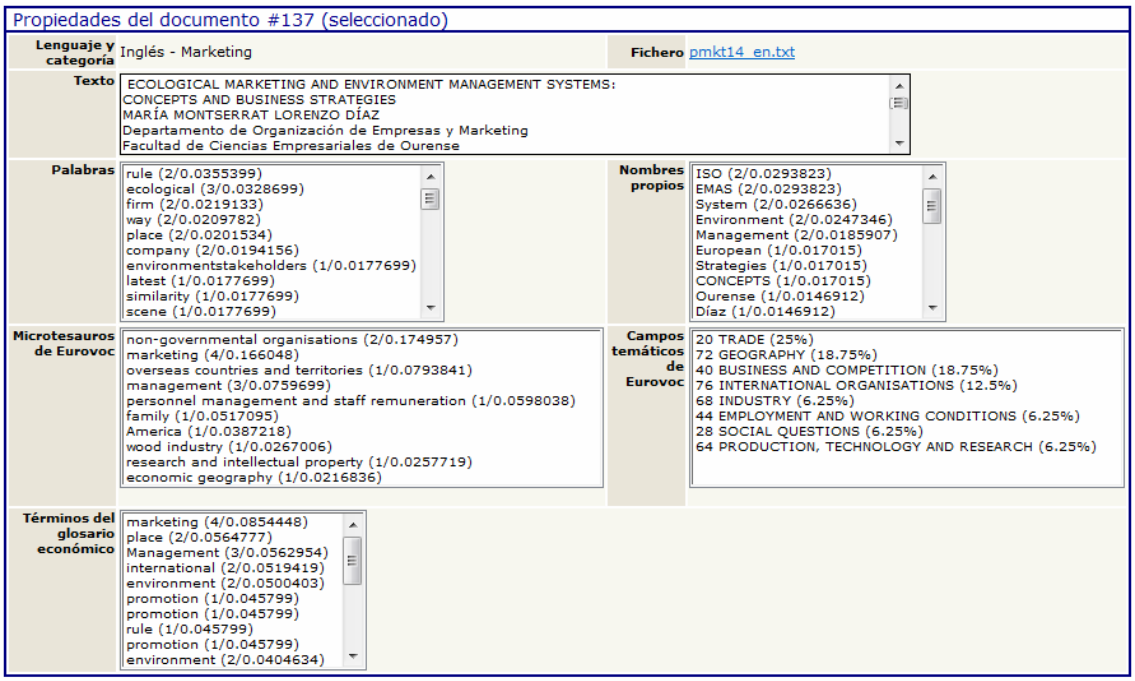


**Figura 3**. Propiedades y descripción detallada de los rasgos de un documento seleccionado en la colección de documentos.

En cuanto a la integración de técnicas de Swarm Intelligence, se han implementado procedimientos de clasificación automática de documentos mediante la taxonomía JEL y algoritmos de clasificación no supervisada (clustering) mediante técnicas basadas en PSO, ACO y ant clustering. La Figura 4 muestra la interfaz del motor de cálculo del sistema, como se observa, a través de un sencillo menú el usuario puede seleccionar la técnica de clasificación deseada. Se puede utilizar un algoritmo de clasificación no supervisada clásico como es el algoritmo de las k-medias, o diferentes algoritmos basados en *Swarm Intelligence*, la descripción de los algoritmos puede encontrarse en (Cobo y Rocha, 2007) y (Cobo y Rocha, 2008). Los algoritmos basados en estrategias de optimización ACO o PSO exigen conocer de antemano el número de grupos a crear. Si ese dato no está disponible, el usuario podrá seleccionar las estrategias de ant clustering, que además de identificar automáticamente el número de grupos proporcionan una visualización gráfica de los mismos.

