Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 383-390 recibido 29-04-2005; aceptado 01-06-2005

Uso del punto de transición en la selección de términos índice para agrupamiento de textos cortos*

Héctor Jiménez-Salazar & David Pinto
B. Universidad Autónoma de Puebla
Facultad de Cs. de la Computación
Ciudad Universitaria, Puebla, México
(hjimenez, dpinto)@cs.buap.mx

Paolo Rosso
Universidad Politécnica de Valencia
Sistemas Informáticos y Computación
Camino de Vera s/n, Valencia, España
prosso@dsic.upv.es

Resumen: La aplicación de los métodos de agrupamiento de textos considera la decisión crítica sobre cuáles términos serán usados para representar a cada instancia de la colección. Abordamos el agrupamiento de resúmenes de textos de un dominio específico. Así, el problema se complica porque hay pocos elementos que pueden usarse en la selección de términos, y se tratan textos muy parecidos. Nuestro enfoque ha sido utilizar los términos cuya frecuencia está en una vecindad del llamado punto de transición; la frecuencia que divide al vocabulario del texto en términos de alta y baja frecuencia. En las pruebas se utilizó una variante del método *vecino más cercano* sobre una colección de resúmenes del evento CICLing-2002. Evaluamos nuestros resultados con el estándar dado en el mismo evento y observamos un alto índice de desempeño con el método de selección de términos que proponemos.
Palabras clave: selección de términos, agrupamiento, resumen.

Abstract: Nowadays a wide variety of clustering methods exist. The critical decision of what keywords will be used in the representation of the collection is considered in those methods. In this paper we deal with the problem of clustering a set of short texts from an specific domain. Thus, the problem become to be more complex because of the small number of terms that can be used in term selection process; besides, all texts of the collection are very similar. Our approach uses a neighborhood of terms around of the named *transition point* (frequency that divides vocabulary in terms of high and low frequency). In our tests over a collection of abstracts from CICLing-2002, a modified method of Nearest Neighbour (NN) was used. We used a Gold Standard for the evaluation, observing a high performance for the proposed method.
Keywords: term selection, clustering, abstract.

1. Introducción

Los algoritmos de agrupamiento son utilizados principalmente con fines de clasificación. En la Recuperación de Información (RI) se tiene la finalidad de analizar grandes colecciones de documentos, subdividiéndolas en grupos que posean documentos similares. Un enfoque bien conocido para el Agrupamiento de Textos (AT), consiste en representar dichos textos por medio de un vector compuesto de un conjunto de términos (*palabras clave*) y, a partir de ellos, usar alguna función de similitud para generar grupos de textos similares (Manning y Schütze, 1999). Es necesario emplear métodos de selección de términos índice para conseguir la representación de los textos. En los algoritmos de agrupamiento supervisado, el conjunto de términos se construye a partir de un conjunto de instancias de entrenamiento que pertenecen al mismo domino. En caso de que el dominio sea desconocido, pueden emplearse técnicas de agrupamiento no supervisado. En este caso, el conjunto se construye directamente a partir de la misma colección de textos. Se sigue, por ejemplo, el modelo de espacio vectorial para la representación de textos; esto es, la asignación de un peso a cada uno de los términos del conjunto usado en la representación de los documentos, por ejemplo con la ponderación clásica $tf \cdot idf$

* Este trabajo fue parcialmente apoyado por BUAP-VIEP # III9-ING/G.

(Salton y Buckley, 1988); esto es, el peso de un término, para un determinado documento, está en función directa de su frecuencia de aparición en el documento (tf), y en función inversa del número de documentos que lo utilizan (idf).

