Agrupamiento de Documentos Textuales mediante Métodos Concatenados

Leticia Arco, Rafael Bello, Juan M. Mederos, Yoisy Pérez

Centro de Estudios de Informática,
Universidad Central "Marta Abreu" de Las Villas
Carretera a Camajuaní, km 5 ½
Santa Clara, Villa Clara, Cuba 54830
{leticiaa,rbellop,jmm,ypo}@uclv.edu.cu

Resumen

Este trabajo tiene como objetivo mostrar una propuesta de agrupamiento de corpus textuales mediante métodos concatenados y su evaluación a partir de resultados experimentales. Los algoritmos incluidos son *Extended Star*, SKWIC y *Fuzzy* SKWIC. El algoritmo *Extended Star* es considerado un método interno, mientras que los algoritmos SKWIC y *Fuzzy* SKWIC constituyen los dos métodos exteriores en las dos variantes de concatenación propuestas: *Extended Star* – SKWIC y *Extended Star* – *Fuzzy* SKWIC. El primer método concatenado emplea una técnica de agrupamiento dura y determinista y el segundo es un método borroso. Es ventajoso usar estos métodos concatenados principalmente cuando se desean realizar procesamientos posteriores a los grupos de documentos creados y cuando no se tiene un conocimiento previo del dominio. Finalmente, se muestra la viabilidad de los métodos concatenados propuestos a partir de la aplicación, a la herramienta CorpusMiner que soporta dichos métodos, de un caso de estudio construido a partir de una colección de la agencia de noticias Reuters. Se evaluó la propuesta utilizando pruebas estadísticas no paramétricas y se demostró que las variantes concatenadas superan los resultados del agrupamiento respecto a los algoritmos originales.

Palabras clave: Agrupamiento, Corpus de Documentos, Minería de Textos.

1. Introducción

La minería de datos (*data mining*) ha sido definida como la extracción de información implícita, previamente desconocida y potencialmente útil desde datos estructurados. La mayoría de los esfuerzos se han focalizado en el descubrimiento de conocimiento desde bases de datos estructuradas, sin embargo, una gran cantidad de información aparece solamente en colecciones de textos no estructurados [Feldman et al. 96].

La minería de textos (*text mining*), también conocida como minería de datos textuales o descubrimiento de conocimiento desde bases textuales no estructuradas, pretende algo similar a la minería de datos: identificar relaciones y modelos en la información, pero a diferencia de la minería de datos, lo hace a partir de información no cuantitativa. Es decir, proveer una visión selectiva y perfeccionada de la información contenida en documentos, sacar consecuencias para la acción y detectar patrones no triviales e información sobre el conocimiento almacenado en las mismas [Berry04].

Existen varias áreas que conforman la minería de textos. Entre ellas, la recuperación y extracción de información, el análisis de textos, el resumen, el agrupamiento, la categorización y la clasificación de documentos.

El análisis de clusters o agrupamiento de datos es una actividad humana muy importante. Esta actividad usualmente forma las bases del aprendizaje y del conocimiento. La minería de textos no constituye una excepción respecto a la importancia de la aplicación de técnicas de análisis de clusters. Éste puede ser usado eficientemente para encontrar los vecinos más cercanos de un documento, para mejorar la calidad de sistemas de recuperación de información, en la organización y personalización de la información en motores de búsqueda, en la verificación de la homogeneidad de un corpus textual, en el resumen de colección de documentos y en la categorización de términos, entre otros.

Este trabajo tiene como objetivo mostrar una propuesta de agrupamiento de corpus textuales mediante la concatenación de métodos de agrupamiento ya existentes, así como evaluar, a partir de la experimentación, la propuesta realizada.

A continuación se formaliza en qué consiste el agrupamiento, así como se describen las principales técnicas de agrupamiento existentes. En la sección 3 se citan los trabajos relacionados y se muestra la motivación para el desarrollo de esta investigación. En la sección 4 se describe la propuesta de agrupamiento de documentos mediante los métodos concatenados: *Extended Star* – SKWIC y *Extended Star – Fuzzy* SKWIC. Finalmente, en la sección 5 se validan los resultados.

2. Técnicas de agrupamiento

El objetivo del agrupamiento de documentos es formar una colección de clusters (subconjuntos, grupos, clases) que cumplan las propiedades siguientes [Höppner et al. 99]:

- La homogeneidad dentro de los clusters, i.e. los documentos que pertenecen al mismo cluster deben ser tan similares como se pueda.
- La heterogeneidad entre clusters, i.e. los documentos que pertenecen a clusters diferentes deben ser tan diferentes como se pueda.

El agrupamiento es un proceso de división de un conjunto de datos u objetos en un conjunto de subclases significativas, llamadas clusters. Formalmente, dada una colección de n objetos descritos por un conjunto de p atributos, el objetivo del agrupamiento es derivar una división útil de los n objetos en un número de clusters. Un cluster es una colección de objetos similares entre sí, basado en los valores de sus atributos, y puede ser tratado colectivamente como un grupo. El agrupamiento es útil para obtener una distribución interna del conjunto de datos [Fung02].

Al agrupar los objetos de un conjunto de datos, se requieren algunas medidas para cuantificar el grado de asociación entre ellos. Con este propósito, se pueden utilizar distancias, o medidas de similitud o disimilitud. Algunos algoritmos de agrupamiento tienen un requerimiento teórico para el uso de una medida específica, pero lo más común es que el investigador seleccione qué medida utilizará con determinado método. Una distancia ampliamente utilizada es la Euclidiana, sin embargo en dominios textuales no arroja buenos resultados. La mayoría de los estudios coinciden que en minería de textos es adecuado utilizar similitud Coseno, coeficiente Dice o Jaccard, preferiblemente la similitud Coseno expresada en la ecuación 2.1 [Frankes et al. 92]

$$S(d_i, d_j) = \frac{\sum_{h=1}^{k} \left(peso_{ih} \cdot peso_{jh} \right)}{\sqrt{\sum_{h=1}^{k} peso_{ih}^2 \cdot \sum_{h=1}^{k} peso_{jh}^2}} \qquad (2.1)$$

donde d_i y d_j son los documentos a comparar, k es el número de términos que caracterizan los documentos y $peso_{xy}$ es el peso del término y en el documento x, calculado teniendo en cuenta la frecuencia de aparición de ese término en el documento, siguiendo diversos criterios.

Existen varios tipos de técnicas de agrupamiento, entre ellas, técnicas de agrupamiento incompleto o heurístico, técnicas de agrupamiento duro y determinista, técnicas de agrupamiento duro y con solapamiento, técnicas de agrupamiento probabilísticas, técnicas de agrupamiento borroso, técnicas de agrupamiento jerárquico, técnicas de agrupamiento basadas en funciones objetivos y técnicas de estimación de grupos [Höppner et al. 99].

En este trabajo se muestran algoritmos que aplican técnicas de agrupamiento duro y determinista, duro y con solapamiento, y borroso. En el agrupamiento duro y determinista se asigna cada dato exactamente a un cluster de modo que la partición de clusters defina una partición ordinaria del conjunto de los

datos. Mientras que en el agrupamiento duro y con solapamiento cada dato será asignado al menos a un cluster, o puede ser simultáneamente asignado a varios clusters. Los algoritmos de agrupamiento borroso puro trabajan con grados de pertenencia que indican en qué medida un dato pertenece a los clusters. La suma de las pertenencias de cada dato a todos los clusters es igual a uno [Höppner et al. 99].

3. Trabajos relacionados y motivación

Varios algoritmos que aplican técnicas duras y deterministas, las técnicas duras y con solapamiento, y las técnicas de agrupamiento borroso, se han aplicado al agrupamiento de documentos.

Algunos ejemplos de algoritmos que aplican técnicas duras y deterministas son: la red de Kohonen (*Self-Organizing Maps*) para el agrupamiento de documentos [Nürnberger et al. 01], el algoritmo *Autoclass* [Larocca et al. 00], el *k-means* que es un clásico en el análisis de clusters y también se ha utilizado en dominios textuales, así como algunas de sus variantes *Batch k-means, Incremental k-means* y *Means* [McQueen67] [Jang97] [Kogan01], y el algoritmo *Simultaneous Clustering and Attribute Discrimination* (SCAD) [Frigui et al. 00].

Los algoritmos *Star* [Aslam et al. 98] y *Extended Star* [Gil-García et al. 03] son ejemplos de algoritmos que aplican técnicas duras y con solapamiento, mientras que algunos muy usados en el agrupamiento borroso son *Fuzzy c-means* [Bezdek73] [Jang97] y *Relational Alternating Cluster Estimation* [Runkler et al. 01].

Todos estos algoritmos tienen ventajas y desventajas que son necesarias tener en cuenta al aplicarlos en la solución de un problema. Algunos de ellos requieren que el número de clusters a obtener sea especificado a priori, por tanto es necesario un cierto conocimiento del dominio y en muchos casos, la calidad de la partición final depende de una buena selección de la partición inicial.

Al agrupar documentos es muy útil obtener simultáneamente los grupos de documentos y las palabras claves de cada grupo. Así, se divide la colección de documentos en categorías más significativas y se genera automáticamente una descripción compacta de cada cluster en términos no sólo de los valores de los atributos, sino también de su relevancia.

En este trabajo se propone una variante de agrupamiento que permite simultáneamente agrupar los documentos y obtener las palabras claves que caracterizan cada grupo, sin necesidad de tener conocimiento previo del dominio.

En [Berry04] se muestran los algoritmos *Simultaneous Keyword Identification and Clustering of Text Documents* (SKWIC) y *Simultaneous Soft Clustering and Term Weighting of the Text Documents* (*Fuzzy* SKWIC) que devuelven la colección de clusters y la relevancia de las palabras por clusters, pero requieren que sea especificado el número inicial de clusters a obtener. Mientras que el algoritmo *Extended Star* [Gil-García et al. 03] no requiere especificar inicialmente el número de clusters a obtener, pero no calcula las palabras claves en el propio proceso de agrupamiento.

4. Métodos concatenados para agrupar documentos

Para el desarrollo de esta investigación se han seleccionado los algoritmos SKWIC y *Fuzzy* SKWIC [Berry04] porque simultáneamente logran agrupar y calcular la relevancia de las palabras por grupos, cuestión fundamental cuando el objetivo del agrupamiento es obtener posteriormente resúmenes extractos de los grupos homogéneos obtenidos, y por tanto, la relevancia de las palabras es fundamental. *Extended Star* [Gil-García et al. 03] fue seleccionado debido a que no requiere un conocimiento previo del dominio en la determinación de grupos a obtener.

Como resultado de este trabajo se proponen dos métodos de agrupamiento a partir de la concatenación de algoritmos, para lograr de esta forma, explotar las ventajas de los métodos ya existentes en el desarrollo de métodos concatenados que produzcan un agrupamiento de mayor calidad.

Los tres algoritmos seleccionados tienen en común que parten de una representación espacio vectorial (*Vector Space Model* VSM) [Salton et al. 75] del corpus textual y devuelven una colección de clusters. La estructura de los clusters varía en dependencia de los algoritmos de agrupamiento. SKWIC y *Fuzzy* SKWIC retornan una colección de clusters, donde cada cluster tiene el vector centro de cluster (el centro de cada cluster es un documento ideal), una lista de la relevancia de las palabras en el cluster y una lista de documentos con sus

particularidades si es SKWIC o *Fuzzy* SKWIC. En el caso específico de SKWIC, al ser duro y determinista, cada cluster tiene la lista de los documentos que lo conforman y el algoritmo crea una partición. En el caso del algoritmo *Fuzzy* SKWIC, todos los documentos de la colección pertenecen a todos los clusters, pero con un grado de pertenencia asociado, por tanto, la lista de documentos de cada cluster tiene por cada documento su grado de pertenencia al cluster. El algoritmo *Extended Star* también retorna una colección de clusters, donde cada cluster tiene señalado el documento que es centro o estrella del cluster (el centro de cada cluster coincide con un documento de la colección), así como una lista con los documentos de la colección que pertenecen a él. Este algoritmo permite que un documento pertenezca a más de un cluster, ya que aplica una técnica dura y con solapamiento.

En la propuesta de agrupamiento concatenado que se realiza en este trabajo, se han clasificado los métodos que se concatenan en métodos interiores y métodos 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.

Los métodos externos son SKWIC y *Fuzzy* SKWIC porque ambos logran calcular la relevancia de las palabras a los clusters simultáneamente al proceso de agrupamiento de documentos, característica que tiene gran importancia cuando se desea resumir los grupos de documentos afines o determinar los términos que caracterizan cada grupo. Estos métodos son ventajosos porque utilizan una variante del coeficiente Coseno [Frankes et al. 92] para calcular la disimilitud entre documentos, medida que tiene muy buenos resultados en el agrupamiento de textos. Los documentos, por sus características, pueden pertenecer a más de un grupo, de ahí la utilidad de incorporar *Fuzzy* SKWIC, aunque SKWIC haya sido incorporado.

Ambos algoritmos tienen la desventaja que requieren que el número de clusters a obtener sea especificado a priori, por tanto, es necesario tener conocimiento del dominio para poder definir ese valor y en la mayoría de las aplicaciones no contamos con el conocimiento suficiente. Otra desventaja de estos algoritmos es que en la primera iteración se seleccionan aleatoriamente los centros de clusters, por tanto, lograr que los centros queden estabilizados puede consumir varias iteraciones, además, debido a esa aleatoriedad, el algoritmo SKWIC puede devolver centros de clusters vacíos, es decir, que se haya especificado una cantidad k de clusters a obtener y realmente retorne solo t clusters ($t<k$), ya que los centros de clusters aleatoriamente generados al incio no sean representativos de los grupos y por tanto, en las iteraciones iniciales no se les asigne ningún documento y esos clusters desaparezcan.

Las desventajas antes mencionadas pueden ser resueltas cuando estos métodos son inicializados con la salida del algoritmo que se ha seleccionado como interior, *Extended Star* [Gil-García et al. 03]. Una gran ventaja de este algoritmo es que no requiere un conocimiento previo del dominio, porque no es necesario fijar el número de clusters inicialmente. Esta ventaja es la que la da la condición de método interior en esta propuesta. El algoritmo *Extended Star* tiene otra ventaja, y es que permite calcular la similitud entre vectores utilizando diferentes medidas. Cuando el algoritmo se utiliza para el agrupamiento de documentos, es conveniente calcular la similitud entre los vectores documentos con las medidas Dice, Jaccard y Coseno, que son las que arrojan los mejores resultados en los dominios textuales [Frankes et al. 92]. Independientemente de la calidad del agrupamiento que se pueda obtener con *Extended Star*, este algoritmo no calcula simultáneamente al proceso de agrupamiento la relevancia de las palabras a los clusters, por tanto, esto reafirma la incorporación del método como un método interior. Otra desventaja del algoritmo *Extended Star* es que se generan muchos clusters con un único documento, situación que se regula en la concatenación, porque esos centros de clusters no se considerarán como centros de clusters en la inicialización de los métodos externos.

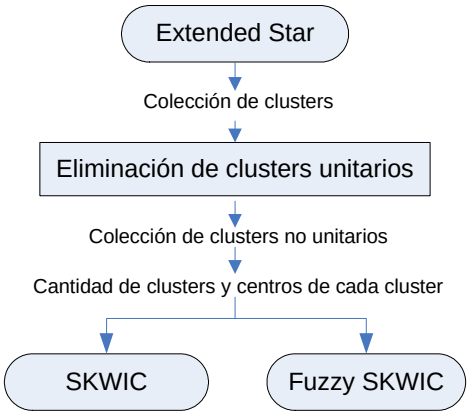

Figura 1. Concatenación del método de agrupamiento *Extended Star* con los métodos SKWIC y *Fuzzy* SKWIC

Obsérvese en la figura 1 que en el proceso de agrupamiento concatenado se ha considerado *Extended Star* como algoritmo interior y los algoritmos externos pueden ser SKWIC o *Fuzzy* SKWIC. A partir de la salida del método *Extended Star*, se eliminan los clusters unitarios y se se inicializa el número de clusters, así como sus centros, en los algoritmos SKWIC y *Fuzzy* SKWIC. A partir de esta concatenación, ya no es necesario tener un conocimiento previo del dominio para definir el número de clusters a obtener, ni generar aleatoriamente los centros de clusters. Por tanto, se superan las deficiencias de los algoritmos SKWIC y *Fuzzy* SKWIC con la concatenación. Además, al no generar los centros de clusters aleatoriamente en la primera iteración, se reduce la cantidad de clusters vacíos en SKWIC. Se supera la deficiencia de la gran cantidad de clusters aislados que produce *Extended Star*, al sólo considerar en la inicialización de los métodos externos el número de clusters con presencia de dos documentos o más.

Es importante analizar en esta concatenación de métodos la complejidad temporal de la propuesta. La complejidad del algoritmo *Extended Star* es de un $O(n^2m)$, donde n es el número de documentos y m el número de términos que describen los documentos, por tanto, éste es un tiempo que se le adiciona a la tradicional ejecución de SKWIC y *Fuzzy* SKWIC. Sin embargo, la complejidad temporal de la concatenación no es la suma de las complejidades de ambos algoritmos, porque al considerar como método interior *Extended Star*, se disminuye sustancialmente el número de iteraciones a realizar por los algoritmos SKWIC y *Fuzzy* SKWIC. Así, la concatenación no sólo logra mejores resultados sacrificando complejidad temporal, sino que logra reducciones de tiempo en los métodos externos, ya que éstos logran converger en un menor tiempo.

5. Validación de los resultados

Los métodos concatenados propuestos para el agrupamiento de documentos (*Extended Star* – SKWIC y *Extended Star* – *Fuzzy* SKWIC) son soportados por el software CorpusMiner [Arco05]. CorpusMiner es un software desarrollado por los autores de este trabajo que consta de cuatro etapas en el procesamiento textual: representación VSM del corpus, reducción de la dimensionalidad, agrupamiento y obtención de resúmenes extractos de los grupos textuales. Obsérvese en la figura 2 las etapas a las que es sometido el corpus textual antes de aplicarle métodos de agrupamiento.

Al ser los documentos datos no estructurados, se hace necesario un preprocesamiento de los mismos. En CorpusMiner el preprocesamiento se realiza automáticamente y permite la transformación, pesado y reducción de la dimensionalidad del corpus. La transformación incluye las operaciones siguientes: convertir todos los caracteres a mayúscula, la sustitución de las contracciones por sus expansiones, de las abreviaturas por sus formas completas y la eliminación de números y símbolos, la lematización y verificación de la homogeneidad ortográfica. La mayoría de las formas de pesado se basa en alguna variación de la fórmula TF-IDF. La idea de una expresión TF-IDF es 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 permite el pesado de la representación con una variante de TF-IDF publicada en [Manning et al. 00]. La reducción de la dimensionalidad se realiza a partir de la eliminación de las palabras gramaticales y la selección de términos. Las palabras gramaticales (*function words*) son aquellas que proveen estructura en lugar de contenido en el documento y tienen una alta frecuencia de aparición (e.g. artículos, preposiciones, conjunciones). La selección se realiza de aquellos términos que tengan una calidad superior a determinado umbral considerando la medida de calidad de términos referenciada en [Dhillon, I. et al. 01].

A partir de casos de estudio diseñados y su aplicación en el software, se pudo evaluar el desempeño del agrupamiento teniendo en cuenta las

métricas entropía, F-*Measure* (*Precision* y *Recall*) y *Overall Similarity* [Stein et al. 00] [Forman03].

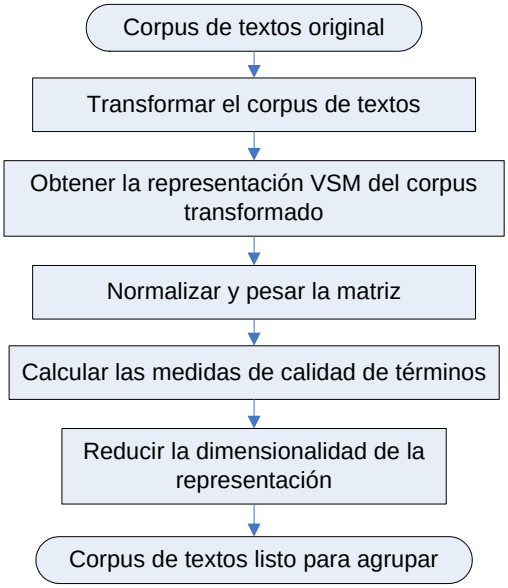

Figura 2. Etapas del preprocesamiento del corpus textual antes de someterlo al agrupamiento

5.1 Métricas para evaluar la calidad del agrupamiento

Para evaluar el agrupamiento, dos tipos de métricas de calidad son ampliamente utilizadas [Stein et al. 00]. Un tipo de métrica permite comparar diferentes conjuntos de grupos sin referenciar el conocimiento externo y es llamada métrica de calidad interna. En tal sentido, se ha seleccionado la métrica *Overall Similarity* basada en calcular la similitud de pares de documentos en un cluster. Existe otro tipo de métricas que permite evaluar cuán bueno es el método de agrupamiento mediante la comparación de los grupos producidos por la técnica de agrupamiento con las clases que se identifican en la colección. Este tipo de métrica es llamada métrica de calidad externa. Ejemplos de métricas externas son la Entropía y F-*Measure*. Estas métricas requieren que los corpus estén previamente etiquetados, es decir, utilizan una clasificación humana de referencia para evaluar el agrupamineto.

En la literatura se reportan varias métricas para validar la calidad del agrupamiento de documentos y los resultados obtenidos al aplicarlas a diferentes algoritmos de agrupamiento puede variar sustancialmente dependiendo de cual métrica fue usada. Sin embargo, si un algoritmo de agrupamiento funciona mejor que otros para la mayoría de las métricas, entonces se puede decir que ciertamente ese algoritmo es el mejor para la situación que fue evaluada.

5.1.1 Entropía

En [Stein et al. 00] usan la entropía como métrica de calidad de los clusters (con la advertencia que la menor entropía es obtenida cuando cada cluster contiene exactamente un documento). Sea *CS* un resultado de agrupamiento. Para cada cluster es calculada primero la distribución de las clases (i.e. para un cluster j se calcula p_{ij}, la probabilidad que un miembro del clusters j pertenezca a la clase i)

$$p_{ij} = \frac{n_j^i}{n_j} \tag{5.1}$$

donde n_j^i es el número de documentos de la clase i que están asignados al cluster j. Entonces, usando esa distribución de la clase, la entropía de cada cluster j es calculada usando la fórmula estándar

$$E_j = -\sum_i p_{ij} \log\left(p_{ij}\right) \tag{5.2}$$

donde la sumatoria es aplicada sobre todas las clases. La entropía total para el conjunto de clusters i es calculada como la suma de las entropías de cada cluster y han sido pesadas por el tamaño de cada cluster

$$E_{CS} = \sum_{j=1}^{m} \frac{n_j * E_j}{n} \tag{5.3}$$

donde n_j es el tamaño del cluster j, m es el número de clusters y n es el número total de documentos.

5.1.2 F-*Measure*

F-*Measure* es una métrica que combina las ideas de *Precision* y *Recall* de la recuperación de información [Frankes et al. 92] [Stein et al. 00]. Observe en las expresiones 5.4 y 5.5 el cálculo de *Precision* y *Recall* para el cluster j y la clase i

$$\Pr(i, j) = n_{ij} \Big/ n_j \tag{5.4}$$

$$\mathrm{Re}(i, j) = n_{ij} \Big/ n_i \tag{5.5}$$

donde n_{ij} es el número de miembros de la clase i en

el cluster j, n_j es el número de miembros del cluster j y n_i es el número de miembros de la clase i.

Los niveles más altos de *Precision* generalmente se obtienen con valores bajos de *Recall*. La expresión 5.6 permite calcular el valor F-*Measure*, proporcionando un parámetro α ($0 \leq \alpha \leq 1$) que permite ponderar *Precision* y *Recall*.

$$F_\alpha(i,j) = \frac{1}{\alpha \frac{1}{\Pr(i,j)} + (1-\alpha)\frac{1}{\mathrm{Re}(i,j)}} \qquad (5.6)$$

Un valor de F-*Measure* es calculado para toda la colección calculando un promedio pesado de todos los valores de las métricas F-*Measure*, según la expresión 5.7.

$$F = \sum_i \frac{n_i}{n} \max_j \{F_\alpha(i,j)\} \qquad (5.7)$$

5.1.3 *Overall Similarity*

En la ausencia de información externa, tales como las etiquetas de clases, la cohesión de los clusters puede ser usada como una métrica de similitud de clusters [Stein et al. 00]. Un método para calcular la cohesión de un cluster es usar la similitud pesada de la similitud interna del cluster,

$$OS = \frac{1}{|C_j|^2} \sum_{\substack{d \in C_j \\ d' \in C_j}} \mathrm{distance}(d', d) \qquad (5.8)$$

donde C_j es el cluster que se va a evaluar [Frankes et al. 92]. En este trabajo se ha utilizado la similitud Coseno para comparar los vectores documento. Para tener un resultado por agrupamiento se ha calculado la media de los valores *Overall Similarity* por clusters.

5.2 Descripción del caso de estudio: Corpus textuales de la agencia de noticias Reuters

Todas las métricas utilizadas para evaluar la calidad del agrupamiento, excepto *Overall Similarity*, requieren que los corpus textuales hayan sido previamente etiquetados. Colecciones de documentos existen varias, pero que estén previamente etiquetas es menos probable. Es por esto que se utilizó una colección de la agencia Reuters de noticias previamente clasificada (http://www.daviddlewis.com/resources/testcollections). A partir de la colección 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.

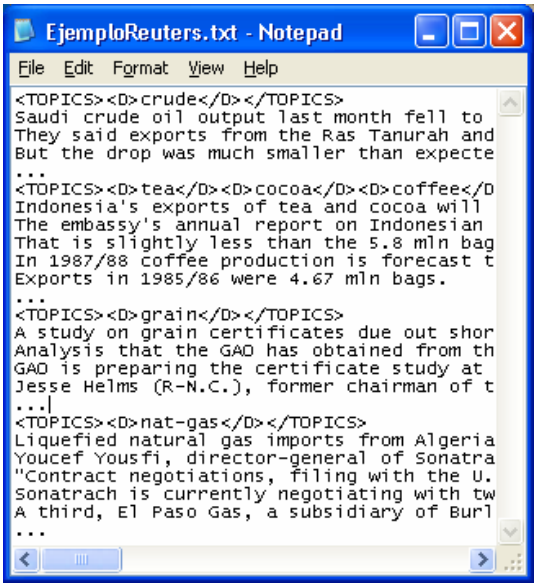

Figura 3. Ejemplo fragmento de corpus de Agencia Reuters de noticias

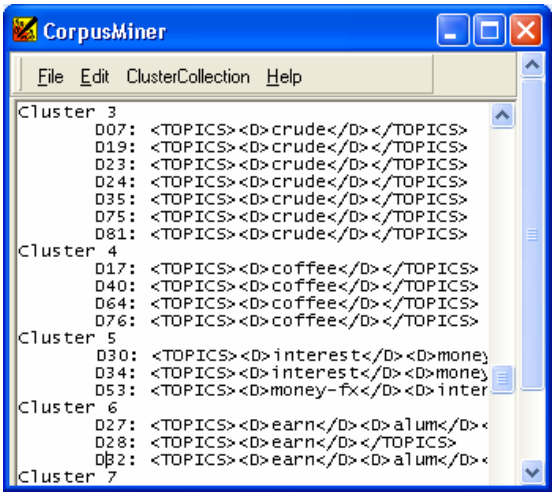

Figura 4. Ejemplo fragmento de resultado del agrupamiento de un corpus de Agencia Reuters de noticias según CorpusMiner

Obsérvese en la figura 3 un fragmento de los corpus utilizados en la experimentación y en la figura 4 un fragmento de cómo CorpusMiner presenta el resultado del agrupamiento.

5.3 Resultados experimentales del agrupamiento para el caso de estudio definido

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. Las métricas utilizadas en la evaluación fueron las descritas en la sección 5.1.

A partir del caso de estudio descrito se construyó la tabla 1 para realizar el análisis estadístico del agrupamiento.

Corpus	**SKWIC**	**Fuzzy SKWIC**	***Extended Star* – SKWIC**	***Extended Star* – Fuzzy SKWIC**
	1..5	1..5	1..5	1..5
C_1				
…				
C_{20}				

Tabla 1. Organización de los datos para realizar análisis estadístico

En la tabla 1 los C_i, con $1 \leq i \leq 20$ representan los 20 corpus formados a partir de la agencia Reuters de noticias. Por cada algoritmo a comparar se registraron los valores de las métricas consideradas en la evaluación en el siguiente orden:

1. *Precision*
2. *Recall*
3. F-*Measure*
4. Entropía
5. Media O*verall Similarity*

Las métricas son calculadas a partir de los resultados del agrupamiento al aplicar cada algoritmo o concatenación de éstos a los 20 corpus textuales.

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.

El análisis se ha focalizado en verificar si el método concatenado *Extended Star* – SKWIC logra superar al algoritmo SKWIC y si el método concatenado *Extended Star – Fuzzy* SKWIC logra superar al algoritmo *Fuzzy* SKWIC.

Métricas	**SKWIC v.s. *Extended Star* – SKWIC**	**Fuzzy SKWIC v.s. *Extended Star* – Fuzzy SKWIC**
1	.025	.031
2	.723	.214
3	.002	.045
4	.023	.052
5	.160	.232

Tabla 2. Significación a partir de pruebas pareadas no paramétricas de Wilcoxon

Como se observa en la tabla 2, respecto a las medidas *Precision* 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 en la tabla 2 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.

A continuación 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. Obsérvese en la tabla 3 los valores medios de estas métricas para los 20 corpus evaluados en cada caso. 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 clusters más compactos, resultado también deseado.

Métricas	**SKWIC**	***Extended Star* - SKWIC**
Precision	.894	.943
F-*Measure*	.849	.888
Entropía	.346	.313

Tabla 3. Valores medios de las métricas que arrojaron diferencias significativas entre los algoritmos SKWIC y *Extended Star* – SKWIC

Este análisis se puede ampliar al observar en la tabla 4 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. Obsérvese que 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 tiene 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 los valores de Entropía 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. Obsérvese en la tabla 5 los valores medios de *Precision* y F-*Measure* que fueron las medidas que mostraron diferencias significativas entre los algoritmos borrosos concatenados o no, y los valores medios de Entropía que es una medida que reflejó diferencias medianamente significativas.

Rangos

		N
Precision ES-SKWIC - Precision SKWIC	Rangos neg	4[a]
	Rangos pos	16[b]
	Empates	0[c]
	Total	20
F-Measure ES-SKWIC - F-Measure SKWIC	Rangos neg	3[d]
	Rangos pos	17[e]
	Empates	0[f]
	Total	20
Entropía ES-SKWIC - Entropía SKWIC	Rangos neg	16[g]
	Rangos pos	4[h]
	Empates	0[i]
	Total	20

a. Precision ES-SKWIC < Precision SKWIC
b. Precision ES-SKWIC > Precision SKWIC
c. Precision ES-SKWIC = Precision SKWIC
d. F-Measure ES-SKWIC < F-Measure SKWIC
e. F-Measure ES-SKWIC > F-Measure SKWIC
f. F-Measure ES-SKWIC = F-Measure SKWIC
g. Entropía ES-SKWIC < Entropía SKWIC
h. Entropía ES-SKWIC > Entropía SKWIC
i. Entropía ES-SKWIC = Entropía SKWIC

Tabla 4. Rangos para las métricas que reflejan diferencias significativas entre los algoritmos SKWIC y *Extended Star* – 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 clusters 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 clusters que no son aislados (i.e., que tiene más de un documento), éste 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.

Métricas	*Fuzzy* SKWIC	*Extended Star* – *Fuzzy* SKWIC
Precision	.873	.933
F-*Measure*	.907	.930
Entropía	.311	.288

Tabla 5. Valores medios de las métricas que muestran diferencias significativas entre los algoritmos *Fuzzy* SKWIC y *Extended Star* – *Fuzzy* SKWIC

¿Por qué seleccionar *Extended Star* como un método interior en la concatenación? Una razón es que el algoritmo *Extended Star* forma clusters muy homogéneos y compactos lo que hace que sea muy adecuado para inicializar los algoritmos SKWIC y *Fuzzy* SKWIC.

Métricas	*Extended Star* – SKWIC	*Extended Star* – *Fuzzy* SKWIC
Precision	.002	.001
Recall	.002	.005
F-*Measure*	.432	.306
Entropía	.002	.002
Overall Similarity	.014	.003

Tabla 6. Significación a partir de pruebas pareadas no paramétricas de Wilcoxon

Obsérvese en las tablas 6 y 7 los valores de significación y valores medios de las métricas respectivamente. Existen diferencias altamente significativas entre los algoritmos *Extended Star* y SKWIC, y *Extended Star* y *Fuzzy* SKWIC para las métricas *Precision*, *Recall*, Entropía, y *Overall Similarity*.

Las métricas *Precision*, Entropía y *Overall Similarity* que miden cuán compactos y homogéneos son los clusters reportan que el algoritmo *Extended Star* supera a SKWIC y *Fuzzy* SKWIC en los 20 corpus agrupados. Sin embargo, SKWIC y *Fuzzy* SKWIC reportan los mejores resultados para los 20 corpus teniendo en cuenta la métrica *Recall, ya que* el algoritmo *Extended Star* forma muchos clusters aislados y clusters pequeños que no cubren las clases. Esto demuestra por qué no existen diferencias significativas según F-*Measure*, con valores de significación de 0.432 y 0.306 respectivamente, porque esta métrica es una combinación de *Precision* y *Recall*. De ahí que haya sido conveniente utilizar los centros de clusters compactos de *Extended Star* para inicializar SKWIC y *Fuzzy* SKWIC, y como éstos tienen un buen cubrimiento de las clases, se logra obtener buenos resultados de F-*Measure*.