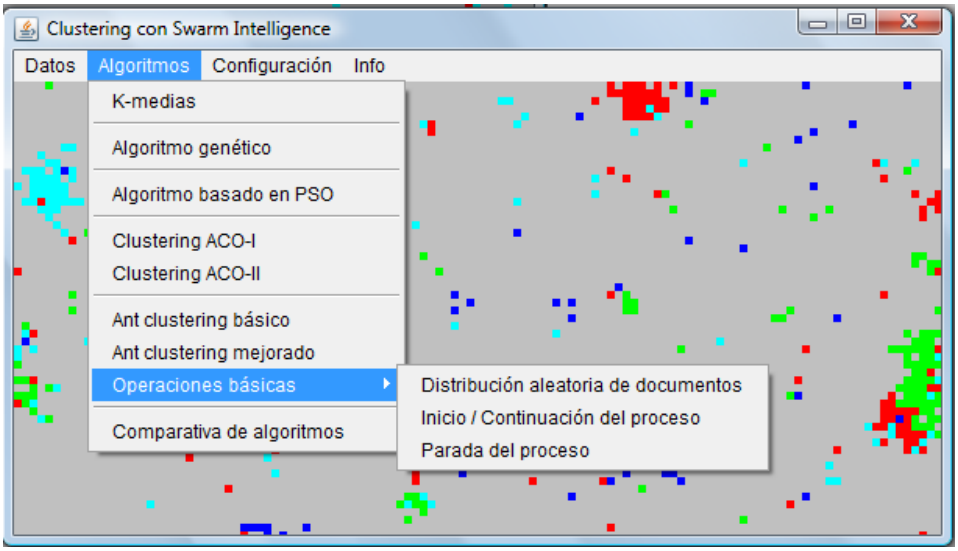


**Figura 4.** Algoritmos de Swarm Intelligence implementados en el sistema.

## 4. EJEMPLO DE APLICACIÓN

A modo de ejemplo se presentarán en esta sección algunos resultados experimentales del uso del modelo y la aplicación en el proceso de clasificación de una colección de 250 documentos científicos de 5 categorías diferentes relacionadas con la economía y la gestión empresarial, extraídos de publicaciones científicas de las áreas correspondientes. Para aumentar aún más la complejidad del proceso de clasificación se seleccionaron 125 documentos escritos en inglés y otros tantos en español. En este caso al utilizarse para la representación de los documentos dos herramientas lingüísticas multilingües (glosario FMI y tesauro Eurovoc) la barrera lingüística puede ser superada.

Un proceso de clasificación documental puede verse como un proceso de optimización en el que se trata de minimizar las distancias medias de las representaciones vectoriales de los diferentes documentos a los centroides de los grupos a los que son asignados. En este sentido, es posible utilizar metaheurísticas para abordar el proceso de optimización. Sobre esa base, en el modelo desarrollado se han implementado técnicas de clasificación basadas en ACO y en PSO. La Tabla 1 muestra los valores medios de diferentes indicadores de calidad clásicos en minería de datos al realizar 20 procesos de clasificación independientes sobre el corpus documental. Como puede apreciarse, las técnicas basadas en ACO producen mejores resultados.

| | PSO | ACO | k-medias |
|---|---|---|---|
| **Entropía** | 0.8737 | 0.9249 | 1.0274 |
| **Medida F** | 0.5921 | 0.6268 | 0.5676 |
| **Pureza** | 0.5816 | 0.6382 | 0.5550 |
| **Tiempo (miliseg.)** | 36986.95 | 7569.25 | 1507.05 |

**Tabla 1.** Comparativa de resultados de aplicación de procesos de clasificación no supervisada basados en técnicas ACO y PSO, en comparación con un algoritmo de clustering clásico.

