

Técnicas de representación de textos para clasificación no supervisada de documentos

Germán Cobo, Xavier Sevillano, Francesc Alías y Joan Claudi Socoró

Departamento de Comunicaciones y Teoría de la Señal
Enginyeria i Arquitectura La Salle, Universitat Ramon Llull
Po. Bonanova nº8 08022 - Barcelona (España)
{gcobo, xavis, falias, jclaudi}@salle.url.edu

Resumen: En este artículo se estudia el impacto de la representación del texto en el ámbito de la clasificación no supervisada (CNS) de documentos. Tomando como referencia una representación basada en un modelo de espacio vectorial de términos, se analizan diferentes técnicas de representación de los datos sobre espacios de menor dimensionalidad (obtenidas mediante técnicas de extracción de términos como el Análisis de Semántica Latente, la Factorización en Matrices No Negativas y el Análisis en Componentes Independientes) con el objetivo de mejorar la CNS de un corpus de documentos. El rendimiento ofrecido por cada una de estas técnicas de representación de textos se analiza sobre diferentes corpus de documentos y problemas de clasificación, evaluando tanto el coste computacional de los algoritmos, como los resultados de la clasificación conseguidos mediante distintas métricas de evaluación.

Palabras clave: Clasificación no supervisada de documentos, modelo de espacio vectorial, LSA, NMF, ICA.

Abstract: This paper analyzes the influence of text representation in the document clustering problem. Taking a term-based vector space model representation as a reference, several low-dimensionality data representation techniques are analyzed (derived by means of terms extraction techniques such as Latent Semantic Analysis, Non-negative Matrix Factorization and Independent Component Analysis) in order to improve clustering results. The performance of these text representation techniques is analyzed over different text corpora and several classification tasks, evaluating their computational cost and classification efficiency by means of different performance metrics.

Keywords: Document clustering, vector space model, LSA, NMF, ICA.

1. Introducción

La gestión automática de documentos de texto basada en su contenido suscita un gran interés en el seno de la comunidad científica. Esto es debido al continuo crecimiento, tanto en número como en tamaño, de las bases de datos textuales existentes en la actualidad.

La literatura sobre el análisis de textos cubre un amplio espectro de aplicaciones tales como la clasificación supervisada, la recuperación de información o la clasificación no supervisada (CNS), objeto del presente trabajo. La mayoría de las técnicas propuestas en este ámbito se basan en el paradigma del aprendizaje artificial (Sebastiani, 2002). Uno de los pilares en los que reposa su correcto funcionamiento es el uso de representaciones de los documentos que reflejen los rasgos distintivos de su contenido de la mejor manera posible. Esta cuestión resulta especialmente relevante cuando se trabaja con colecciones

de documentos no etiquetados, ya que esto implica no conocer a priori la correspondencia entre los documentos y las categorías a las que pertenecen, por lo que nos encontramos ante un problema de CNS (Jain, Murty, y Flynn, 2002).

En este contexto, una de las representaciones textuales más elementales es la basada en el Modelo del Espacio Vectorial (MEV) (Salton, 1989), que representa cada documento como un vector en un espacio multidimensional en base a los términos que lo forman. No obstante, existen en la literatura diversas técnicas de extracción de términos que permiten transformar el espacio vectorial de partida en otro de baja dimensionalidad, mediante técnicas de extracción de características, tales como: *i*) el Análisis de Semántica Latente (*Latent Semantic Analysis* ó LSA) (Deerwester et al., 1990), *ii*) el Análisis en Componentes Independientes (*Independent Compo-*

ment Analysis ó ICA) (Kolenda, Hansen, y Sigurdsson, 2000), y *iii*) la Factorización en Matrices No Negativas (*Non-Negative Matrix Factorization* ó NMF) (Lee y Seung, 1999). Las dimensiones de este nuevo espacio describen mejor las características distintivas de las temáticas a las que pertenecen los documentos. Este trabajo se centra en analizar el impacto del uso estas técnicas de extracción de términos en el ámbito de la CNS de documentos de texto.

Para ello, se comparan los algoritmos en términos de *i*) bondad de la clasificación y *ii*) coste computacional (del propio proceso de extracción de términos y de su impacto en la ejecución del algoritmo de CNS).

A modo de sumario, en este artículo se describen las técnicas de representación de documentos previamente mencionadas (sección 2). A continuación, se evalúa la bondad de la clasificación obtenida por un algoritmo clásico de CNS (sección 3) en base a diversas métricas (sección 4) y a lo largo de distintos experimentos realizados sobre dos corpus de documentos (sección 5). Finalmente, se exponen las conclusiones y líneas de futuro de este trabajo (sección 6).

2. Métodos de representación de documentos

En esta sección, se describe la representación textual MEV, por constituir el espacio vectorial de partida para la derivación del resto de representaciones (modelo de referencia). Seguidamente, se exponen los métodos de extracción de términos estudiados en este trabajo: LSA, ICA y NMF.

2.1. Modelo de Espacio Vectorial (MEV)

En este modelo, la unidad de información es el término, considerando cada documento como una colección de palabras (*bolsa de palabras*). Así, en el MEV, cada documento se representa como un vector \mathbf{d}_j de términos ponderados w_{ij} ($\mathbf{d}_j = [w_{1j} \ w_{2j} \ \dots \ w_{|\Psi|j}]^T$, donde $|\Psi|$ es el número de términos del diccionario). De este modo, una colección de $|D|$ documentos queda representada mediante una matriz de términos por documento ($\mathbf{P} = [\mathbf{d}_1 \ \mathbf{d}_2 \ \dots \ \mathbf{d}_{|D|}]$) (Salton, 1989).

Para definir completamente este espacio vectorial de representación, es necesario determinar el valor que deben tomar los pesos w_{ij} . En (Sebastiani, 2002) se plantean varias

posibilidades, tales como pesos binarios, pesos que consideran el número de apariciones del término en el documento (*term frequency - tf*) o pesos que ponderan la singularidad del término respecto del resto de términos del diccionario (*inverse document frequency - idf*). En este artículo se emplea la ponderación $tf \times idf$:

$$w_{ij} = tf_{ij} \cdot idf_i = tf_{ij} \cdot \log \left(\frac{|D|}{n_i} \right), \quad \forall n_i > 0 \quad (1)$$

siendo tf_{ij} el *tf* del término i del documento j y n_i el número total del apariciones del término i en todo el corpus.

2.2. Análisis de Semántica Latente (LSA)

Esta técnica de extracción de términos realiza una reducción de dimensiones del espacio de representación mediante la proyección de los documentos sobre un espacio ortogonal de baja dimensionalidad ($M \ll |\Psi|$) (Deerwester et al., 1990). Dicha proyección se obtiene mediante la descomposición en valores singulares (SVD de rango M) de la matriz de términos por documento \mathbf{P} :

$$\mathbf{P} \approx \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T \quad (2)$$

La matriz \mathbf{V}^T , de dimensiones $M \times |D|$, contiene la representación de los $|D|$ documentos del corpus en un nuevo espacio ortogonal de dimensión M .

2.3. Análisis en Componentes Independientes (ICA)

Este método de representación asume un modelo de variables latentes basado en la condición de independencia estadística de las mismas (Hyvarinen, Karhunen, y Oja, 2001). Trasladar este modelo al ámbito de la CNS de textos supone asumir que una colección de documentos es el resultado de la mezcla de diversas temáticas, que son modeladas como variables aleatorias estadísticamente independientes (Kolenda, Hansen, y Sigurdsson, 2000).

En este sentido, parece interesante aplicar ICA sobre el espacio definido por LSA, en un intento de conseguir una mayor separabilidad temática de los documentos. El modelo matemático que propone ICA es el siguiente:

$$\mathbf{Y} = \mathbf{W} \cdot \mathbf{V}^T \quad (3)$$

donde \mathbf{V}^T es la matriz de documentos representada según LSA, \mathbf{W} es la matriz de separación e \mathbf{Y} es la matriz de componentes independientes que representa los documentos en el espacio ICA. De entre las múltiples aproximaciones existentes que resuelven este modelo (Hyvarinen, Karhunen, y Oja, 2001), en este trabajo se ha utilizado un algoritmo que maximiza el cumulante de tercer orden (*skewness*) de las componentes independientes (Kabán y Girolami, 2000).

2.4. Factorización en Matrices No Negativas (NMF)

Al igual que ICA, la técnica NMF, originalmente propuesta en (Lee y Seung, 1999), asume la existencia de un modelo generativo de variables latentes. Pero a diferencia de ICA, NMF asume la no negatividad de los datos. En este sentido, se ha conjeturado que NMF se asemeja a un algoritmo ICA no negativo (Plumbley, 2002).

Desde un punto de vista matemático, NMF realiza la siguiente factorización de la matriz no negativa de términos por documento (\mathbf{P}):

$$\mathbf{P} \approx \mathbf{W} \cdot \mathbf{H} \quad (4)$$

Como en LSA, se realiza una aproximación de \mathbf{P} , dado que se efectúa una reducción a $M \ll |\Psi|$ dimensiones. De este modo, la matriz \mathbf{H} , de dimensiones $M \times |D|$, pasa a ser la nueva representación de los documentos del corpus. En este trabajo se ha utilizado un algoritmo NMF multiplicativo que minimiza el error de reconstrucción en base a la distancia de Kullback-Leibler (Lee y Seung, 1999).

3. El algoritmo de clasificación

En este trabajo, se ha empleado el algoritmo *k-means* para la clasificación de los documentos, dada su condición de método de CNS estándar (Jain, Murty, y Flynn, 2002). Esta técnica intenta hallar, de manera iterativa, un número predeterminado de centroides (cada centroide representa una categoría) a los que asignar aquellos datos de entrada respecto a los cuales se minimiza una cierta métrica. El algoritmo *k-means* converge cuando se alcanza una situación en la que la posición de los centroides se estabiliza. No obstante, su correcto funcionamiento depende de una serie de aspectos que se comentan seguidamente.

En primer lugar, la puesta en condiciones iniciales del algoritmo incide significati-

vamente en el resultado final. Esto se refiere al posicionamiento inicial los C centroides alrededor de los cuales el algoritmo agrupará todos los datos. Esta inicialización suele tener una componente aleatoria, que provoca que distintas ejecuciones del algoritmo sobre unos mismos datos den lugar a resultados diferentes. A este respecto, en este trabajo *i*) se han inicializado los C centroides a partir de C pequeñas perturbaciones aleatorias de un centroide único, resultante de promediar todos los datos (en nuestro caso, documentos) y *ii*) se han promediado los resultados de ejecutar el algoritmo 10 veces sobre los mismos datos (reinicializando cada vez) con el fin de obtener una solución que se independice, en cierta medida, de este problema.

En segundo lugar, el resultado de la clasificación también depende de la métrica en base a la cual el algoritmo realice la agrupación. En este artículo, teniendo en cuenta que se trabaja sobre un MEV (y sus representaciones derivadas), se ha escogido la distancia del coseno como métrica que usa el algoritmo *k-means* para establecer la distancia entre los objetos (Sebastiani, 2002).

Y por último, otro aspecto clave que afecta al correcto funcionamiento del algoritmo es la representación de los datos a clasificar. Por ello, la representación óptima de los datos deberá *i*) separar los pertenecientes a categorías distintas y agrupar a los miembros de una misma categoría y *ii*) conseguir que la separación de los datos sea detectable en base a la métrica empleada por el algoritmo. Este es el motivo del presente trabajo y se estudia en la sección 5.

4. Métricas de evaluación

En este apartado se describen las métricas que permiten evaluar la bondad de la clasificación obtenida por el algoritmo de CNS empleado en función de las distintas representaciones textuales utilizadas.

4.1. La función F_1

En base a las medidas clásicas de precisión (π) y cobertura (ρ), que evalúan la pureza e integridad del clasificador, respectivamente, la función F_1 permite estimar, a través de la media armónica de ambas métricas, la bondad del clasificador mediante un único valor (Sebastiani, 2002):

$$F_1 = \frac{2 \cdot \pi \cdot \rho}{\pi + \rho} \quad \forall \pi, \rho, F_1 \in [0, 1] \quad (5)$$

DOMINIOS					
comp	rec	sci	misc	talk.politics	religion
<i>graphics</i>	<i>autos</i>	<i>crypt</i>	<i>forsale</i>	<i>misc</i>	<i>talk.misc</i>
<i>os.ms-windows.misc</i>	<i>motorcycles</i>	<i>electronics</i>		<i>guns</i>	<i>alt.atheism</i>
<i>sys.ibm.pc.hardware</i>	<i>sport.baseball</i>	<i>medicine</i>		<i>midwest</i>	<i>soc.christian</i>
<i>sys.mac.hardware</i>	<i>sport.hockey</i>	<i>space</i>			
<i>windows.x</i>					

 Tabla 1: Distribución de *categorías* en **dominios** del corpus *MiniNewsgroups*.

4.2. Información Mutua

Dado un problema de clasificación de $|D|$ documentos en C categorías, un clasificador nos retorna un vector λ de asignaciones documento-categoría de dimensiones $|D| \times 1$, cuyas componentes adoptan valores comprendidos en $\{1, \dots, C\}$. Para evaluar la bondad de dicho clasificador, se calcula la Información Mutua (IM) entre el vector λ y un vector λ^* , correspondiente al etiquetado real de los documentos:

$$\text{IM}(\lambda, \lambda^*) = \frac{\sum_{h=1}^C \sum_{l=1}^C n_{hl} \log \left(\frac{|D|n_{hl}}{n_h n_l^*} \right)}{\sqrt{\left(\sum_{h=1}^C n_h \log \frac{n_h}{|D|} \right) \left(\sum_{l=1}^C n_l^* \log \frac{n_l^*}{|D|} \right)}} \quad (6)$$

donde n_h es el número de documentos que en λ se asignan a la categoría h , n_l^* es el número de documentos que en λ^* se asignan a la categoría l y n_{hl} es el número de documentos que, estando asignados a la categoría h en λ , también lo están a la categoría l en λ^* , siendo $\text{IM} \in [0, 1]$.

5. Experimentos y resultados

En esta sección se presenta el conjunto de experimentos realizados, comenzando por la descripción de las colecciones de documentos con los que se ha trabajado.

5.1. Corpus de documentos

En este artículo se han utilizado dos colecciones de documentos de texto. En primer lugar, se ha usado el corpus *MiniNewsgroups*, una versión reducida del corpus *20 Newsgroups*¹, ya que contiene solamente 100 documentos en inglés para cada una de las 20 categorías. Al igual que en el corpus *20 Newsgroups*, estas 20 categorías están agrupadas en 6 grandes dominios (ver Tabla 1).

¹Disponible en línea en <http://www.ics.uci.edu/~kdd/databases/20newsgroups/20newsgroups.html>

En segundo lugar, se ha utilizado un corpus de documentos formado por 286 artículos periodísticos en catalán extraídos del periódico AVUI². Estos documentos se reparten entre 6 categorías: *Política* (62 documentos), *Sociedad* (62), *Música* (45), *Teatro* (37), *Economía* (40) y *Deportes* (40).

5.2. CNS binaria

En este primer experimento se comparan los distintos métodos de representación de documentos (MEV, LSA, NMF, ICA) en el marco de tres problemas de CNS binaria de dificultad creciente extraídos del corpus *MiniNewsgroups* (Srinivasan, 2002). La dificultad de estos tres problemas se ha modulado en base al nivel de solapamiento entre categorías. Así pues, el primer problema considera los 200 documentos de dos categorías que pertenecen a dominios distintos (*alt.atheism* y *comp.graphics*, 10184 términos), que, a priori, deberían ser fácilmente separables. El segundo problema contempla dos categorías de un mismo dominio con un nivel medio de solapamiento (*rec.sport.baseball* y *rec.sport.hockey*, 7094 términos). Por último, se intentan clasificar los documentos de dos categorías muy solapadas pertenecientes a un mismo dominio (*talk.politics.midwest* y *talk.politics.misc*, 9181 términos).

