CorpusMiner 1.0: Herramienta para el agrupamiento de documentos

CorpusMiner 1.0: Tool for Document Clustering

[1]Leticia Arco García, [1]Rafael Bello Pérez, [1]Manuel Llanes Abeijón, [2]Libernys Valdés Vera, [3]Juan Manuel Mederos Martínez, [3]Yoisy Pérez Olmos

[1]Departamento de Ciencia de la Computación, Universidad Central "Marta Abreu" de Las Villas (UCLV). {leticiaa, rbellop, manuela}@uclv.edu.cu

[2]Empresa Datys, libernys@gmail.com

[3]Departamento de Programación, Universidad de las Ciencias Informáticas (UCI), {juanm, yoisy}@uci.cu

REVISTA CUBANA DE CIENCIAS INFORMÁTICAS *VOL.1 No.2 ABRIL 2007 p. 18-31*

Resumen

CorpusMiner 1.0 permite formar grupos de documentos similares en un corpus textual. El agrupamiento puede ser duro o borroso. Mediante el uso de métodos de agrupamiento concatenados, no es necesario tener un conocimiento del dominio para inicializar los métodos a utilizar. El agrupamiento se realiza a partir de una representación espacio-vectorial del corpus. Se permite la aplicación de técnicas de selección de rasgos, así como diferentes funciones para el cálculo de la similitud de documentos que mejoran la eficiencia del mismo. Esta herramienta es útil en la extracción de resúmenes, categorización, clasificación, y verificación de homogeneidad de un corpus textual.

Palabras clave: Agrupamiento, corpus textuales, minería de textos.

Abstract

CorpusMiner 1.0 allows the formation of groups holding similar documents in a textual corpus. The clustering may be either crisp or fuzzy. By means of the use of concatenated clustering methods, it is not necessary any a priori knowledge on the domain so as to initialize such methods. Clustering is carried out starting with a Vector Space Model (VSM) representation of the corpus. The application of feature selection techniques, as well as different functions for computing similarity of documents is permitted in order to improve the overall performance. This tool is handy for extracting summaries, categorizing, classifying and verifying the homogeneity of a textual corpus.

Key words: Clustering, text mining, textual corpora

Introducción

Se vive una Revolución de la Información, donde ha surgido una tarea difícil: los humanos no se han diseñado para procesar cantidades masivas de información. Es por eso que en los últimos años se han desarrollado técnicas de minería de datos (Data Mining) que permiten el procesamiento de grandes volúmenes de información estructurada. Pero ha surgido otro problema: más del 80% de la información disponible en Internet es información textual y por tanto, no estructurada. Este nuevo reto propició el desarrollo de técnicas de minería de textos (Text Mining).

La minería de textos pretende identificar relaciones y modelos en la información no estructurada, así como proveer de una

visión selectiva y perfeccionada de la información contenida en documentos escritos y sacar consecuencias para la acción, detectar patrones no triviales e incluso, información sobre el conocimiento almacenado en las mismas (Tan, 1999).

Para lograr sus propósitos, la minería de textos necesita combinar varias técnicas, de ahí que sea un campo multidisciplinario que incluye la recuperación de información, el análisis de textos, la extracción de información, el agrupamiento, la construcción de resúmenes, la categorización, la clasificación, la visualización, la tecnología de bases de datos, el aprendizaje automático y la minería de datos (Tan, 1999) (Dixon, 1997).

Dentro de estas técnicas, el agrupamiento permite encontrar grupos de documentos que están relacionados por tópicos similares y extraer las palabras clave más importantes que son consideradas en esa clasificación (Berry, 2004).

El objetivo de este trabajo es presentar la herramienta CorpusMiner que permite formar grupos de documentos similares en un corpus textual. Una visión general de CorpusMiner será presentada en la sección de **Materiales y Métodos**, desglosada en una descripción de la estructura general de la herramienta, y la descripción detallada de cada uno de los módulos de la herramienta. Por ejemplo, se describirá la utilidad de trasformar el corpus inicial y cómo se lleva a cabo esta transformación. Asimismo, se mencionará de qué manera se realizó la representación del corpus y la selección de los rasgos y se hará la descripción de los métodos de agrupamiento implementados, resaltando las ventajas que brindan las variantes concatenadas propuestas. En la sección de **Resultados** se ilustrará, a partir de la definición de dos casos de estudio, cómo funciona la herramienta CorpusMiner. La sección **Discusión** estará focalizada a mostrar las ventajas de las variantes concatenadas de agrupamiento y las aplicaciones potenciales de la herramienta CorpusMiner en su primera versión. Finalmente, serán presentadas las conclusiones.

Materiales y Métodos

CorpusMiner: Visión general

Al trabajar con corpus textuales se puede realizar un análisis léxico, sintáctico o semántico, o combinaciones de estos. CorpusMiner realiza un análisis léxico, es decir, se sigue la idea de analizar el corpus como una bolsa de palabras (bag of words), donde únicamente se tiene en cuenta la frecuencia de aparición de los términos en los documentos.

En la Figura 1 se observa que la herramienta CorpusMiner 1.0 tiene cuatro módulos fundamentales. Para su correcto funcionamiento son necesarias las premisas siguientes:

- La colección de documentos en idioma inglés debe estar almacenada en un fichero texto. El formato de este fichero requiere que se especifiquen delimitadores de documentos, párrafos y oraciones (la delimitación automática de oraciones no es un problema trivial, por eso se presupone que ellas están previamente delimitadas).

- Es necesaria la especificación de diccionarios para la lematización, homogeneidad ortográfica, contracciones, abreviaturas y palabras gramaticales (stop-words).

El diseño de CorpusMiner es flexible por lo que es posible, a partir de las adecuaciones pertinentes, aplicarse al procesamiento de textos en otros idiomas.

La entrada al sistema es una colección de documentos en inglés y las salidas principales son la representación espacio-vectorial (Vector Space Model (VSM)) (Salton, 1971) del corpus y la colección de grupos homogéneos de documentos afines.

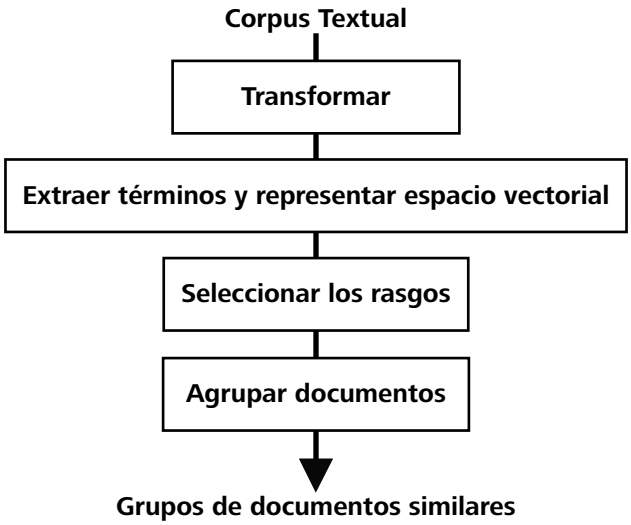

Fig. 1. Módulos principales de CorpusMiner 1.0.

Módulo 1 (Transformar): Las principales transformaciones que se incluyen son la lematización, la sustitución de términos, la homogeneización de las convenciones ortográficas, así como la reexpresión de símbolos numéricos y otros símbolos con las palabras correspondientes. La transformación incluye el etiquetado de las palabras gramaticales.

Módulo 2 (Extraer términos y representar espacio vectorial): Se realiza una representación VSM del corpus transformado, donde se puede especificar si se desea o no la inclusión de las palabras gramaticales en el análisis.

Módulo 3 (Seleccionar los rasgos): La selección de rasgos incluye varios criterios para valorar la calidad de un término, permitiendo la selección de una cantidad fijada de mejores términos, o los términos cuya calidad sobrepase un umbral.

Módulo 4 (Agrupar los documentos): Se implementaron tres algoritmos de agrupamiento, dos siguen técnicas duras deterministas y con solapamiento, y el tercero es borroso. También se crearon e implementaron variantes concatenadas que permiten utilizar al máximo las ventajas de los métodos implementados.

Transformar el corpus

Se pueden considerar dos grandes tipos de operaciones con los corpus: operaciones de conformación y operaciones de transformación.

- El primer tipo incluye las operaciones que tienen el objetivo de conformar el propio corpus mediante la adición de textos, el ordenamiento de estos en el corpus, y su delimitación y segmentación. En este software, este tipo de operación no se lleva a cabo, sino que se considera que los corpus están conformados según los requerimientos del sistema.

• El segundo tipo de operaciones, las de transformación, se realizan sobre un corpus ya conformado y se dividen en dos clases:

  ○ Transformaciones genéricas: Consisten en añadir descripciones, codificaciones, u otras transformaciones al final de cada oración del corpus.

  ○ Las transformaciones específicas incluyen dos subclases:

    ▪Transformaciones descriptivas: Añaden una descripción a cada token del corpus.

    ▪Transformaciones sustitutivas: Cambian ciertos conjuntos de tokens por otros según se especifica en ficheros auxiliares.

En CorpusMiner se realiza un análisis léxico, por tanto, las transformaciones que se desarrollan son sustitutivas e incluyen la lematización (sustitución de cada forma lingüística por el lexema correspondiente), la sustitución de las contracciones por sus expansiones, de las abreviaturas por sus formas completas, la homogeneización de las convenciones ortográficas (según las normas británicas o las norteamericanas), así como la reexpresión de símbolos numéricos (dígitos) y otros símbolos (monetarios y tipográficos) con las palabras correspondientes. También la sustitución incluye la eliminación o el etiquetado de las palabras gramaticales. Esta opción es posible al realizar la representación VSM.

Extraer términos y representar espacio vectorial

VSM es una herramienta efectiva para la representación de información introducida por Salton y sus colegas (1971) hace tres décadas. Obsérvese en la tabla 1 la representación VSM que se realiza en CorpusMiner a partir de un corpus textual.

Tabla 1. Representación VSM de un corpus textual.

	Término 1	Término 2	…	Término *m*
Documento 1	$f_{d_1}(t_1)$	$f_{d_1}(t_2)$		$f_{d_1}(t_m)$
Documento 2	$f_{d_2}(t_1)$	$f_{d_2}(t_2)$		$f_{d_2}(t_m)$
…			…	
Documento *n*	$f_{d_n}(t_1)$	$f_{d_n}(t_2)$		$f_{d_n}(t_m)$

Generalmente, a partir de la frecuencia de aparición de los términos en los documentos se modifica la matriz mediante el cálculo del *Term Frequency Inverse Document Frequency Weighting* (TF-IDF), en el cual el peso del *j*-ésimo término en el *i*-ésimo documento, denotado por *weight(i,j)*, es definido como

$$weight(i,j) = \begin{cases} (1+f_{i,j})\log_2(n/fl_j) & si\ f_{i,j} \geq 1 \\ 0, & si\ f_{i,j} = 0 \end{cases} \quad (1)$$

donde $tf_{i,j}$ es definido como el número de ocurrencias del *j*-ésimo término en el documento d_i, y df_j es número de documentos en los cuales el término aparece (Manning et. al., 2000).

Hay muchas variaciones de la fórmula TF-IDF, pero todas basadas en la idea de que el peso de los términos deba reflejar la importancia relativa de un término en un documento (con respecto a los otros términos en el documento).

En CorpusMiner se calcula TF-IDF en la representación VSM. Pero previamente se hace el cálculo de la frecuencia relativa de aparición de los términos en los documentos, para que el tamaño de estos no influya en la futura formación de grupos. Para lograrlo, las frecuencias absolutas son sometidas a un proceso de normalización.

Seleccionar los rasgos

La selección de rasgos usada para representar un dominio tiene un efecto profundo en la calidad del modelo producido, las soluciones en la minería de textos no son una excepción. Los rasgos bien seleccionados pueden mejorar la exactitud de las técnicas de minería de textos sustancialmente y reducir la cantidad de datos necesarios para obtener el nivel de funcionamiento deseado (Forman, 2003). Sin embargo, trabajos sobre la relación entre los rasgos usados para representar textos y la calidad del modelo final son menos frecuentes que resultados en dominios no textuales.

Las técnicas de selección de rasgos toman como entrada un conjunto *t* de rasgos y producen como salida un subconjunto de esos rasgos, los cuales son relevantes para el problema que se quiera resolver (Lanquillon, 2001). Obviamente, realizar una búsqueda exhaustiva es intratable desde el número de rasgos que es usualmente muy grande en el dominio de textos. Por tal motivo, la selección de rasgos puede ser guiada por heurísticas.

Existen criterios muy sencillos de selección de rasgos en dominios textuales, entre ellos:

- Eliminar las palabras gramaticales (Sam *et al*., 2000).

- Eliminar todos los términos cuyas frecuencias están por encima de un umbral superior o por debajo de un umbral inferior especificado. Estos términos tienen poco poder discriminante.

- Eliminar todos los términos cuya frecuencia de documentos es menor que un umbral predeterminado. Esto es basado en la suposición de que términos que ocurren solamente en muy pocos documentos improbablemente llevan información general de la clase específica y algunas veces tienden a ser ruidosos. Además, usar términos de ocurrencias infrecuentes no es estadísticamente confiable.

Además de estos criterios, en CorpusMiner se implementaron medidas que cuantifican la calidad de los términos y son usadas en la selección de rasgos, considerando aquellos términos que sobrepasen un umbral según la medida calculada. Las medidas implementadas son:

Calidad de un término 1 (Berry, 2004): Sea f_i la frecuencia de un término *t* en el documento d_i, la medida de calidad de un término *t* se define por:

$$q_0(t) = \sum_{i=1}^{n_0} f_i^2 - \frac{1}{n_0}\left[\sum_{i=1}^{n_0} f_i\right]^2 \quad (2)$$

donde n_0 es el número total de documentos en la colección.

Calidad de término 2 (Berry, 2004): Es una modificación de la medida anterior que permite obtener mejores resultados, ya que sustituye n_0 por n_1 que es el número de documentos en los cuales *t* ocurre al menos una vez.

Skewness y Kurtosis son medidas estadísticas que indican una distorsión de una distribución. Son usadas para conocer la parcialidad de los términos porque, la parcialidad de un término *t* se define como $P(t) = w_1 \cdot Skewness(t) + w_2 \cdot Kurtosis(t)$ donde w_1 y w_2 son pesos positivos para Skewness y Kurtosis (Fukuhara et. al., 1999).

Skewness (Fukuhara et. al., 1999): Se define para un término *t* por:

$$Skewness(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i - \bar{x})^3}{s^3} \quad (3)$$

donde x_i es la frecuencia del término *t* en el *i*-ésimo documento, $\bar{x}$ es la media y *s* la desviación estándar.

Kurtosis (Fukuhara et. al., 1999): Se define para un término *t* por:

$$Kurtosis(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i - \bar{x})^4}{s^4} - 3 \quad (4)$$

x_i es la frecuencia del término *t* en el *i*-ésimo documento, $\bar{x}$ es la media y *s* la desviación estándar.

Valores altos de *Skewness(t) y Kurtosis(t)* indican que el término *t* es más general en el corpus de textos y viceversa.

También se puede realizar una extracción de rasgos utilizando la entropía de los términos, como grado de la información que trasmiten.

Agrupar documentos

Existen varios tipos de agrupamiento, entre ellos: técnicas de análisis de grupos heurísticas, técnicas de análisis duras y deterministas, técnicas duras y con solapamiento, técnicas de análisis de grupos probabilísticas, técnicas de análisis de grupos borrosas, técnicas jerárquicas y técnicas de análisis de grupos basadas en heurísticas (Höppner et. al., 1999). En CorpusMiner se combinan tres de estas técnicas: duras deterministas, duras con solapamiento y borrosas.

Los algoritmos *Simultaneous Keyword Identification and Clustering of Text Documents* (SKWIC) y *Simultaneous Keyword Identification and Fuzzy Clustering of Text Documents* (Fuzzy SKWIC) han obtenido muy buenos resultados en el agrupamiento de documentos (Berry, 2004).

Una idea general de ambos algoritmos es la siguiente:

1. Fijar el número inicial de grupos a obtener

2. Inicializar aleatoriamente los centros de grupos

3. REPETIR

    3.1 Asignar cada documento a los grupos (en función de la menor distancia que exista entre un documento y todos los centros de grupos)

    3.2 Recalcular los centros de grupos

4. HASTA (centros estabilizados)

Ambos algoritmos retornan los grupos textuales y la relevancia de los términos en cada grupo. El primero de ellos crea una partición del corpus (duro determinista), mientras que el segundo retorna el grado de pertenencia de los documentos a los grupos (borroso), siendo mucho más efectivo en el dominio textual, porque existen textos que por su contenido pertenecen a más de una categoría.

La desventaja principal de estos algoritmos es que requieren que sea fijado el número inicial de grupos a obtener. En la mayoría de las aplicaciones no se tienen criterios para especificar correctamente este valor, por tal motivo, se propone la aplicación previa de un algoritmo (duro y con solapamiento) que no requiere que sea fijado el número de grupos iniciales como un paso previo en el agrupamiento. La salida de este algoritmo inicial se tomará para definir el número inicial de grupos, así como los centros de los grupos (ya estos no tienen que ser inicializados aleatoriamente), por tanto se reduce el número de iteraciones para lograr la convergencia.

El agrupamiento inicial se realiza mediante una extensión del algoritmo Star (Gil *et al*., 2003). El algoritmo comienza con la construcción de un grafo donde los nodos son los documentos y las aristas entre un par de documentos indican que estos son semejantes. La concepción general es llevar a cabo un proceso iterativo donde se calculan los centros de grupos, a partir de la identificación de los vecinos de los nodos.

Así, quedan definidas las dos variantes concatenadas para el agrupamiento, donde se requiere la definición de métodos interiores y exteriores. Los métodos interiores son aquellos que inicializan el proceso de agrupamiento y los exteriores son los que retornan el agrupamiento final de los documentos. Por tanto, a partir de las ventajas y desventajas que se han mencionado de los algoritmos SKWIC, Fuzzy SKWIC y Extended Star, se ha definido como método interior el algoritmo Extended Star, mientras que los métodos exteriores son los algoritmos SKWIC y Fuzzy SKWIC. De esta forma, los agrupamientos concatenados definidos permiten mejorar los resultados del agrupamiento de documentos y superar las desventajas de los métodos seleccionados. Las variantes concatenadas permiten superar las deficiencias de los métodos seleccionados por lo que ya no se hace necesario un conocimiento del dominio, ni inicializar aleatoriamente los centros de grupos, obteniéndose, no obstante, la relevancia de los términos en el mismo proceso de agrupamiento, salida que ofrecen los métodos SKWIC y Fuzzy SKWIC. Obsérvese la Figura 2.

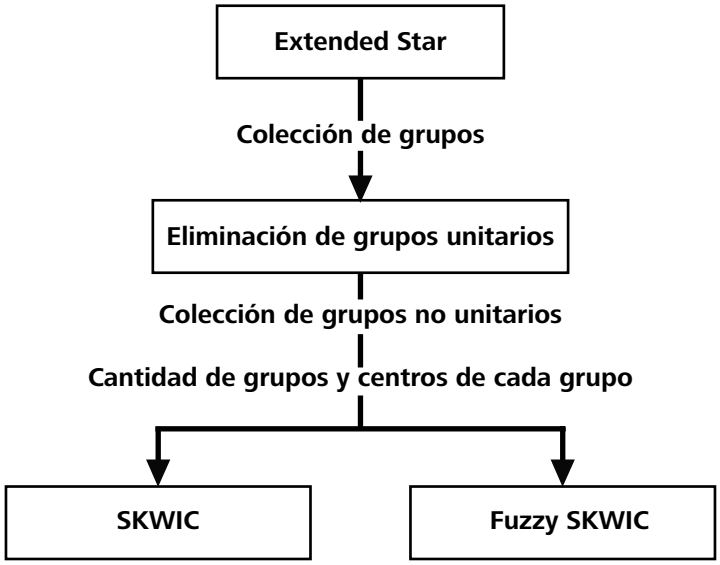

Fig. 2. Métodos de agrupamientos concatenados.

