ISSN 1870-4069

Exploración sobre el máximo desempeño en la selección no supervisada de términos para agrupamiento de textos

Héctor Jiménez-Salazar

Universidad Autónoma Metropolitana, Departamento de Tecnologías de la Información, Unidad Cuajimalpa, México

hgimenezs@gmail.com

Resumen. El agrupamiento de textos es uno reto importante por la diversidad de aplicaciones que se derivan de la solución de dicha tarea. Un elemento indispensable en el agrupamiento es la selección de términos para representar lo mejor posible los textos. Aunque hay muchos métodos orientados a extraer términos de documentos para llevar a cabo categorización de textos, son pocos los que enfrentan la tarea de agrupamiento por la dificultad que se presenta al no contar con la clase de cada uno de los documentos. En este trabajo se propone un nuevo método que extrae los términos para representar los textos y, al ser agrupados, se obtiene el desempeño máximo en una cantidad notable de casos. Las pruebas se llevaron a cabo con un conjunto de varias decenas de colecciones de textos cortos (tuits), lo cual permite observar el comportamiento del método. El planteamiento que subyace al método está basado en el ascenso máximo de la similitud de los documentos y en las propiedades de unificación y diversificación de los términos expuestas por G. Zipf.

Palabras clave: Selección no supervisada de términos, agrupamiento de textos, tuits.

Exploration on the Maximum Performance of Unsupervised Term Selection for Text Clustering

Abstract. Text clustering is a major challenge for a diversity of applications derived from the solution of this task. An indispensable element in the text clustering is the selection of terms in order to get the best representation of texts. Although there are many methods designed to extract terms from documents to carry out categorization of texts, there are few methods that face the task of clustering, due to the difficulty presented by not having the class of each document. In this paper a new method is proposed that extracts the terms to represent the texts and, being grouped, it obtains the maximum performance in a high number of cases. Tests were conducted with a set of several tens of collections composed by short texts (tweets), which allows to know the behavior

of the method. The approach underlying of the method is based on maximum rise of the similarity of documents and properties of unification and diversification of the terms given by G. Zipf.

Keywords. Unsupervised term selection, text clustering, tweets.

1. Introducción

El agrupamiento de textos es uno reto importante no solo por la diversidad de aplicaciones que se derivan de la solución de dicha tarea, como la segmentación de textos [6], la inducción de sentidos [1], y la visualización [7], entre otras, sino además por la alta demanda que hay en la actualidad debido a los volúmenes crecientes de texto en línea que requieren sistematización para el aprovechamiento de su contenido [4].

Dada una colección de textos, el problema de agrupamiento establece como meta reunir en grupos a los textos que satisfagan mayor similitud entre los del mismo grupo y menor similitud entre textos de grupos diferentes. Para resolver este problema se representan los textos mediante algún criterio (bolsa de palabras, vectorial, distribucional, etc.) y se aplica un método de agrupamiento (K-means, k-NN, etc.). La evaluación de la efectividad del agrupamiento puede hacerse con una colección de textos clasificados manualmente o *gold standard*, de tal manera que los grupos obtenidos puedan compararse con el *gold standard*.

Uno de los factores que influye de manera crucial en el resultado del agrupamiento es la representación de los textos; por ejemplo, es conocido que si se incluyen todos los términos de los textos de la colección, muchos de ellos resultarán "ruidosos": sesgarán la construcción de los grupos, incluyendo o excluyendo textos en forma incorrecta. Así, se concibe representar los textos utilizando los términos que logran mejor desempeño en la tarea de agrupamiento. Éste es un problema de optimización combinatoria que por su alta dimensionalidad (tamaño del vocabulario) se ha enfrentado con un enfoque intuitivo a través de dos pasos: definir un criterio de importancia de los términos y, con otro criterio, elegir una parte de los términos que mejor representación hacen de los textos a agrupar.

Por tanto, debe adoptarse un criterio que asigne la importancia a cada término, usando un valor numérico y, finalmente, tomar una parte de los términos más importantes. Los métodos de selección de términos, entonces, deberán proporcionar un ordenamiento de los términos según su importancia (*ranking*) y un criterio de selección, es decir, la cantidad de términos que se tomarán de la lista ordenada para representar solamente con ellos todos los textos de la colección.