Por otro lado, se analiza el efecto producido al variar la dimensionalidad (M) de las representaciones textuales basadas en extracción de términos. Para ello, se realiza un barrido de 2 a 20 dimensiones (se ha observado un progresivo empeoramiento en los resultados para $M > 20$).

La Figura 1 muestra los resultados obtenidos en términos de las métricas F_1 e IM. En ella se observa que los métodos de representación que reportan mejores resultados (tanto en F_1 como en IM) son LSA e ICA, igualando, o superando en la mayoría de los casos,

²<http://www.avui.com>

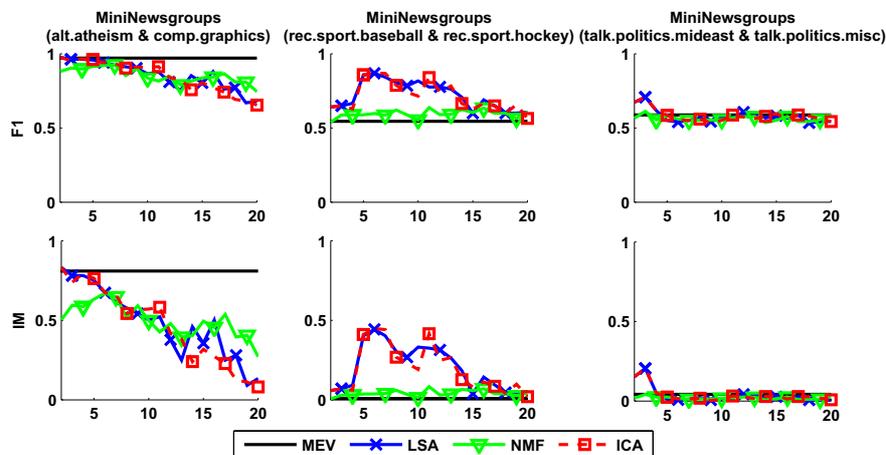


Figura 1: Resultados de CNS binaria en función del método de representación y la dimensionalidad de la misma.

a la representación de referencia (MEV). Los máximos de rendimiento se obtienen para valores bajos de M (entre $M = 2$ y $M = 7$, en función de la dificultad del problema abordado), dado el bajo número de categorías consideradas ($C = 2$).

Nótese la sensible diferencia entre los valores de F_1 e IM que se obtienen para un mismo problema de clasificación. Por ejemplo, allí donde la F_1 se sitúa alrededor de 0.5, la IM es prácticamente nula. Esto indica que ambas trabajan sobre escalas de valores diferentes. Numéricamente, IM penaliza más con más dureza los errores de clasificación que F_1 . No obstante, el hecho de tener una F_1 cercana a 0.5 en problemas de CNS binaria implica una clasificación muy pobre, prácticamente equivalente a realizar una agrupación aleatoria de los documentos.

5.3. CNS multicategoría

Este segundo experimento evalúa el impacto de las distintas metodologías de representación textual en el contexto de problemas de CNS con un mayor número de categorías. En particular se trabaja con 6 categorías y dos corpus: *MiniNewsgroups* y AVUI.

5.3.1. Corpus *MiniNewsgroups*

Para este experimento se han escogido las categorías: *comp.graphics*, *rec.autos*, *sci.crypt*, *misc.forsale*, *talk.politics.misc* y *talk.religion.misc*, pertenecientes a 6 dominios distintos. Como resultado se obtiene una

matriz \mathbf{P} de 18645 términos \times 600 documentos. Para este análisis, se ha ampliado el barrido de dimensiones llegando hasta $M = 50$, dado el aumento del número de categorías. Los resultados de este segundo experimento se muestran en la Figura 2.

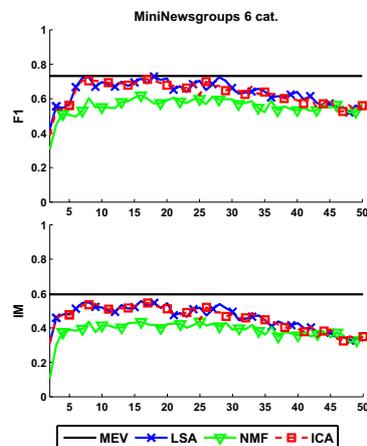


Figura 2: Resultados de CNS con 6 categorías en función del método de representación y su dimensionalidad (corpus *MiniNewsgroups*).