La Figura 5 muestra la evolución de la clasificación no supervisada en una de las ejecuciones del algoritmo basado en ACO. El proceso ACO es un proceso iterativo en el que en cada una de las iteraciones se construye un conjunto de posibles agrupamientos cuya calidad debe ir mejorando a medida que avanza el proceso. En el gráfico se muestra la evolución de la distancia media a los centroides de los grupos creados en cada iteración. Cuanto menor sea este valor mejor se considerará el agrupamiento realizado, al estar en los agrupamientos los documentos muy próximos entre sí dentro de los grupos correspondientes. Se puede apreciar cómo a medida que avanzan las iteraciones del proceso los agrupamientos creados son mejores, observándose una clara convergencia en la calidad media de los agrupamientos creados por las "hormigas" de la colonia en las diferentes iteraciones.

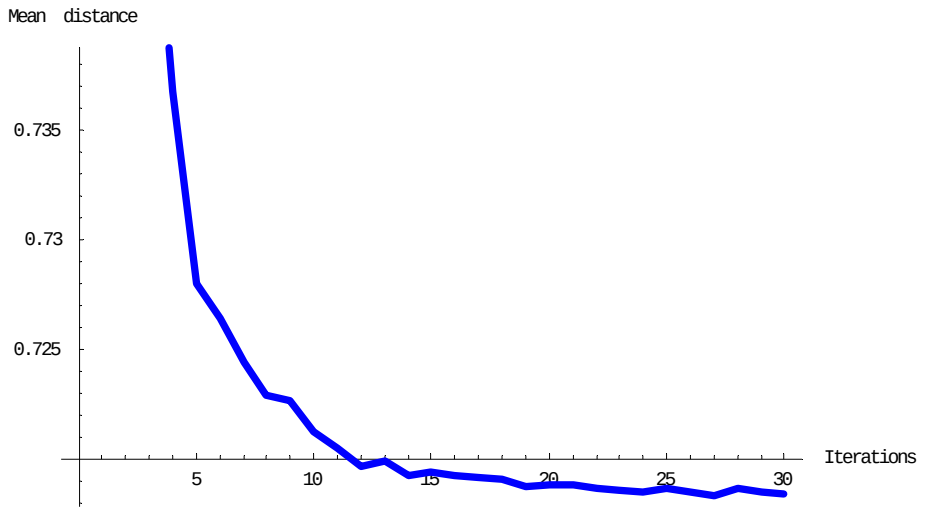


**Figura 5.** Ejemplo de la evolución de un algoritmo ACO para la clasificación no supervisada de un conjunto de 250 documentos.

## 5. CONCLUSIONES

Los modelos de optimización basados en Swarm Intelligence han demostrado su valía en una gran variedad de contextos, en este caso, combinados con las metodologías de la minería de texto favorecen la adecuada gestión de los enormes volúmenes de información no estructurada que se genera en el contexto de las organizaciones. Este tipo de metodologías pueden servir para generar verdadero conocimiento en el contexto de las organizaciones y aprovechar las potencialidades de los sistemas actuales de comunicación y de acceso en línea a recursos de información. La integración de estas metodologías en las herramientas de gestión documental permite su uso por personal sin excesivos conocimientos técnicos.

La principal aportación de este trabajo ha sido la elaboración de una herramienta en la que se han integrado novedosas técnicas de clasificación documental basadas en estrategias de optimización y modelos de clustering inspiradas en comportamientos observados en colonias de insectos. Gracias a esa herramienta han podido realizarse diversas pruebas experimentales que ponen de manifiesto la utilidad de los enfoques de optimización de colonias de hormigas para abordar este tipo de problemas.

## 6. REFERENCIAS BIBLIOGRÁFICAS


- AZZAG, H.; GUINOT, C. y VENTURINI, G. (2004). “How to use ants for hierarchical clustering”. 4th International Workshop on Ant Colony Optimization and Swarm Intelligence. Lecture Notes on Computer Science, 3172, pp 350-357.
- BAEZA, R. y RIBEIRO, B. (1999). “Modern Information Retrieval”. Addison Wesley.
- BONABEAU, E.; DORIGO, M.; y THERAULAZ, G. (1999). “Swarm Intelligence: From Natural to Artificial Systems”. Oxford University Press.
- BRÜCHER, H.; KNOLMAYER, G. y MITTERMAYER, M. A. (2002). “Document Classification Methods for Organizing Explicit Knowledge”. Proceedings of the Third European Conference on Organizational Knowledge, Learning, and Capabilities.