Métricas	*Extended Star*	SKWIC	*Fuzzy* SKWIC
Precision	.958	.894	.873
Entropía	.190	.311	.311
Recall	.666	.947	.998
Overall Similarity	.501	.221	.157

Tabla 7. Valores medios de las métricas que muestran diferencias entre los algoritmos *Extended Star*, SKWIC y *Fuzzy* SKWIC

El algoritmo *Extended Star* continúa siendo superior a *Extended Star* – SKWIC y a *Extended Star* – *Fuzzy* SKWIC respecto a *Precision*, Entropía y *Overall Similarity*. Esto es lógico, por las características de *Extended Star* al lograr clusters muy compactos y homogéneos. No obstante, el objetivo de la concatenación de métodos no es superar el algoritmo *Extended Star*, sino lograr mejorar los algoritmos SKWIC y *Fuzzy* SKWIC que son los que simultáneamente calculan la relevancia de las palabras en el proceso de agrupamiento, y por tanto, son de gran utilidad en procesamientos posteriores como la obtención de resúmenes extractos.

6. Conclusiones

El resultado de esta investigación refleja que existe una creciente base teórica conceptual sobre el agrupamiento de documentos, sin embargo, aún esta es un área donde es necesario continuar trabajando para lograr algoritmos que agrupen eficientemente los documentos y además retornen salidas importantes como palabras claves por cluster, necesarias en procesamientos textuales posteriores, sin necesidad de tener conocimiento del dominio y logrando converger rápidamente.

En el trabajo se propusieron los métodos concatenados para el agrupamiento *Extended Star* – SKWIC y *Extended Star – Fuzzy* SKWIC, siendo en ambos casos el algoritmo de entrada *Extended Star* y los de salida SKWIC o *Fuzzy* SKWIC si se desea aplicar técnicas duras y deterministas o borrosas.

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.

Agradecimientos

Agradecemos a Libernys Valdés por su arduo trabajo en el desarrollo de CorpusMiner, y a los revisores por sus valiosas sugerencias.

Referencias

[Arco05] Arco, L. Modelo para el agrupamiento de documentos afines y su ulterior resumen a través de la representación espacio vectorial de un corpus textual. Tesis de Maestría en Ciencia de la Computación. Universidad Central "Marta Abreu" de Las Villas. Cuba. (2005).

[Aslam et al. 98] Aslam, J. Pelekhov, K. Rus, D. Static and dynamic information organization with star clusters. Proceedings of the Conference of Information Knowledge Management, Baltimore, MD. (1998).

[Berry04] Berry, M. Survey of Text Mining. Clustering, Classification, and Retrieval. Springer-Verlag. ISBN 0-387-95563-1. (2004).

[Bezdek73] Bezdek, J.C. Fuzzy mathematics in pattern classification. Ph. D. Thesis, Applied Math. Center, Cornell University, Ithaca. (1973).

[Feldman et al. 96] Feldman, R. Hirsh, H. Exploiting background information in knowledge discovery from text. Journal of Intelligente Information Systems. (1996).

[Forman03] Forman, G. An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research 3. pp. 1289-1305. (2003).

[Frankes et al. 92] Frankes, W.B. Baeza-Yates, R. Information retrieval. Data structures & algorithms. Prentice Hall PTR. (1992).

[Frigui et al. 00] Frigui, H. Nasraoui, O. Simultaneos clustering and attribute discrimination. Proceedings of the IEEE Conference on Fuzzy Systems, pp. 158-163. (2000).

[Fung02] Fung, B. Hierarchical document clustering using frequent itemsets. Master of Science in the School of Computing Science. Simon Fraser University. (2002).

[Gil-García et al. 03] Gil-García, R. Badía, J.M. Pons, A. Extended star clustering algorithm. Proceedings of CIARP. (2003).

[Höppner et al. 99] Höppner, F. Klawonn, F. Rudolf, K. Runkler, T. Fuzzy cluster analysis. Methods for classification, data Analysis and image recognition. John Wiley & Sons. (1999).

[Jang97] Jang, J.S. Neuro fuzzy and soft computing. Prentice Hall. (1997).

[Kogan01] Kogan, J. Means clustering for text data. Proceedings of the First SIAM International Conference on Data Mining, pp. 47-57. (2001).

[Larocca et al. 00] Larocca, J. Santos, A. Kaestner, C. Freitas, A. Generating text summaries through the relative importance of topics. Proceedings of 7th Iberoamerican Conference on Artifificial Inteligence, pp. 300-309. (2000).

[Manning et al. 00] Manning, C. Shütze, H. Foundations of statistical natural language Processing. MIT Press, Cambridge, MA. (2000).

[Nürnberger et al. 01] Nürnberger, A. Klose, A. Kruse, R. Clustering of document collection to support interactive text exploration. Proceedings of 25th Annuals Conference of the Gesellschaft für Klassification. pp. 291-299. (2001).

[McQueen67] McQueen, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. pp. 182-297. (1967).

[Runkler et al. 01] Runkler, T.A. Bezdek, J.C. Relational clustering for the analysis of internet newsgroups. Proceedings of the 25th Annuals Conference of the Gesellschaft für Klassifikation, pp. 291-299. (2001).

[Salton et al. 75] Salton, G. Wong, A. Yang, C.S. A vector space model for automatic text retrieval. Communications of the ACM. 18(11). pp. 613-620. (1975).

[Stein et al. 00] Stein, G. Bagga, A. Wise, G.B. Multi-document summarization: methodologies and evaluations. Proceedings of TALN. (2000).