En este caso, los métodos de representación basados en extracción de términos que siguen reportando mejores resultados (tanto en F_1 como en IM) son LSA e ICA. Sin embargo, no logran mejorar a la representa-

ción de referencia (MEV). En este caso, dado que ha aumentado el número de categorías ($C = 6$) respecto al experimento anterior, los mejores resultados se obtienen para dimensionalidades más altas (comprendidas entre $M = 15$ y $M = 20$).

Se observa también que, a pesar de trabajar con más categorías que en el apartado 5.2, los resultados obtenidos son mejores que en los experimentos en que ambas clases pertenecen a un mismo dominio, ya que ahora el grado de solapamiento entre categorías es menor.

5.3.2. Corpus AVUI

Se ha realizado un experimento idéntico al anterior sobre las 6 categorías del corpus AVUI (la matriz \mathbf{P} contiene 20079 términos \times 286 documentos). Los resultados obtenidos se muestran en la Figura 3.

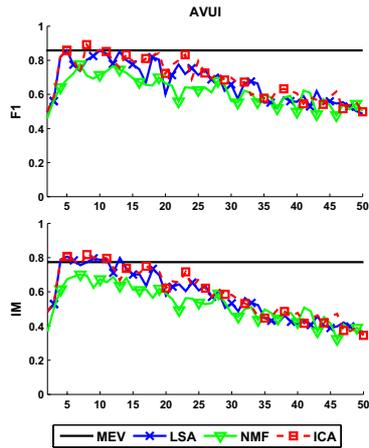


Figura 3: Resultados de CNS con 6 categorías en función del método de representación y su dimensionalidad (corpus AVUI).

En este experimento, las representaciones LSA e ICA son capaces de mejorar ligeramente los resultados obtenidos en base a MEV, correspondiendo la mejor clasificación a la representación ICA con $M = 8$. En comparación con la Figura 2, se aprecia que el corpus AVUI resulta más fácilmente clasificable que el corpus *MiniNewsgroups*. Unido este motivo al hecho de que el corpus AVUI contiene menos de la mitad de documentos (286 contra 600), el número de dimensiones necesario para alcanzar la representación óptima de los datos se reduce prácticamente a la mitad.

5.4. Análisis visual en 2D

Este experimento pretende analizar visualmente las representaciones textuales obtenidas por los distintos métodos analizados, a fin de intentar explicar por qué unas técnicas arrojan mejores resultados de clasificación que otras. Con el objetivo de facilitar el análisis visual de las representaciones de los documentos, nos centramos en un caso de reducción a $M = 2$ dimensiones, para poder representar los documentos sobre un plano.

A tal efecto, se utiliza el experimento de CNS binaria de 2 categorías muy separables (*alt.atheism* y *comp.graphics*). Tal y como se muestra en la Figura 1, las representaciones basadas en LSA e ICA alcanzan su máximo rendimiento en $M = 2$ dimensiones. Por su parte, aunque NMF tenga su máximo en $M = 6$, su comportamiento para 2 dimensiones es razonablemente cercano al óptimo.

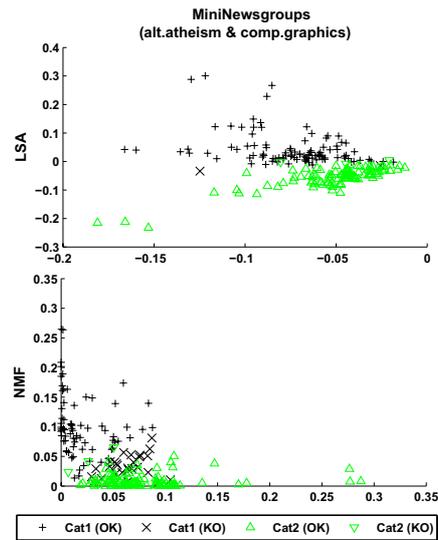


Figura 4: Representación visual sobre un plano de los documentos obtenida mediante LSA (arriba) y NMF (abajo).