- COBO, A. y ROCHA, R. (2007). "Application of Ant-Based Algorithms for Clustering in Multilingual Document Collections". 3nd European-Latin American Workshop on Engineering Systems - Curicó (Chile).
- COBO, A. y ROCHA, R. (2008). "Clustering de documentos mediante técnicas híbridas de Swarm Intelligence". CLAIO'08 - XIV Congreso Latino-iberoamericano de Investigación de Operaciones.
- DENEUBOURG, J., GOSS, S., FRANKS, N., SENDOVA-FRANKS, A., DETRAIN, C., y CHRETIEN, L. (1990). "The dynamic of collective sorting robot-like ants and ants-like robots". In Proceedings of the First Conference on Simulation of Adaptive Behavior, pages 356-363.
- DORIGO, M. (1992). "Optimization, learning and natural algorithms". PhD Tesis. Politechnico di Milano.
- EBERHART, R. y KENNEDY, J. (1995). "A new optimize using particle swarm theory". Proc. IEEE Sixth International Symposium on Micromachine and Human Science, pp 39-43.
- EGGHE, L. y MICHEL, C. (2002). "Strong similarity measures for ordered sets of documents in information retrieval". Information Processing and Management. N. 38. pp. 823-848.
- GANTZ, J.; CHUTE, C.; MANFREDIZ, A.; MINTON, S.; REINSEL, D.; SCHLICHTING, W y TONCHEVA, A. (2008). "The Diverse and Exploding Digital Universe. An Updated Forecast of Worldwide Information Growth Through 2011". Technical Report, Consultora IDC-EMC.
- GARNIER, S.; GAUTRAIS J.; y THERAULAZ, G. (2007). "The biological principles of Swarm Intelligence". Swarm Intelligence, vol 1, pp 3-31.
- GRUBER, T.R. (1995). "Towards Principles of the Design of Ontologies Used for Knolewdge Sharing". International Journal of Human Computer Studies", num 43, pp 907-928.
- HANDL, J. y MEYER, B. (2002). "Inproved ant-based clustering clustering sorting in document retrieval interface". 7th International Conference on Parallel Problem Solving from Nature. Lecture Notes on Computer Science, 2439, pp 913-923.

- KENNEDY, J. y EBERHART, R. (1995). "Particle Swarm Optimization". Proc. IEEE International Conference on Neural Networks, Perth (Australia), vol 4, pp 1942-1948.
- POLI, R.; KENNEDY, J.; y BLACKWELL, T. (2007). "Particle Swarm Optimization: An overview". Swarm Intelligence, vol 1, pp 33-57.
- ROCHA, R., ALONSO, M., y COBO, A. (2008). "Using Swarm Intelligence Techniques in Document Management Systems". In Proceedings of ICAI'08 – International Conference on Artificial Intelligence - Las Vegas (USA).
- SALTON, G. (1971). "The SMART Retrieval Sistem - Experiments in Automatic Document Processing. Prentice Hall.
- STEINBERGER, R.; POULIQUEN, B. y IGNAT, C. (2005). "Navigating multilingual news collections using automatically extracted information". Journal of Computing and Information Technology, Vol. 13, No. 4, pp. 257-264.
- VIZINE, A.L., de CASTRO, L.N.; HRUSCHKA, E.R. y GUDWIN, R.R. (2005). "Towards Improving Clustering Ants: An Adaptive Ant Clustering Algorithm". Informatica, vol 29, pp 143-154.
- WANG, Z., ZHANG, Q. y ZHANG, D. (2007)."A PSO-based Web document classification algorithm". 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing.