Desarrollo de una aplicación para la gestión, clasificación y agrupamiento de documentos económicos con algoritmos bio-inspirados

Cobo Ortega, Ángel; Rocha Blanco, Rocío
{acobo;rochar}@unican.es
Departamento de Matemática Aplicada y Ciencias de la Computación
Departamento de Administración de Empresas
Universidad de Cantabria

RESUMEN

Este trabajo describe el desarrollo de una aplicación Web que utiliza técnicas bio-inspiradas para clasificar y agrupar colecciones multilingües de documentos en el campo de la economía y los negocios. La aplicación identifica grupos relacionados de documentos económicos, escritos en español e ingles, utilizando algoritmos de clustering inspirados en el comportamiento de las colonias de hormigas. Para la generación de una representación vectorial de los documentos que resulte independiente del idioma, se utilizan varios recursos lingüísticos y herramientas de procesamiento de documentos textuales. Cada documento es representado utilizando cuatro vectores de rasgos independientes del idioma, y la similitud entre ellos es calculada mediante combinaciones lineales convexas de las similitudes de esos vectores de rasgos. El trabajo presenta resultados experimentales obtenidos en la clasificación de un corpus de 250 documentos científicos en diversas áreas de la economía y administración de empresas.

***Palabras claves*:**
Clustering; minería de texto; metaheurísticas; algoritmos bio-inspirados.
***Clasificación JEL (Journal Economic Literature)*:**
C63; C80; C88.

***Área temática*:** Informática aplicada.

1. INTRODUCCIÓN

Las tecnologías de la información y Comunicaciones (TIC) están transformando la manera en la que las organizaciones y las personas realizan sus negocios y actividades. Los sistemas de información de cualquier organización capturan y almacenan grandes cantidades de información y datos sobre la propia organización y su entorno, y los gerentes utilizan dicha información en los procesos de toma de decisiones, planificación y control los procesos. Es por ello que hoy en día la información se ha convertido en un recurso estratégico de primer orden para las organizaciones, y un adecuado uso de ella puede suponer para la organización ventajas competitivas.

La actividad económica en nuestros días viene caracterizada por la internacionalización y la globalización de los mercados. En este contexto, las organizaciones obtienen grandes volúmenes de información de diversas fuentes y de manera automática (suscripción a revistas o periódicos, sindicación de contenidos, recuperación de información de bases de datos, consultas a Internet, etc.). Frecuentemente esta información se refleja en documentos de texto escritos en diferentes idiomas. En tales casos, los sistemas de información y las herramientas de computación pueden ayudar a romper barreras lingüísticas y permitir a las organizaciones administrar, consultar y extraer información de grandes sistemas de documentos, descubriendo el conocimiento global en un entorno multilingüe. La disciplina de minería de texto es especialmente apropiada para este propósito.

La minería de texto hace referencia generalmente a procesos para la obtención de información de alta calidad a través del análisis de textos, descubriendo tendencias, desviaciones y asociaciones entre la información textual, utilizando patrones estadísticos de aprendizaje, y teniendo como objetivo revelar conceptos y relaciones dentro de la colección de textos. Típicamente las tres grandes tareas o áreas de la minería de texto son la recuperación de información, la categorización o clasificación de documentos, y el agrupamiento de documentos en *clusters*. Un problema de clasificación surge cuando se quiere decidir si un documento pertenece a una categoría preestablecida de documentos; en el clustering de documentos, sin embargo, las categorías no están previamente definidas y el objetivo es encontrar grupos de documentos relacionados, con una alta similitud entre ellos y una alta disimilitud con

los documentos de los restantes grupos. Las técnicas de clustering en minería de texto han ido ganando popularidad con el aumento de la disponibilidad de documentos electrónicos escritos en diferentes idiomas. Sin embargo, actualmente la mayoría de las técnicas de clustering se centran principalmente en procesos de clustering de documentos monolingües, una menor atención se ha puesto en diseñar y aplicar técnicas que manejen grupos de documentos escritos en diferentes idiomas. En este trabajo se intenta trabajar en esta línea, aplicando técnicas bio-inspiradas para el agrupamiento documentos relacionados escritos en diferentes idiomas. Además, se ha optado por el diseño de una herramienta especialmente orientada hacia el análisis de documentos de carácter económico o empresarial. El prototipo desarrollado se limita al análisis de documentos escritos en español e inglés, sin embargo las herramientas lingüísticas que utiliza permitirían ampliar el abanico de idiomas incluyendo otros idiomas ampliamente utilizados como el francés o el alemán.

2. PROCESAMIENTO Y EXTRACCIÓN DE RASGOS DE DOCUMENTOS

2.1. Almacenamiento de documentos

La aplicación desarrollada se ejecuta bajo un entorno Web, lo que facilita su integración en los sistemas de información de cualquier empresa y hace que el usuario se encuentre familiarizado con la interfaz. Ha sido desarrollada utilizando tecnologías de programación del lado del servidor (lenguaje PHP) y del lado del cliente (lenguaje Java); además trabaja sobre una base de datos con el gestor MySQL. El usuario puede consultar en todo momento los documentos almacenados, accediendo no solo al texto completo sino también a los rasgos identificados en el mismos, así por ejemplo, puede consultar las palabras más representativas o el área más afín de acuerdo a un tesauro especializado. El usuario puede igualmente clasificar manualmente el documento dentro de un conjunto de categorías predefinidas.

2.2. Recursos lingüísticos

Para salvar las barreras lingüísticas y poder comparar documentos similares aunque estén escritos en diferentes idiomas, se hace necesario disponer de herramientas que permitan obtener una representación de cada documento independiente del idioma.

En este caso se ha optado por la utilización de un glosario multilingüe de términos económicos y un tesauro que organiza jerárquicamente términos y conceptos relacionados con un área determinada. En este caso, se ha integrado en la aplicación web el glosario multilingüe del Fondo Monetario Internacional (FMI[1]) y el tesauro *Eurovoc*[2] desarrollado por la Comisión Europea. Los respectivos organismos han otorgado licencias de uso para fines de investigación en este trabajo y ambos recursos han sido integrados en la base de datos sobre la que trabaja la aplicación. El glosario del FMI contiene más de 11.500 registros de términos: palabras, frases, y títulos institucionales comúnmente encontrados en documentos del Fondo Monetario Internacional en las áreas de moneda y banca, finanzas públicas, balance de pagos y crecimiento económico. Estás versiones de términos están disponibles en varios idiomas, aunque en el caso del presente proyecto se ha limitado a los dos idiomas anteriormente citados. El glosario del FMI ha sido ampliado con términos de un diccionario electrónico de términos español-inglés en el ámbito de los negocios. De esta menara, la aplicación gestiona un total de 18.724 términos. El tesauro multilingüe *Eurovoc* cubre los campos en los cuales la Comunidad Europea es activa, proporcionando los medios de indexación de documentos en los sistemas de documentación de las instituciones europeas y de sus usuarios. Eurovoc 4.2 existe en 21 idiomas oficiales de la Unión Europea y es utilizado en proyectos de investigación sobre recuperación de información, clustering de documentos y clasificación [Steinberger et al., 2005]. El tesauro tiene una lista estructurada con más de 6.600 descriptores y 127 microtesauros en 21 campos temáticos.

2.3. Procesamiento de los documentos

En el momento de insertar un nuevo documento en la base de datos, una serie de operaciones de preprocesamiento son iniciadas de manera automática:

- *Extracción del texto de cada documento*. El documento original puede estar en formato PDF o TXT y la aplicación Web utiliza la herramienta *pdftotext* para extraer el texto de cada archivo; este texto es insertado en una campo de la base de datos.

[1] http://www.imf.org/external/np/term/index.asp

[2] http://europa.eu/eurovoc/

- *Eliminación de Stopwords*. Una lista de *stopwords* es utilizada para eliminar términos comunes que no aportan ninguna información sobre el contenido o la temática del documento. Ejemplos de tales palabras son los artículos, conjunciones, preposiciones, etc. La aplicación cuenta con una lista predefinida de 463 términos considerados como *stopwords* (359 en español y 104 en ingles).
- *Lematización, clasificación de palabra e identificación de nombres propios*. La herramienta *TreeTagger* ha sido integrada en la aplicación para identificar los lemas de las palabras y sus categorías gramaticales (verbos, adjetivos, nombres, etc.). Esta herramienta clasifica como nombres propios a las palabras que empiezan con una letra mayúscula, si bien este criterio de clasificación es claramente mejorable, los resultados obtenidos son aceptables a efectos de realizar procesos de clasificación y agrupamiento.
- *Búsqueda de términos en el glosario económico bilingüe*. Se localizan aquellos términos del glosario que aparecen en el contenido del documento.
- *Aplicación del tesauro Eurovoc*[3]. La aplicación busca descriptores del tesauro Eurovoc y los microtesauros asociados utilizando las propias reglas de relación del tesauro. Además esta identificación permite calcular el grado de asociación del documento con cada una de las 21 áreas temáticas que cubre el tesauro.

Tras la realización de las etapas anteriores, el documento está listo para ser representado vectorialmente a partir de los rasgos identificados.

3. REPRESENTACIÓN DE DOCUMENTOS Y CÁLCULO DE SIMILITUDES

3.1. Modelo Vectorial

La extracción de rasgos permite obtener una representación independiente del idioma de los documentos del corpus. Tradicionalmente, cada documento es representado por un vector de términos ponderados (rasgos) [Salton, 1971], [Baeza and Ribeiro, 1999]; en este caso se ha empleado un modelo vectorial modificado para

[3] http://europa.eu/eurovoc/

representar cada documento por 4 vectores de rasgos diferentes. El modelo vectorial se ha convertido en una herramienta estándar en sistemas de recuperación de información, se basa en una idea simple: dado un grupo de términos de un documento, no todos ellos son igualmente importantes para describir los contenidos del, esto conduce a la asignación de pesos numéricos $w_{ij} \geq 0$ a cada término o palabra k_i de un documento d_j. De esta forma, el documento puede ser representado por el vector $\mathbf{d}_j = (w_{1j}, w_{2j}, ..., w_{tj})$. Este vector puede ser normalizado para facilitar el cálculo posterior de la similitud con otros documentos. Si N es el número de documentos de la colección y t el número de términos o palabras presentes; el corpus de documentos puede ser representando por una matriz $W=(w_{ij})$ de dimensión $t \times N$, donde cada columna representa un documento, y cada entrada representa el peso de una palabra en un documento. Esta matriz es conocida como matriz de términos-documentos.

En la fase de preprocesamiento de los documentos la aplicación identifica cuatro grupos de rasgos: los términos asociados al glosario económico, los descriptores del tesauro y a partir de ellos los microtesauros, los nombres propios (personas, lugares y organizaciones) y las palabras en el lenguaje nativo. Utilizando esos rasgos, un documento es representado por cuatro vectores. Siempre que un nuevo documento es insertado en la base de datos, dichos vectores son construidos automáticamente. Al utilizar los recursos bilingües señalados anteriormente, los 2 primeros vectores de representación son independientes del idioma; además el vector de nombres propios suele resultar también independiente del idioma.

Para el cálculo de los pesos o coordenadas de los vectores se ha utilizado un *esquema tf-idf*. El peso de un término en un documento se obtiene como producto de dos factores; el primero de ellos, conocido como *factor tf*, mide la frecuencia de aparición del termino 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 términos con poco valor discriminante por aparecer en muchos documentos de la colección. El *esquema TF-IDF (Term Frequency Inverse Document Frequency Weighting)* es definido por:

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

donde $freq_{i,j}$ representa el número de veces que la palabra o término k_i aparece en el texto del documento d_j, N es el número total de documentos en la colección y n_i es el número de documentos que contienen el término k_i. Hay muchas variaciones de la formula TF-IDF, pero todas ellas están basadas en la misma idea: tener en cuenta la frecuencia de aparición de cada término en el documento pero también cómo de frecuente es ese término en los documentos de la colección.

3.2. Cálculo de similitudes

Con los cuatro vectores asociados a cada documento, se puede estimar la similitud entre un par de documentos *(p,q).* La función de similitud utilizada es una combinación lineal convexa de las similitudes entre los cuatro vectores de rasgos:

$$\begin{aligned} SimML(\mathbf{p},\mathbf{q}) = {} & \lambda_1\, Sim(V_{glossary}(\mathbf{p}), V_{glossary}(\mathbf{q})) + \lambda_2\, Sim(V_{Eurovoc}(\mathbf{p}), V_{Eurovoc}(\mathbf{q})) + \\ & \lambda_3\, Sim(V_{pnames}(\mathbf{p}), V_{pnames}(\mathbf{q})) + \lambda_4\, Sim(V_{words}(\mathbf{p}), V_{words}(\mathbf{q})) \\ \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1 \quad & \text{con} \quad \lambda_i \geq 0 \end{aligned} \tag{2}$$

La similitud entre dos vectores es calculada con la clásica *medida del coseno*, también conocida como separación angular, y cuya definición es:

$$Sim(v(\mathbf{p}), v(\mathbf{q})) = \cos(\sigma) = \frac{\mathbf{p} \circ \mathbf{q}}{\|\mathbf{p}\| \times \|\mathbf{q}\|} = \frac{\sum_{i=1}^{t} w_{ip} w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ip}^2} \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \tag{3}$$

Teniendo en cuenta que los pesos son todos no negativos, *Sim(v(**p**),v(**q**))* varía entre 0 y 1. Cuando la similitud es 1, los vectores se consideran totalmente similares. Si los vectores han sido previamente normalizados, el cálculo de esta expresión se reduce al cálculo del producto escalar. La métrica (3) es relativamente simple, y resultados experimentales han demostrado que es una de las que resulta más efectiva para los propósitos de medir la similitud entre documentos.

4. ALGORITMOS DE CLUSTERING BASADOS EN COLONIAS DE HORMIGAS

Los algoritmos de clustering basados en colonias de hormigas, introducidos inicialmente por [Deneubourg, 1990] y [Lumer and Faieta, 1994], se encuentran entre

las primeras técnicas metaheurísticas inspiradas en el comportamiento de las hormigas. Una colonia de hormigas tiene muchas características que se pueden considerar útiles, está compuesta por muchos agentes simples que pueden realizar tareas complejas en grupo, pero sin una coordinación centralizada. Las hormigas de la vida real realizan grupos y clasifican objetos entre sus muchas actividades cotidianas.

La aplicación desarrollada utiliza una técnica de clustering inspirada en colonias de hormigas para agrupar los documentos relacionados. En concreto, se utiliza un algoritmo denominado *ant clustering* [Monmarché, 1999], [Handl, 2006], adaptándole a las particularidades y el contexto del tipo de documentos analizados. El proceso de agrupamiento se realiza sobre una rejilla bidimensional toroidal, donde los objetos (documentos) son colocados aleatoriamente. Tras esa disposición inicial, un grupo de hormigas artificiales exploran la rejilla realizando operaciones de recolocación de objetos. La probabilidad de recoger un documento o colocar uno recogido previamente en una celda de la rejilla dependerá de la similitud entre ese documento y los documentos de un entorno de la celda. Cuando un documento es similar con sus vecinos en la rejilla, la probabilidad de recogerlo es baja, sin embargo, si la similitud es baja las hormigas van a recogerlo con una alta probabilidad y buscarán una buena posición en la rejilla para recolocarlo. Esas probabilidades se definen por las expresiones:

$$P_{pick}(\mathbf{d}_i) = \left(\frac{k^+}{k^+ + f(\mathbf{d}_i)} \right)^2 \qquad P_{drop}(\mathbf{d}_i) = \left(\frac{f(\mathbf{d}_i)}{k^- + f(\mathbf{d}_i)} \right)^2 \tag{4}$$

donde k^+ es un parámetro de recogida, k^- es un parámetro de colocación y $f(d_i)$ es una función de similitud con el vecindario definida por:

$$f(\mathbf{d}_i) = \frac{1}{\sigma^2} \sum_{\mathbf{d}_j \in \Omega} \frac{SimML(\mathbf{d}_i, \mathbf{d}_j)}{\alpha} \tag{5}$$

Tras una operación de recogida o colocación de un documento, la hormiga se moverá en una posición próxima en la rejilla en una dirección elegida aleatoriamente, y el proceso continuará. Siguiendo estas reglas, los documentos relacionados tenderán a situarse en posiciones vecinas de la rejilla, obteniéndose además una visualización gráfica de los grupos obtenidos (ver Figura 1).

Como modificaciones al algoritmo básico, en la aplicación se han considerado algunas mejoras propuestas por [Handl, 2006], como el uso de una memoria a corto

plazo que permita a las hormigas recordar las posiciones de los últimos documentos colocados y la actualización dinámica de determinados parámetros del algoritmo. Las técnicas de clustering basadas en hormigas han sido utilizadas en una amplia variedad de aplicaciones, y diferentes estudios muestran evidencias de que los mecanismos de clustering basados en hormigas son una alternativa robusta y viable, comparada con otras técnicas [Handl et al., 2006]. Un completo estudio sobre la efectividad de estos algoritmos de clustering puede ser encontrado en [Handl and Dorigo, 2003].

5. RESULTADOS EXPERIMENTALES Y CONCLUSIONES

5.1. Corpus de documentos

Con objeto de mostrar la efectividad de las técnicas expuestas, se realizó un experimento tomando como base un *corpus* de documentos formado por 250 documentos extraídos de bases de datos de artículos científicos en el campo de la empresa y la economía. Los artículos han sido publicados en revistas nacionales e internacionales de las áreas involucradas. Se seleccionaron artículos de las diferentes áreas funcionales de la empresa: marketing, contabilidad y finanzas, recursos humanos, sistemas de información y economía escritos en español e inglés. El corpus cuenta con 25 artículos de cada idioma y área funcional. Conociendo la clasificación a priori en esas 5 áreas, la efectividad de los algoritmos de clustering puede analizarse con diferentes medidas de calidad clásicas, como la cobertura, precisión, o la medida F.

5.2. Agrupamiento de los documentos

La Figura 1 muestra la disposición inicial de los 250 documentos en la rejilla y a la derecha la disposición final de dichos documentos tras 250.000 operaciones básicas del algoritmo de ant clustering sobre la rejilla. A pesar del número de operaciones básicas el tiempo de ejecución de las mismas es de tan solo 1943 milisegundos sobre un procesador Intel Pentium M 1.50 Ghz. En concreto los valores de los diferentes parámetros de configuración del algoritmo fueron los siguientes:

Coeficientes de la combinación lineal convexa de cálculo de similitudes: $\lambda_1 = 0.45$ $\lambda_2 = 0.45$ $\lambda_3 = 0.05$ $\lambda_4 = 0.05$; *t*amaño de la colonia (número de hormigas): *25*; tamaño de la memoria (número de posiciones recordadas): *20*; paso máximo de

avance en la rejilla: *25*; parámetros de recogida y colocación: $k^+ = 0.0015$ $k^- = 0.05$; radio de percepción del entorno: $\sigma = 5$ y dimensiones de la rejilla: 60×60.

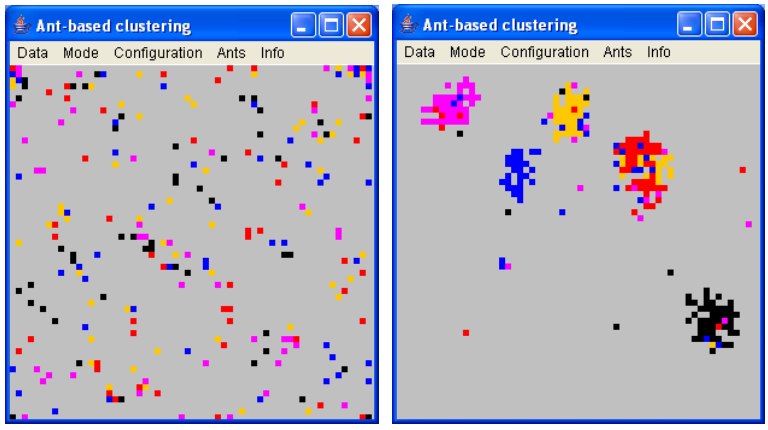

Figura 1: Situación inicial (izda.) y final (derecha) de los documentos en la rejilla.

En el agrupamiento final se observan claramente 5 grupos, correspondiendo a las 5 categorías predefinidas. La Tabla 1 muestra el nivel de pureza de cada uno de los grupos, es decir el porcentaje de miembros pertenecientes a la categoría dominante en él. Finalmente, como medida de calidad final del conjunto del agrupamiento se obtuvo un valor de la medida F de 0.576.

Cluster	*Temática dominante*	*% pureza*	*Num docs.*	*% español*	*% inglés*
C1	Economía	84,09	44	59,09	40,91
C2	Sistemas de información	71,74	46	26,09	73,91
C3	Marketing	96,3	27	44,44	55,56
C4	Recursos Humanos	57,53	73	64,38	35,62
C5	Contabilidad y finanzas	89,8	49	46,94	53,06

Tabla 1: Resultados del agrupamiento.

5.3. Conclusiones

En este trabajo se ha presentado una aplicación que permite al usuario administrar y gestionar una colección de documentos económicos escritos en español e inglés. La aplicación extrae rasgos de cada documento utilizando un glosario de términos económicos y el tesauro Eurovoc, calcula la similitudes entre los documentos escritos en diferentes lenguajes utilizando cuatro tipos de rasgos e implementa algoritmos de *ant clustering*. Estos algoritmos han sido aplicados en contextos de recuperación de documentos, sin embargo no se han encontrado aplicaciones en el ámbito multilingüe. Los resultados obtenidos sobre un corpus de 250 artículos científicos de diferentas áreas de la empresa y economía han sido satisfactorios.

6. AGRADECIMENTOS

Agradecimiento por la concesión de licencias de uso para fines de investigación de sus recursos lingüísticos a la Oficina Oficial de Publicaciones de la Comunidad Europea y al Servicio de Idiomas del Fondo Monetario Internacional (FMI). Se agradece igualmente a la profesora Julia Handl la cesión del código fuente del algoritmo de ant clustering.

7. REFERENCIAS BIBLIOGRÁFICAS

- BAEZA, R. y RIBEIRO, B. (1999). "Modern Information Retrieval". Addison Wesley.
- 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.
- HANDL, J. y DORIGO, M. (2003). "On the performance of ant-based clustering". In Proceedings of the 3rd International Conference on Hybrid Intelligent Systems.
- HANDL, J., KNOWLES, J., y DORIGO, M. (2006). "Ant-based clustering and topographic mapping". Artificial Life, 12:35-61.
- LUMER, E. y FAIETA, B. (1994). "Diversity and adaptation in population of clustering ants". In Proceedings of 3rd International Conference on Simulation of Adaptive Behaviour: From Animals to Animats, pages 501-508.
- MONMARCHÉ, N. (1999). "On data clustering with artificial ants". In Freitas, A., editor, AAAI-99 & GECCO-99 Workshop on Data Mining with Evolutionary Algorithms: Research Directions, pages 23-26.
- SALTON, G. (1971). "The SMART Retrieval Sistem - Experiments in Automatic Document Processing". Prentice Hall.