Es importante señalar que la tarea de agrupamiento textual difiere de la tarea de categorización textual. En el primer caso las aplicaciones de los métodos se limitan a un conjunto de textos sin ninguna información sobre la clase de ellos y justamente se trata de definir la clase de cada uno de los textos, mientras que en el segundo caso se cuenta con textos y la clase a la cual pertenecen, a partir de lo cual se espera determinar a cual de las clases definidas pertenece uno o varios nuevos textos.

Al resolver el problema de agrupamiento, no se tiene ninguna información previa. Puede ser, incluso, que tengamos una colección de textos de la cual tenemos clasificada una parte y se desee agrupar el resto, sin considerar las clases conocidas; esto es, no se podrá tomar en cuenta la importancia de los términos observada en la parte clasificada de los textos.

Por ejemplo, en la figura 1 aparecen dos líneas, éstas representan la efectividad del agrupamiento de dos colecciónes del mismo dominio (autos). Cada curva se define por puntos que consideran, en el eje horizontal, un múltiplo del 5% del total de términos y, en el vertical, la efectividad del agrupamiento. Lo que se observa es que si tomamos como referencia el porcentaje para el cual se obtuvo máximo desempeño en la colección 13 para usarlo en la selección de términos de la colección 9, tendremos un gran fracaso. En conclusión, en el agrupamiento no basta que un método que determina la importancia de los términos sea superior a otros, además deberá contarse con una forma de determinar cuántos términos elegir.

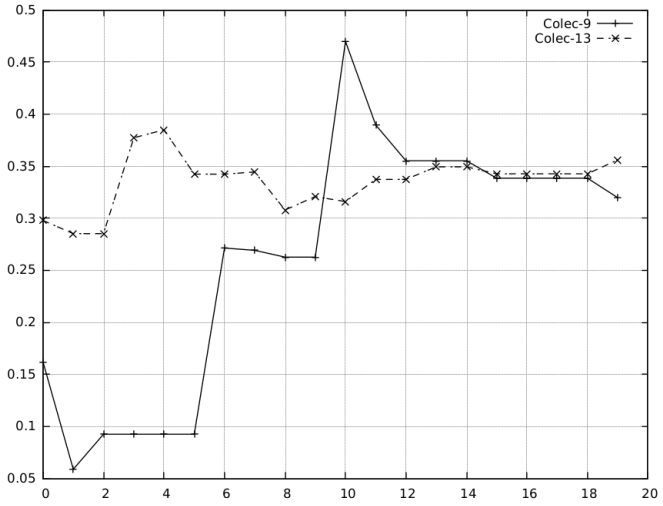

Fig. 1. Desempeño de un método de selección en dos colecciones.

Este trabajo expone algunos experimentos realizados sobre agrupamiento de textos en un conjunto de colecciones con el fin de analizar las regularidades de un método de selección de términos y otro de pesado de términos (que permite hacer el *ranking*). Solamente se usa un método de agrupamiento: *K-star* (una variante de *K-Nearest Neighbor*) [10], el cual determina en forma heurística el número de grupos. El presente trabajo aporta resultados orientados hacia el conocimiento sobre la representación de textos para mejorar la tarea de agrupamiento a través de:

1. Un método para pesado de términos.

2. Un criterio para seleccionar los términos pesados.

lo anterior sin considerar ninguna información previa: umbrales de selección para otras colecciones, información sobre indicadores en una parte clasificada de la colección, etc.

En este trabajo se utilizó una parte de la colección de RepLab-2013, competencia internacional sobre monitoreo de reputación (organizaciones, personalidades, etc.) y los detalles sobre ella se exponen en la sección 2. Si bien se propone y explora un método de selección, se utilizan como contraste otros métodos de pesado de términos, todo esto se describe en la sección 3. La sección 4, por su parte, contiene el procedimiento seguido para realizar el experimento y los resultados obtenidos. Las conclusiones de este trabajo aparecen en la última sección.

2. Descripción de las colecciones de prueba

En los experimentos se utilizó el conjunto de datos de la competencia RepLab 2013 [2]. Para las tareas previstas en RepLab-2013 este conjunto de datos fue anotado manualmente con (a) pertenencia o no a la entidad, (b) tema, (c) polaridad, y (d) grado de prioridad. De estas anotaciones solamente interesa el tema, y se usa con el fin de medir el desempeño de los métodos propuestos.

Este conjunto de datos está compuesto por cuatro dominios: autos, bancos, universidades, y música. En cada uno de los dominios hay entidades, y por cada entidad varios temas. Finalmente cada tema contiene tuits. En total son 61 colecciones de texto. Además, en todas las colecciones aparecen tuits que no tienen ninguna relación con el dominio o tema.

Como se sabe, cada tuit contiene información que puede ser de importancia para tareas diversas, como el lenguaje usado, *hashtag*, etc. En nuestros experimentos, de esta información se toma en cuenta solamente el lenguaje para dividir en dos las colecciones: inglés y español.

Con la anterior estructura se tienen dos conjuntos de datos, de entrenamiento y prueba. Nuestros experimentos trabajan únicamente con la colección de entrenamiento en español. Los tuits fueron preprocesados eliminando las palabras cerradas, asimismo, los enlaces dentro de los tuits y nombres de usuario fueron removidos, pero la conversación derivada de cada tuit se conservó como parte del tuit inicial.

Después de preprocesar las colecciones se obtuvo el material al cual se aplicaron los métodos que se presentan en la sección 3.. La tabla 1 muestra los siguientes valores: en la primera columna, el número total de colecciones; en la segunda, el número promedio de clases por colección; en la tercera, el número promedio de textos por colección; y en la última, el número promedio de palabras que ocurren por colección.

Tabla 1. Composición de las colecciones de prueba utilizadas.

Colecciones	Prom.clases	Prom.textos	Prom.palabras
54	11.20	117.27	1288.07

3. Métodos de selección utilizados

En esta sección se presenta la base sobre la que se apoya el método de ponderación de términos propuesto. Asimismo, se describen los métodos de pesado de términos que fueron analizados. El método que a continuación se presenta fue utilizado en la competencia RepLab13 [2,9] (UAMCLyR-7); aunque solamente en el presente trabajo aparece la descripción detallada del método y nuevos experimentos que muestran su grado de efectividad.

3.1. Método propuesto

El método surge de la relación que guarda el promedio de la similitud entre todas las parejas de documentos de una colección con la entropía [3]. Esta propiedad puede enunciarse como: a mayor similitud menor entropía; es decir, los términos que componen los documentos son más informativos, cuando la entropía es menor. Relacionado con lo anterior se consideran dos conceptos de G. Zipf, *diversificación* y *unificación* [13].

Para trabajar con estas ideas considérese una colección de documentos, $\mathcal{C} = \{d_1, \ldots, d_n\}$. Los términos de $\mathcal{C}$ manifiestan su propiedad unificadora a través de una alta similitud entre los documentos, y su propiedad diversificadora mediante baja similitud entre los elementos de $\mathcal{C}$. Un ejemplo de términos diversificadores son los *hapax*, y términos muy frecuentes entre documentos serán unificadores.

Se representa cada término t de $\mathcal{C}$ por el conjunto de clases que contienen a t, $\bar{t}$. Un término es diversificador si $\#\bar{t}$ es bajo, y es unificador si $\#\bar{t}$ es alto. Ciertamente ambos tipos de términos son necesario en la tarea de agrupamiento.

El enfoque seguido está basado en la cuantificación de la propiedad unificadora de los términos, a través de la fórmula:

$$U(t_i) = \frac{1}{r} \sum_{j \neq i} sim(\bar{t_i}, \bar{t_j}),$$

donde $r = \#\{t_j | sim(\bar{t_i}, \bar{t_j}) \neq 0)\}$, y sim es una medida de similitud. También es útil considerar:

$$S(\mathcal{C}) = \frac{2}{n(n-1)} \sum_{i \neq j} sim(d_i, d_j).$$

$S(\mathcal{C})$ da un valor global sobre la unificación a partir de los términos que componen los documentos de $\mathcal{C}$. Faltaría explicar cómo elegir un conjunto de términos que tenga una proporción adecuada de ambos términos, unificadores y diversificadores.

Como un primer paso de la representación de documentos podemos incluir la mayor parte de los términos diversificadores, lo cual se consigue eligiendo términos con los valores más bajos de U. Llamemos V a este conjunto, y el $p\%$ de los términos con los valores más bajos de U será denotado por V_p. Dado p, se calcula $S(\mathcal{C}_p)$, donde $\mathcal{C}_p$ tiene los mismos documentos que $\mathcal{C}$ pero representados unicamente por sus términos que pertenecen a V_p. Cuando p crece, los valores de U también, pero para cierto porcentaje q, $\mathcal{C}_q$ se satura con términos unificadores provocando un gran descenso en el valor $S(\mathcal{C}_q)$. Este descenso se interpreta como un indicador de una selección balanceada con ambos tipos de términos.

En las colecciones de prueba se observó correlación entre el valor máximo de F (usando precisión y exhaustividad) y el descenso abrupto de $S(\mathcal{C})$. Este descenso está asociado con la similitud máxima de los documentos. En suma, el método sigue dos pasos:

1. Determinar un conjunto balanceado de términos usando U y S:
   (a) Calcular $U(t)$ para todos los términos de $\mathcal{C}$ y ordenar en forma creciente: $T_U = [U(t_1), \ldots, U(t_n)]$.
   (b) Dividir T_U en m partes, para proveer m conjuntos de términos: $V_i, (1 \leq i \leq m)$ representa las primeras i partes de términos (en el experimento se usó $m = 10$).
   (c) Calcular $S(\mathcal{C}_i)$, correspondiente a cada conjunto de selección V_i y determinar el punto de máximo descenso; j.
2. Aplicar el algortimo de agrupamiento a $\mathcal{C}_j$.

Notemos que en el anterior procedimiento puede usarse cualquier otro método de pesado. El que se ha presentado, definido por U, será llamado DU en lo sucesivo.

Por otro lado, las clases usadas para representar cada término ($\bar{t}$), fueron determinadas usando el resultado de la aplicación del algoritmo *K-Star* sin seleccionar términos de los documentos del mismo corpus de trabajo.

3.2. Otros métodos de selección

Adicionalmente al método DU, se utilizaron tres métodos de selección de términos: TA (*term average*), DF (*document frequency*), y TC (*term contribution*).

En primer lugar se define el peso *frecuencia en documentos*, DF, como $DF(t_i) = |\{d_j | t_i \text{ ocurre en } d_j\}|$ [12]. DF se considera en el el cálculo de un peso muy utilizado en recuperación de información [8] y su normalización:

$$tfidf_{ij} = f_{ij} \cdot \log\left(\frac{2 * n}{DF(t_i)}\right), \qquad tfin_{ij} = \frac{tfidf_{ij}}{\sqrt{\sum_{k=1}^{n} tfidf_{ik}^2}},$$

donde f_{ij} es la frecuencia del término t_i en el documento d_j. Con lo anterior se define el peso *contribución del término* t_i [5]:

$$TC(t_i) = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} tfin_{ij} \cdot tfin_{ik}.$$

Y el peso *promedio del término* basado en $tfidf_{ij}$, TA [11]:

$$TA(t_i) = \frac{1}{n} \sum_{j=1}^{n} tfidf_{ij}.$$

A continuación se describe la forma en que se midió el desempeño de los métodos de selección utilizados en el agrupamiento. Dada una colección $\mathcal{C}$ con m clases $C_1, \ldots, C_m$ y un agrupamiento $\mathcal{G} = \{G_1, \ldots, G_s\}$, con base en las siguientes medidas $P_{ij} = \frac{|Ci \cap Gj|}{|Ci|}$, $R_{ij} = \frac{|Ci \cap Gj|}{|Gi|}$, $1 \leq i \leq m$ y $1 \leq j \leq s$, se define la medida $F_{ij} = \frac{2 \cdot Pij \cdot Rij}{Pij + Rij}$ de la clase C_i con respecto al grupo G_j. Con ello se calcula una medida global de todo el agrupamiento $\mathcal{G}$ como:

$$F = \sum_{i=1}^{m} \frac{|C_i|}{|\mathcal{C}|} F_i,$$

donde $F_i = \max\{F_{ij}\}_{j=1}^{m}$. Con esta medida se realizó la evaluación de los agrupamientos realizados.

4. Experimentos y resultados

De acuerdo con lo expuesto en la sec. 3., el primer paso fue calcular los pesos por cada uno de los métodos utilizados para ordenar términos por su importancia; TA, TC, DF y DU. Las listas de términos, ordenados de mayor a menor según su peso (excepto DU, que es de menor a mayor), se utilizan para tomar porcentajes crecientes, desde 5%, e incrementando 5% hasta tener la totalidad de términos. Representando cada documento con únicamente los términos seleccionados se llevó a cabo el agrupamiento con *K-star*. El método de agrupamiento calcula la similitud promedio entre todos los textos y utiliza este valor como criterio para decidir si dos textos pertenecen a la misma clase: se toma como la similitud mínima que deben satisfacer textos de un mismo grupo. Como se verá, el ascenso de la similitud promedio entre los documentos es un indicador de una buena selección en ciertas condiciones.

Ya que se cuenta con 54 colecciones de textos, un primer experimento fue conocer el comportamiento de los métodos de pesado de términos. Como ejemplo en la figura 2 se muestran los resultados de la selección obtenida por el criterio de similitud en una muestra de seis colecciones. Cada punto del eje horizontal corresponde a una colección, y la altura al desempeño obtenido por uno de los métodos.

En esta gráfica se observa la variabilidad de los valores de F para varias colecciones y no parece haber regularidad de un método a través de las colecciones. Por ejemplo, para las primeras dos colecciones el método TC resulta mejor, no así en las demás colecciones. Algo semejante sucede si exploramos la totalidad de las colecciones.

Por otro lado, en la figura 3 se observa el máximo desempeño que obtuvo cada método en las seis colecciones. En este caso sí se tiene que el método TA

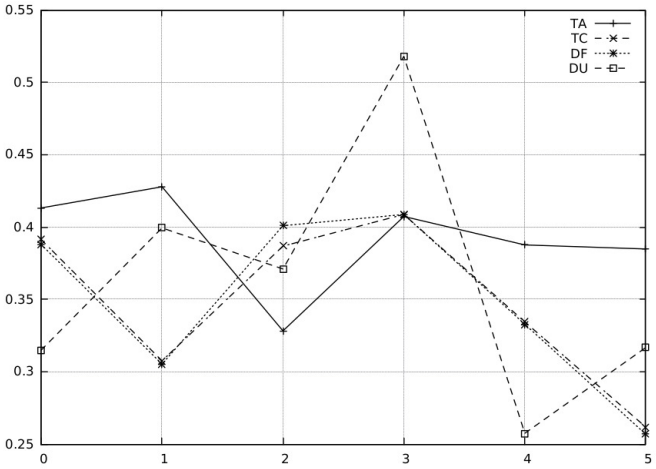

Fig. 2. Valor de F obtenido con el criterio de similitud máxima sobre una muestra de seis colecciones.