En la Figura 4 se muestra la posición, sobre un plano, de los 200 documentos del corpus una vez transformados en vectores bidimensionales mediante LSA y NMF, al ser éstas las técnicas de extracción de términos que presentan comportamientos más dispares.

Concretamente, se representan con el símbolo + los documentos de la categoría

alt.atheism que han sido correctamente clasificados, mientras que el símbolo \times denota los documentos mal clasificados pertenecientes a esta categoría. Análogamente, el símbolo \triangle identifica los documentos bien clasificados de la categoría *comp.graphics* y los clasificados erróneamente son representados mediante el símbolo ∇ .

Se observa que, bajo la representación LSA, los documentos de cada categoría quedan alineados a lo largo de dos direcciones distintas del espacio, mientras que en el caso de NMF la distribución de los datos es más dispersa, de modo que las categorías se discriminan con mayor dificultad. Esto queda reflejado en un mayor número de documentos mal clasificados (símbolos \times y ∇) cuando el clasificador opera sobre la representación NMF.

Así mismo, la observación de la Figura 4 muestra que la distancia del coseno es una buena métrica para discernir entre las distintas categorías del corpus, debido al mencionado alineamiento direccional de los documentos pertenecientes a categorías diferentes.

5.5. Coste computacional

Este último experimento compara las distintas técnicas de representación textual en función de su coste computacional. La implementación de todos los algoritmos se ha realizado bajo Matlab 7.1 sobre un PC PIV a 3 GHz con 1 GB de memoria RAM y sistema operativo Windows XP.

El análisis del coste computacional se realiza en términos del tiempo de CPU necesario para ejecutar *todo* el proceso de clasificación de los textos. Más concretamente, dicho cómputo de tiempo se ha fraccionado en *i*) tiempo de ejecución del proceso de extracción de términos, y *ii*) tiempo de ejecución del algoritmo de CNS *k-means*. Esta división está motivada por el interés de comprobar si las características de las distintas representaciones textuales afectan no sólo a la bondad de la clasificación (como se ha visto en los apartados precedentes), sino también a la velocidad de ejecución del algoritmo *k-means*.

Pese a que se ha realizado el cálculo de los tiempos de ejecución para todos los experimentos descritos en esta sección, únicamente se presentan los resultados referentes al experimento de CNS binaria (*alt.atheism* vs. *comp.graphics*), ya que el comportamiento era similar para todos los casos.

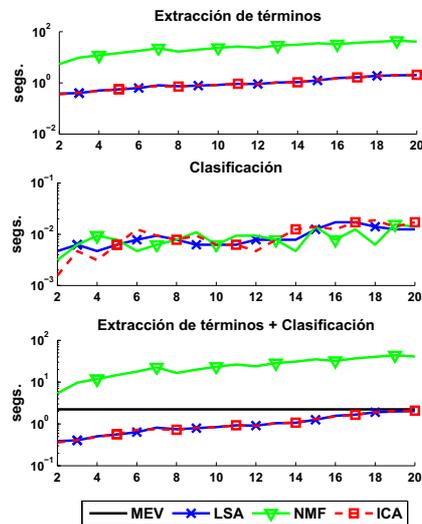


Figura 5: Tiempos de CPU (en segundos) consumidos por los procesos de extracción de términos (arriba), clasificación *k-means* (centro) y la suma de ambos (abajo), en función de la dimensionalidad del análisis.

En el gráfico superior de la Figura 5 se presenta el tiempo de CPU consumido por los procesos de representación textual basados en LSA, ICA y NMF. Se aprecia que el coste temporal del análisis NMF es muy superior al de LSA e ICA, lo que se debe a la propia naturaleza del algoritmo NMF. En concreto, la descomposición NMF se implementa mediante algoritmos iterativos que son muy sensibles a su inicialización. Por ello, para hallar la representación NMF de los documentos, es necesario lanzar el algoritmo NMF 10 veces consecutivas, antes de extraer la solución que minimice la función de coste optimizada (Xu, Liu, y Gong, 2003).