El modelo de espacio vectorial no es solamente usado para agrupar documentos con un alto número de palabras, sino también para agrupar documentos cortos (alrededor de 50 a 100 palabras), por ejemplo, noticias, o información sobre publicidad, resúmenes de artículos científicos, patentes, etc. Los documentos de este tipo son los más interesantes, ya que la mayoría de bibliotecas digitales, y otros almacenes basados en el web que ponen a disposición documentos científicos y de información técnica, actualmente proporcionan acceso libre únicamente a los resúmenes y no al texto completo de los documentos. Sin embargo dichas colecciones de documentos imponen retos importantes. Si suponemos que la colección contiene textos pertenecientes a *dominios diferentes*, tales como deportes, política, etc; entonces éstos tendrán pocos o ningún término en común en sus vocabularios. En este caso, el tamaño de los documentos no es importante para los algoritmos de agrupamiento, ya que cualquier procedimiento de agrupamiento dividirá tales textos en grupos (considerados como dominios) bien definidos (Dhillon, Guan, y Kogan, 2002); los documentos serán mapeados a subespacios de términos completamente disjuntos dentro del espacio total de la colección. Cuando trabajamos con textos de *un solo dominio*, la situación es bastante diferente al caso anterior. Los grupos a identificar tienen una gran cantidad de términos en la intersección de sus vocabularios, y la diferencia entre estos grupos no solamente consiste del conjunto de términos índice sino también de su proporción.

En este trabajo abordamos el problema de agrupamiento de textos cortos, usando el concepto de punto de transición, una frecuencia intermedia del vocabulario de un texto. Teóricamente, alrededor de esta frecuencia se encuentran las frecuencias de palabras con mayor contenido semántico del texto. Por ello, se tiene confianza en que este enfoque permita elegir "mejores" términos que otros métodos de selección. Particularmente, la sencillez del método que será presentado, no supervisado y sin el apoyo de fuentes de conocimiento externas, ofrece ventajas sobre otros métodos, a la vez que reduce el número de términos.

En las secciones que restan en este documento, se presentan algunos trabajos relacionados con el tema de agrupamiento de textos y selección de términos, una breve fundamentación del punto de transición, los métodos de selección de términos que serán utilizados, la descripción del experimento llevado a cabo, y las conclusiones.

2. Trabajos relacionados

Existen muy pocos trabajos relacionados con el agrupamiento de textos cortos. Los trabajos presentados por Hynek *et al.* (Hynek y Rohlikm, 2000) y Zizka *et al.* (Zizka y Bourek, 2002) usan métodos supervisados que obtienen excelentes resultados, sin embargo requieren un conjunto de textos para el proceso de entrenamiento. En nuestro caso, como en el presentado por Mikhail *et al.* (Mikhail, Gelbukh, y Rosso, 2005), se usa un método no supervisado, de tal manera que se desconoce de antemano la cantidad de grupos a generar, así como las categorías de éstos. Makagonov *et al.* (Makagonov, Alexandrov, y Sboychakov, 2000) consideraron el problema de agrupamiento de resúmenes, sin embargo, en su trabajo, la colección de documentos usada contenía textos pertenecientes a dominios fácilmente distinguibles, y además el número de dominios era conocido de antemano.

Makagonov *et al.* (Makagonov, Alexandrov, y Gelbukh, 2004) usaron criterios fuertes para la selección de términos y una medida combinada de cercanía entre los documentos (medidas del coseno y polinomial). Estos criterios pueden dar mayor confiabilidad a los términos con frecuencias absolutas bajas de ocurrencia en los resúmenes; la medida combinada puede acercar los resultados a la opinión del experto. Sin embargo, ambas técnicas no son totalmente confiables ya que no son justificadas adecuadamente, además de haberse probado sobre situaciones en donde se conoce de antemano el número de grupos a generar.

Es importante remarcar que los métodos para encontrar términos índice pueden ser, también supervisados o no. Un trabajo en esta dirección es (Kerner, Gross, y Masa, 2005). En él se presenta un conjunto de métodos supervisados y no supervisados para encon-

trar *frases clave* de un texto. En este trabajo, como es de suponerse, los supervisados son mejores pero, además, se parte de los textos completos, y no solamente de los resúmenes.

Mikhail *et al.* (Mikhail, Gelbukh, y Rosso, 2005), proponen un método basado en fuentes de conocimiento externas (*corpus* general balanceado) para la selección de términos en documentos cortos y, posteriormente, usan algoritmos de agrupamiento no supervisados para generar grupos, particularmente, el algoritmo de agrupamiento MajorClust.

Liu *et al.* (Liu et al., 2003) evaluaron algunos métodos de selección de términos para agrupamiento de textos aplicado a una subcolección de Reuteres 21578. Señalan la dificultad de realizar una buena selección para el caso de los métodos de selección no supervisados, y proponen una técnica iterativa para elegir términos.

3. El punto de transición

El *punto de transición* (PT) es una consecuencia de las observaciones de George Kinsley Zipf, quién formuló la ley de frecuencias de palabras de un texto (Ley de Zipf), la cual establece que el producto del rango por la frecuencia de una palabra es constante (Zipf, 1949). Esta regularidad estadística proviene de la tensión entre dos fuerzas inherentes a los lenguajes naturales: *unificación* y *diversificación*. La primera conduce a emplear términos de índole general, mientras que la segunda al uso de términos específicos. Los términos ligados a la primera fuerza establecen nexos con el entorno del texto, y los de la segunda detallan su contenido. Esto sugiere que las palabras que caracterizan un texto no sean ni las más frecuentes ni las menos frecuentes, sino las que se encuentran en una frecuencia media de ocurrencia dentro del texto (Luhn, 1958).

Algunos autores, llevaron a cabo experimentos con las ideas anteriores; la indización automática de textos, y la identificación de palabras clave de un texto (Urbizagástegui, 1999). A partir de la ley de ocurrencia de palabras con baja frecuencia propuesta por Booth (Booth, 1967), fue posible derivar una fórmula para localizar la frecuencia que divide en dos al vocabulario de un texto: las palabras de baja, y alta frecuencia; justamente, el llamado punto de transición. La fórmula para calcular el PT es:

$$PT = \frac{\sqrt{1 + 8 \times I_1} - 1}{2}, \qquad (1)$$

donde I_1 representa el número de palabras con frecuencia 1. De acuerdo con la caracterización de las frecuencias medias (Booth, 1967), el PT puede localizarse, en el vocabulario de un texto, identificando la frecuencia más baja, de las altas, que no se repita. Este método es particularmente útil para textos cortos; en la obtención del extracto de un texto (Bueno, Pinto, y Jiménez-Salazar, 2005), y la identificación de las palabras clave de un texto (Pinto y Pérez, 2004).

Ha habido algunas aplicaciones que revelan la utilidad del PT. Específicamente, en el corte de la selección de términos por los métodos clásicos de selección (Moyotl y Jiménez, 2004), y la selección de términos para categorización de textos (Moyotl-Hernández y Jiménez-Salazar, 2005).

Debido a que un resumen reúne las características de cualquier texto, el problema de frecuencia baja de los términos, decisivo en la representación para procesamiento, puede atenuarse considerando que se cumplen las leyes derivadas de la de Zipf. En esencia, esta hipótesis es la que se pretende reforzar en el presente trabajo.

4. Elección de términos índice

En numerosas tareas de procesamiento de texto (CT, RI, y AT, entre otras) es necesario representar los textos usando los términos contenidos en ellos. Sin embargo, suele hacerse una reducción de estos términos, debido a la gran cantidad de términos que ocurren en una colección; además de que el empleo de todos los términos vicia el procedimiento, sea éste de clasificación, resumen, etc. Así, se usan variados métodos para elegir los términos que representarán a los textos; es decir los términos índice. La selección se hace con base en una puntuación que el método asigna a cada término: se toma un porcentaje del total de términos de los textos con la más alta puntuación.

Los métodos de selección pueden ser supervisados o no supervisados; esto es, los supervisados utilizan información acerca de los términos que tienen mayor capacidad para determinar una clase, según la colección de entrenamiento (Sebastiani, 2002). Dos de los métodos supervisados más efectivos son:

CHI, que mide la independencia entre la clase de un texto y un término contenido en el texto; e IG cuya puntuación representa la carencia de información que provee un término para predecir la clase del texto en el que ocurre. En este trabajo utilizaremos métodos no supervisados puesto que resulta más útil para el tipo de problema que se pretende resolver. Consideremos una colección de textos $D = \{T_1, \ldots, T_k\}$. Tres son los métodos que abordaremos:

Frecuencia entre documentos (DF). Asigna a cada término t el valor df_t, que es el número de textos de D en los que ocurre t. Se supone que los términos raros (baja frecuencia) difícilmente ocurrirán en otro texto y, por tanto, no tienen capacidad para predecir la clase de un texto.