La similitud entre documentos está estrechamente relacionada con los algoritmos de agrupamiento, porque los algoritmos de agrupamiento que se implementaron en CorpusMiner tienen en cuenta la similitud entre documentos en la formación de los grupos. Muchas medidas de similitud entre documentos pueden ser utilizadas, las implementadas en CorpusMiner son similitud de Dice, Jaccard y Coseno, por ser las que han reportado los mejores resultados en dominios textuales (Fakes *et al*., 1992). Entre ellas, la similitud Coseno ha sido la más utilizada para comparar vectores de frecuencias de documentos para un vocabulario de *n* términos (Korfhage, 1977):

$$S(d_i, d_j) = \frac{\sum_{k=1}^{n} f_{k,i} \times f_{k,j}}{\sqrt{\sum_{k=1}^{n} f_{k,i}^2} \sqrt{\sum_{k=1}^{n} f_{k,j}^2}} \quad (5)$$

donde f_k,i y $f_{k,j}$ son la frecuencia de aparición del término *k* en el documento *i* y en el documento j, respectivamente.

Resultados

Se han conformado dos casos de estudio para ilustrar el funcionamiento de los cuatro módulos principales de CorpusMiner.

Primer caso de estudio. Incluye un corpus textual de la Colección de la Agencia Reuters de Noticias[1]. El corpus creado tiene un tamaño de 353 KB y posee 113 noticias previamente etiquetadas. Las noticias abordan 6 tópicos; 12 noticias se refieren a *cocoa*, 23 noticias tratan de *acq*, 12 noticias se refieren a *money-supply*, 17 noticias abordan *trade*, 24 noticias se refieren a *crude* y 25 noticias tienen como tópico *earn*.

Segundo caso de estudio. Este caso incluye un corpus textual conformado a partir de la colección Central BioMed[2] con artículos científicos sobre Medicina, Biología y Bioinformática. El corpus creado tiene un tamaño de 3.08MB y posee 123 artículos

[1]Reuters-21578 Text Categorization Collection, 135 tópicos. http://www.daviddlewis.com/resources/testcollections/
[2]BioMed Central, 22003 artículos publicados. http://www.biomedcentral.com/info/about/datamining/

científicos previamente etiquetados. Estos documentos abordan 7 tópicos; 16 artículos se refieren a *Cystic fibrosis*, 12 sobre *Genic therapy*, 6 abordan *Diabetes mellitus (therapy and diet)*, 32 se refieren a *Diabetes mellitus (research and molecular biology)*, 31 sobre el *AID*, 16 artículos abordan *Lung cancer* y 10 se refieren a *Microarrays*.

En esta sección se muestran los resultados de aplicar los dos casos de estudio definidos a CorpusMiner. Como se observa en las Figuras de la 3 a la 5, la interfaz gráfica de CorpusMiner está dividida en cuatro partes fundamentales: menú principal, árbol de resultados, pantalla de resultados y panel con la historia de las operaciones realizadas.

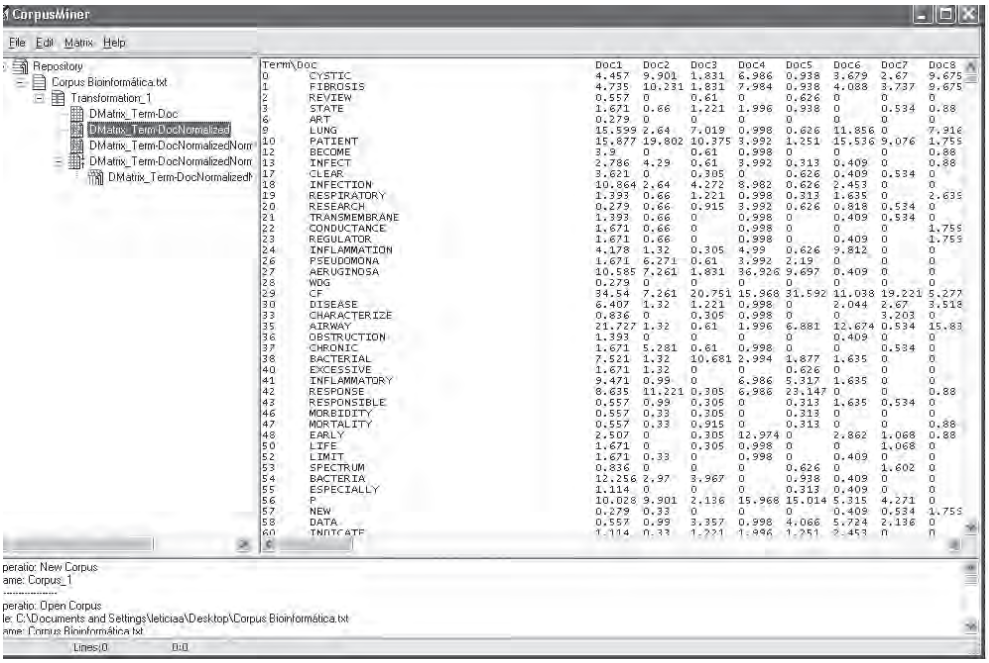

Fig. 3. Representación VSM, pesado y normalización, y reducción de dimensionalidad del corpus del segundo caso de estudio.

En la Figura 3 se muestra el corpus textual del segundo caso de estudio después de finalizada la aplicación del segundo módulo de CorpusMiner, es decir, vencidas las etapas de transformación y representación. La salida de este módulo es una matriz siguiendo la representación VSM donde se aplicaron métodos de pesado y normalización, así como la reducción de dimensionalidad, en este caso que se muestra se seleccionaron los mejores 600 términos siguiendo la medida calidad de término II (Berry, 2004).

Las figuras 4 y 5 muestran fragmentos de resultados de la aplicación del módulo 4 de CorpusMiner, es decir, la aplicación de los algoritmos de agrupamiento a ambos casos de estudio. En el primer caso fue aplicado el método concatenado Extended Star – SKWIC, mientras que en el segundo caso se aplicó SKWIC. Nótese que existe prácticamente total coincidencia entre la clasificación manual de los corpus y la que realiza CorpusMiner. Para verificar esta aclaración nótese que se publicaron por grupos textuales la primera línea de cada documento, donde se muestra él tópico al que pertenece siguiendo la clasificación manual realizada.

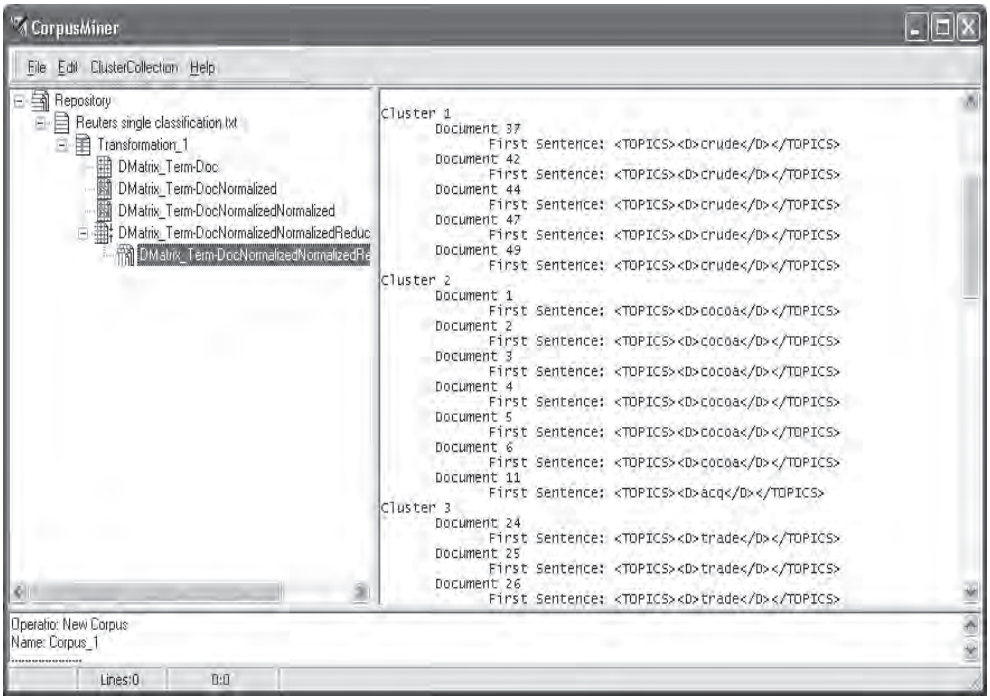

Fig. 4. Fragmento del resultado del agrupamiento concatenado Extended Star – SKWIC para el corpus del primer caso de estudio.

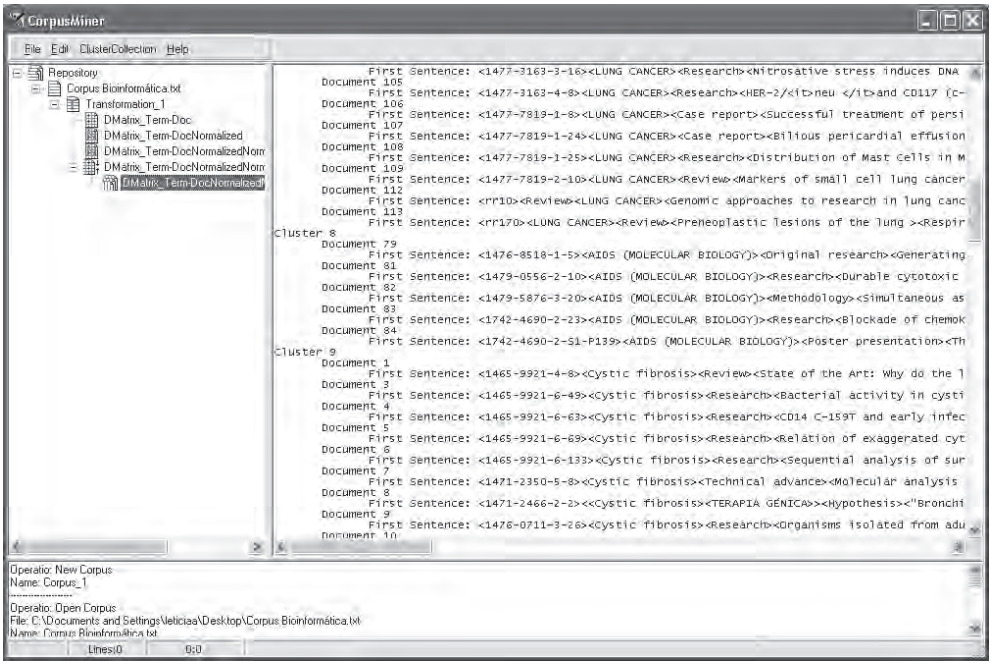

Fig. 5. Fragmento del resultado del algoritmo de agrupamiento SKWIC para el corpus del segundo caso de estudio.