En el diagrama central de la Figura 5 se muestra el tiempo de ejecución del algoritmo de CNS *k-means* cuando es alimentado por las distintas representaciones textuales. En este gráfico se observa cómo, para una dimensionalidad M dada, el coste de ejecución del algoritmo sobre cualquiera de las representaciones estudiadas es comparable. Así pues, las características de las distintas representaciones textuales no parecen influir en la velocidad de ejecución del algoritmo *k-means*. Por otra parte, y como es lógico, el coste computacional del algoritmo de CNS aumenta al in-

crementarse la dimensionalidad del análisis.

Por último, el gráfico inferior de la Figura 5 presenta el coste total de los procesos previamente descritos, añadiendo, en este caso, el coste de clasificación resultante de alimentar al algoritmo *k-means* con la representación MEV de los documentos. Se puede apreciar como el coste computacional de este último se sitúa ligeramente por encima del coste computacional de LSA e ICA, a pesar de que no se realiza ningún proceso intermedio de extracción de términos.

6. Conclusiones

En este trabajo se ha presentado un estudio sobre distintas representaciones textuales basadas en extracción de términos para clasificación no supervisada (CNS) de documentos. El estudio ha comparado los distintos métodos en términos de la bondad de clasificación resultante y su coste computacional, en función de la dimensionalidad de la representación de los datos y sobre distintos escenarios de clasificación, variando el número de categorías y su grado de solapamiento.

Como conclusión, se ha observado que, en los problemas de CNS binaria abordados, resulta aconsejable utilizar las representaciones basadas en extracción de términos, especialmente LSA. Por contra, en los problemas de CNS multicategoría, los índices de clasificación obtenidos en base a la representación MEV se asemejan en gran medida (cuando no superan) a los conseguidos a través de LSA, ICA y NMF.

En otro orden de cosas, el coste computacional resultante de emplear técnicas de extracción de términos no conlleva una ralentización del proceso de clasificación (excepto en el caso de NMF, debido a su especial naturaleza algorítmica), sino que, muy al contrario, lo acelera (en comparación al uso del MEV, dada la muy elevada dimensionalidad de éste).

A tenor de estas conclusiones, se pretende acometer nuevas investigaciones en el ámbito de la CNS estudiando *i*) la aplicación de técnicas de selección automática de la dimensionalidad óptima del modelo (vista su influencia en los resultados) y *ii*) el uso de algoritmos de CNS más sofisticados que *k-means*, con el fin de validar la generalidad de los resultados obtenidos en este trabajo.

Bibliografía

- Deerwester, S., S.-T. Dumais, G.-W. Furnas, T.-K. Landauer, y R. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal American Society Information Science*, 6(41):391–407.
- Hyvarinen, A., J. Karhunen, y E. Oja. 2001. *Independent Component Analysis*. John Wiley and Sons.
- Jain, A., M. Murty, y P. Flynn. 2002. Data Clustering: a Survey. *ACM Computing Surveys*, 31(3):264–323.
- Kabán, A. y M. Girolami. 2000. Unsupervised Topic Separation and Keyword Identification in Document Collections: a Projection Approach. Informe técnico, Dept. of Computing and Information Systems. University of Paisley.
- Kolenda, T., L.K. Hansen, y S. Sigurdsson. 2000. Independent components in text. En M.Girolami, editor, *Advances in Independent Component Analysis*. Springer-Verlag, páginas 241–262.
- Lee, D.D. y H.S. Seung. 1999. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature*, 401:788–791.
- Plumbley, M. 2002. Conditions for non-negative independent component analysis. *IEEE Signal Processing Letters*, páginas 177–180.
- Salton, G. 1989. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley.
- Sebastiani, F. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1–47.
- Srinivasan, S.H. 2002. Features for Unsupervised Document Classification. En *Proceedings of the Conference on Computational Natural Language Learning*, páginas 36–42, Taipei, Taiwan.
- Xu, W., X. Liu, y Y. Gong. 2003. Document Clustering Based on Non-Negative Matrix Factorization. En *Proceedings of the 26th ACM SIGIR Conference*, páginas 267–273, Toronto, Canada.