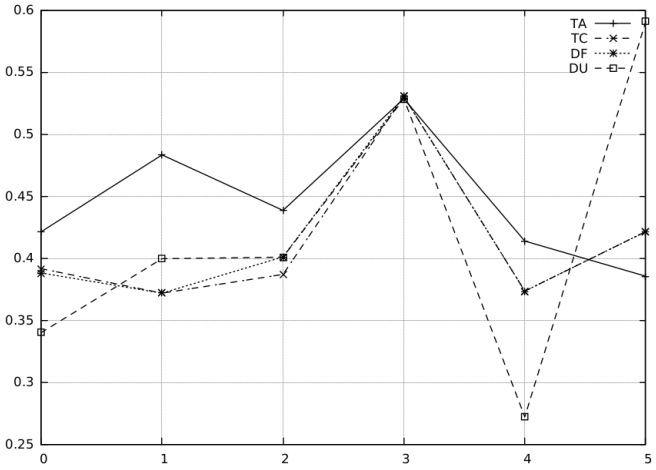

Fig. 3. Máximo valor de F que puede obtenerse con cada método de selección sobre una muestra de seis coleciones.

supera en la mayoría de casos. Sin embargo, no tenemos acceso a esta información pues requiere del conocimiento de la clase de cada texto. Se ha comentado que el porcentaje de términos de un ordenamiento, dado por un método de ponderación, que obtuvo buen desempeño en una colección no es útil en otras, aún cuando se observe un mismo porcentaje conveniente para varias colecciones, no puede este hecho asegurar que funcionará tal porcentaje en una nueva colección.

Para discernir sobre el mejor desempeño de los métodos de selección se

diseño un experimento para conocer la relación entre el desempeño obtenido y el máximo posible que podía alcanzar cada método: qué tan cerca están estos dos valores.

El experimento consistió en efectuar el agrupamiento de cada una de las colecciones utilizado los métodos de selección TA, TC, DF y DU. Además se aplicó el criterio de similitud máxima para determinar la selección que presumiblemente obtendría el mejor desempeño.

Enseguida se exponen los pasos realizados para encontrar los valores que permiten comparar el desempeño de los métodos utilizando el criterio de similitud máxima. Denotaremos un método de selección por M.

1. Para cada una de las colecciones, $\mathcal{C}_i$ $(1 \leq i \leq 54)$:
   (a) Usando cada selección de términos S_{Mij} $(1 \leq j \leq 20)$, proporcionada por el método M, se representaron los textos de $\mathcal{C}_i$ y se realizó el agrupamiento de $\mathcal{C}_{ij}$.
   (b) En cada agrupamiento se obtuvo el desempeño: F_{Mij}.
   (c) De los valores de desempeño se obtuvo el valor promedio y máximo: $\bar{F}_{Mi}$, $F_{Mi,max}$.

En totalidad se tienen 54 parejas $\bar{F}_{Mi}$, $F_{Mi,max}$ para cada método. Con ellas se llevó a cabo una prueba de hipótesis para conocer en qué medida el método M obtenía un valor cercano al máximo. La prueba entonces utliza una muestra $(n = 54)$ en donde se supone que los valores de $\bar{F}_{Mi}$ y $F_{Mi,max}$ se distribuyen normalmente. Se llevó a cabo la prueba de diferencia de medias:

$$H_0 : \bar{F}_M = \bar{F}_{M,max}$$
$$H_1 : \bar{F}_M \neq \bar{F}_{M,max}$$

donde $\bar{F}_M$ es la media de los valores de F obtenidos con el criterio de similitud máxima para el método M, y $\bar{F}_{M,max}$ es el promedio de los valores máximos de F obtenidos para el método M. La tabla 2 muestra los resultados de las pruebas realizadas con un nivel de significancia del 95%. Las columnas, de izquierda a derecha, corresponden a: el método (M), el promedio de los valores F para M, el promedio de los máximos valores de F que obtuvo M, la diferencia de los anteriores valores, el valor crítico de la normalización de la distribución de la diferencia de valores medios, y la conclusión de la prueba de hipótesis, respectivamente.

Como se aprecia en la tabla 2, el método DU es el que mejor se aproxima al máximo que se puede obtener utlizando el criterio de similitud máxima. Notamos también que los otros métodos, aunque no siguen el criterio de máxima similitud, pueden obtener desempeños altos, lo cual se aprecia en la columna $\bar{F}_{M,max}$. Sin embargo, se ha dicho que no se tiene un acceso seguro a dichos valores en forma no supervisada.

5. Conclusiones

Se ha presentado un nuevo método para pesado de términos, DU, que funciona en combinación con el criterio de ascenso máximo de similitud (la similitud

Tabla 2. Resultados de las pruebas de diferencia entre $\bar{F}_M$ y $\bar{F}_{M,max}$.

M	$\bar{F}_M$	$\bar{F}_{M,max}$	DIFER	V.crítico	Acepta
TA	0.460015	0.556161	0.0961463	0.0724493	H_1
TC	0.427633	0.508341	0.0807074	0.0689402	H_1
DF	0.448352	0.533959	0.0856074	0.0756838	H_1
DU	0.473756	0.498837	0.0250815	0.0846659	H_0

promedio entre documentos que usan una selección de términos). La aplicación del método de selección (pesos y criterio de selección) a 54 colecciones de tuits obtuvo una efectividad muy cercana al máximo posible obtenido por este método. Al utilizar otros métodos de pesado de términos con el mismo criterio de selección se obtuvo una diferencia significativa con respecto a la máxima efectividad posible. Es importante destacar que los otros métodos obtienen un $\bar{F}_{max}$ mayor que el de DU. Esto sugiere, por ejemplo, que si pudiera adaptarse el criterio a estos métodos, su efectividad crecería.

En suma, los métodos TA, TC, y DF son menos compatibles con el criterio de ascenso máximo de similitud que el método DU. Aún cuando la afirmación anterior es estadísticamente válida para el conjunto de colecciones empleada, quizá ello no pueda generalizarse para colecciones semejantes. Ciertamente, habría un sustento para usar el método propuesto en agrupamiento de tuits, pero también es necesario conocer el alcance del método, a través de la realización de más pruebas tanto para textos que no sean tuits, como para textos de mayor tamaño.

Agradecimientos. El autor desea expresar su agradecimiento al Departamento de Tecnologías de la Información de la UAM Cuajimalpa por el apoyo parcial recibido al presente trabajo.

Referencias

1. Agirre, E., Soroa A.: Semeval-2007 task 02: evaluating word sense induction and discrimination systems. ACL (2007)
2. Amigó, E.; Carrillo de Albornoz, J.; Chugur, I.; Corujo, A.; Gonzalo, J.; Martín, T.; Meij, E.; de Rijke, M., Spina, D.: Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems. CLEF 2013, LNCS 8138, pp. 333–352 (2013)
3. Dash, M., Liu, H.: Feature Selection for Clustering. PAKDD 2000 (2000)
4. Jain, A.; Murty, M., Flynn, P.: Data Clustering: A Review. ACM Computing Surveys 31 (3) (1999)
5. Liu, T.; Lui, S.; Chen, Z., Ma, W.: An Evaluation of Feature Selection for Text Clustering. Proc. of the 20th Int. Conf. on Machine Learning (2003)
6. Lu, Q.; Conrad, J.; Al-Kofahi, K., Keenan, W.: Legal document clustering with built-in topic segmentation. Proc. of the 20th ACM Int. Conf. on Information and Knowledge Management, ACM (2011)

7. Metsalu, T., Vilo, J.: ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Research (2015)
8. Salton, G., Buckley, C.: Term weighting approaches in automatic text retrieval. Information Processing and Management, 24(5):513–523, issue 5 (1988)
9. Sánchez-Sánchez, C.; Jiménez-Salazar, H., Luna-Ramírez, W.: UAMCLyR at Replab2013: Monitoring Task. Notebook for RepLab at CLEF (2013)
10. Shin, K., Han, S.: Fast clustering algorithm for information organization. Proc. of the CICLing 2003 Conf. Volume 2588 of LNCS Springer-Verlag (2003)
11. Tang, B.; Shepherd, M.; Milos, E., Heywood, M.: Comparing and combining dimension reduction techniques for efficient text clustering. Int. Workshop on Feature Selection for Data Mining (2005)
12. Yang, Y., Pedersen, J. A.: Comparative Study of Feature Selection in Text Categorization. Proc. of SIGIR (1995)
13. Zipf, G.: Human behavior and the principle of least effort. Cambridge, (Mass.), Addison-Wesley (1949)