Discusión

Se quiere centrar la discusión del trabajo en mostrar las ventajas que tienen las variantes concatenadas para el agrupamiento de documentos en CorpusMiner. Para ello se aplicaron medidas que permiten evaluar la calidad del agrupamiento: *Precision, Recall, F-Measure*, Entropía y *Overall Similarity* (Stein *et al*., 2000) (Frankes *et al*., 1992).

A partir de la colección de la Agencia de Noticias Reuters se conformaron 20 corpus textuales con un tamaño promedio de 160 KB, con 88 documentos incluidos como promedio que abordan 32 tópicos aproximadamente. Nótese que el número elevado de tópicos se debe a que las noticias están multiclasificadas.El objetivo de la evaluación del agrupamiento es verificar si el método concatenado propuesto logra superar a los algoritmos SKWIC y Fuzzy SKWIC aplicados de manera independiente. Por tanto, se aplicaron al caso de estudio descrito, los métodos SKWIC y Fuzzy SKWIC, así como los métodos concatenados Extended Star – SKWIC y Extended Star – Fuzzy SKWIC.

La evaluación se realizó utilizando el paquete de análisis estadístico SPSS versión 13.0. A partir de estos datos se realizaron pruebas pareadas no paramétricas de Wilcoxon. El test de Wilcoxon plantea como hipótesis fundamental que las muestras son estadísticamente iguales. A partir de los valores de significación arrojados por test se podrá comprobar si existen o no diferencias significativas en los algoritmos y sus variantes concatenadas.

Respecto a las medidas Precisión y Entropía existen diferencias significativas entre los algoritmos SKWIC y Extended Star – SKWIC, con valores respectivos de significación del test de Wilcoxon 0.025 y 0.023. Considerando la medida *F-Measure* existen diferencias altamente significativas entre SKWIC y Extended Star – SKWIC, reflejado en un valor de significación de 0.002. Los resultados de estos algoritmos son comparables teniendo en cuenta la medida *Overall Similarity*. Una situación similar se aprecia al comparar los algoritmos Fuzzy SKWIC y Extended Star – Fuzzy SKWIC. A partir de los valores de significación del test teniendo en cuenta las métricas Precision y *F-Measure*, se aprecia que la diferencia entre los algoritmos es significativa, ya que los valores de significación son 0.031 y 0.045 respectivamente. Considerando la significación 0.052 al comparar los algoritmos Fuzzy SKWIC y Extended Star – Fuzzy SWCK según Entropía, se aprecian diferencias medianamente significativas. Mientras que los resultados según *Overall Similarity* son comparables, con una significación 0.232.

El test de Wilcoxon logra mostrar dónde existen diferencias, sin embargo se hace necesario observar resultados de una estadística descriptiva y los rangos para determinar cuál algoritmo logra un mejor agrupamiento. Así, se particulariza el análisis comparativo de los algoritmos SKWIC y Extended Star – SKWIC con las métricas *Precision, F-Measure* y Entropía que fueron las que mostraron diferencias entre ambos. Los valores medios de *Precision* y *F-Measure* en Extended Star – SKWIC superan a los obtenidos en SKWIC, resultado que es deseado, ya que valores mayores de estas métricas indican mayor calidad del agrupamiento evaluado. El valor medio de Entropía en el método concatenado duro es menor que en el algoritmo SKWIC, evidenciándose que el método concatenado logra grupos más compactos, resultado también deseado.

Este análisis se puede ampliar considerando los rangos negativos, positivos y empates entre los algoritmos SKWIC y Extended Star – SKWIC al evaluar con *Precision, F-Measure* y Entropía los 20 corpus textuales. En 16 corpus los valores de *Precision* obtenidos por Extended Star – SKWIC superan los obtenidos al agrupar con SKWIC. Similarmente, 17

de los corpus evaluados tienen valores superiores de F-Measure cuando se agrupan con Extended Star – SKWIC que cuando se aplica SKWIC. Por su parte, valores de entropía bajos son deseados, y en correspondencia con los resultados anteriores, en 16 corpus dichos valores son menores cuando se evalúa con Extended Star – SKWIC que cuando se hace el agrupamiento con SKWIC.

Un análisis similar se puede realizar al comparar los algoritmos Fuzzy SKWIC y el método concatenado Extended Star – Fuzzy SKWIC. Al comparar las variantes borrosas, concatenadas o no, se aprecia que el método concatenado logra mayores valores de *Precision* y *F-Measure* que la variante Fuzzy SKWIC. Por su parte, la Entropía es menor al agrupar con el método concatenado, lo que refleja que esta variante logra grupos más compactos que Fuzzy SKWIC.

Estos resultados demuestran que utilizar la salida del algoritmo Extended Star para inicializar SKWIC y Fuzzy SKWIC, reporta mejores resultados que con una inicialización aleatoria de los algoritmos y donde es necesario tener en cuenta conocimiento del dominio. Es importante señalar que en CorpusMiner se consideran en la inicialización los centros de aquellos grupos que no son aislados (i.e., que tienen más de un documento), este es un elemento que contribuye a obtener mejores resultados, debido a que se inicializa con aquellos centros que se corresponden altamente con los tópicos que aborda el corpus textual.

El agrupamiento es una tarea importante en el correcto funcionamiento de muchos sistemas de minería de textos y recuperación de información. El agrupamiento puede ser usado eficientemente para encontrar los vecinos más cercanos de un documento, para mejorar la *Precision* y *Recall* en sistemas de recuperación de información, en la organización de motores de búsqueda y últimamente en la personalización de los resultados de los motores de búsqueda, en la verificación de la homogeneidad de un corpus textual, en la construcción de resúmenes de documentos y en la categorización de términos, entre otros.

CorpusMiner está dirigido al uso futuro de los resultados del agrupamiento en la verificación de homogeneidad, construcción de resúmenes de grupos textuales y categorización de términos con uso en una selección de rasgos por abstracción y no por extracción.

Conclusiones

Se ha presentado la herramienta CorpusMiner para el agrupamiento de documentos en su primera versión. La misma ofrece a los investigadores y desarrolladores en el campo de la minería de textos la posibilidad de realizar el agrupamiento de textos a partir de una representación espacio-vectorial del corpus textual. La herramienta cuenta con cuatro módulos fundamentales y permite en cada uno de ellos que el investigador en el área de la minería de textos seleccione y combine variantes de solución y algoritmos para cada etapa del procesamiento. Las etapas de transformación y representación textual consideran un análisis léxico de corpus en idioma inglés y utilizan una representación VSM. El agrupamiento se realiza mediante los métodos Extended Star, SKWIC y Fuzzy SKWIC, así como las variantes concatenadas Extended Star – SKWIC y Extended Star – Fuzzy SKWIC. Las variantes concatenadas han permitido superar las deficiencias de los métodos seleccionados por lo que ya no se hace necesario un conocimiento del dominio, ni inicializar aleatoriamente los centros de grupos, obteniéndose, no obstante, la relevancia de los términos en el mismo proceso de agrupamiento.

La aplicación de las pruebas estadísticas no paramétricas para el estudio de los métodos de agrupamiento permitió determinar que los métodos concatenados Extended Star – SKWIC y Extended Star – Fuzzy SKWIC reportan mejores resultados que los algoritmos SKWIC y Fuzzy SKWIC respectivamente. Además, que el algoritmo Extended Star logra los mejores valores de *Precision*, Entropía y *Overall Similarity*, lo que reafirma que se haya seleccionado como método interior en la propuesta concatenada. Por tanto, los resultados de esta evaluación permiten trazar las políticas respecto al uso de los algoritmos de agrupamiento para el desarrollo de aplicaciones de usuario final.

Los dos casos de estudio definidos permitieron ilustrar cómo funciona CorpusMiner.

Referencias

Berry, M.W. Survey of Text Mining. Clustering, Classification, and Retrieval. Springer-Verlag, 2003.

Dixon, M. An Overview of Document Mining Technology. http://citeseer.ist.psu.edu/dixon97overview. 1997.

Fakes, W.B. Baeza-Yates Information Retrieval, Data Structures and Algorithms. Prentice Hall. 1992.

Forman, G. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. Journal of Machine Learning Research 3. 2003, p. 1289-1305.

Fukuhara, T. Takeda, H. Multiple-text Summarization for Collective Knowledge Formation. Proceedings of Workshop on Social Aspects of Knowledge and Memory. IEEE Systems, Man and Cybernetics. 1999.

Gil, R. Baldía, J.M. Pons, A. Extended Star Clustering Algorithm. Proceedings of CIARP. 2003.

Hoppner, F. Klawonn, F. Kruse, R. Fuzzy Cluster Analysis. John Wiley & Sons, LTD. 1999.

Korfhage, R.R. Information Storage and Retrieval. Wiley, New York, 1977.

Lanquillon, C. Enhacing Text Classification to Improve Information Filtering. Tesis doctoral. Universidad de Magdeburgo, Alemania. 2001.

Manning, C. Shutze, H. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA. 2000.

Salton, G. The SMART Retrieval System. Prentice-Hall, Englewood Cliffs, NJ. 1971.

Sam, H. Karypis, G. Centroid-Based Document Classification: Analysis & Experimental Results. In Proceedings of the Fourth European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD) Lyon, France, 2000, p. 424-431.

Stein, G. Bagga, A. Wise, G.B. Multi-document summarization: methodologies and evaluations. Proceedings of TALN. 2000.

Tan, A.H. Text Mining: The state of the art and the challenges. 1999.