Fuerza de enlace (TS). La puntuación que se da a un término t está definida por:

$$ts_t = \Pr(t \in T_i | t \in T_j),$$

donde $sim(T_i, T_j) > \beta$, y β es un umbral que debe ajustarse observando la matriz de similitudes entre los textos. Con base en su definción, puede decirse que un valor alto de ts_t significa que t contribuyó a que, al menos, dos documentos fueran más similares que el umbral β.

Punto de transición (PT). Los términos reciben un valor alto entre más cerca esté su frecuencia del PT. Una forma de hacerlo es calcular el inverso de la distancia entre la frecuencia del término y el PT:

$$idtp_t = \frac{1}{|PT - fr(t)| + 1},$$

donde $fr(t)$ es la frecuencia local, (en el texto, y no en la colección); esto es, los términos reciben una puntuación en cada texto.

DF es un método muy simple pero efectivo, por ejemplo, en categorización de textos (CT) compite con los clásicos supervisados CHI e IG.

También el método PT tiene un cálculo simple, y puede usarse de diversas formas. En especial para CT se ha visto mejor desempeño con PT_{df}, o PT global; esto es, se considera df_t, en lugar de la frecuencia local de los términos en cada texto de la colección.

Los métodos DF y PT están en la clase de complejidad lineal con respecto al número de términos de la colección.

El método TS (*Term Strength*) es muy dispendioso en su cálculo, pues requiere calcular la matriz de similitudes entre documentos; cuadrático en el número de textos. Pero se reportan resultados de AT cercanos a los métodos supervisados (Liu et al., 2003).

4.1. Enriquecimiento de términos índice

Es común enriquecer los términos índice, por ejemplo, incluyendo sus sinónimos. Esta idea se emplea en diversos contextos; por ejemplo, en RI se refiere a la expansión de consultas. La expansión de un término t añade términos relacionados con t. El fin es detectar textos relevantes a la consulta mediante los términos relacionados (Voorhees, 1994). La expansión habrá de apoyarse en una fuente que disponga los términos relacionados para cada término, un *thesaurus*. Aunque se dispone de ricas fuentes de información léxica, como WordNet, éstas son de carácter general y no abarcan dominios especializados.

Empleamos una técnica basada en la propuesta de Hindle (Hindle, 1990) que apoya los métodos de construcción de *thesauri*. Se dice que dos términos son *vecinos cercanos* cuando uno de ellos coocurre con el otro entre los de mayor frecuencia, y viceversa. En estos métodos es común utilizar una medida de asociación como la información mutua. Sin embargo, estas medidas se usan en textos grandes, y por ello nos limitamos a utilizar solamente la frecuencia de los términos.

A cada uno de los términos del vocabulario de una colección de textos se asocia una lista de términos que coocurren frecuentemente en las oraciones de la colección. Si consideramos que los términos índice representan a cada texto, entonces los términos asociados a los índice representarán de una manera más rica a los textos.

La lista de asociación para cada término índice se calcula como sigue. Para cada término, x, en el vocabulario de la colección su lista es:

$$L(x) = \{(y, k) | k = \#Ctx(x, y)\},$$

donde $Ctx(x, y)$ es el conjunto:

$$\{O | (\text{existe } T_j \in D) \wedge (O \in T_j) \wedge (x, y \in O)\},$$

i.e. $Ctx(x, y)$ es el conjunto de contextos (tomados como oraciones) en los que coocurren x e y para alguna oración de un texto de la colección.

Denotemos con T' los términos índice de T. Consideramos para cada término índice t ($t \in T'$) su lista de asociaciones, $L(x)$, ordenada por la segunda componente de sus miembros: $[(y_1, k_1), (y_2, k_2), \ldots]$, $k_i \geq k_{i+1}$ $(1 \leq i \leq \#L(x) - 1)$. En ésta se realiza un recorte de las parejas con $df_y = 1$, debido a que son términos que no contribuyen al agrupamiento así como los términos con frecuencias muy altas[1].

Sea $L'(x)$ la lista de palabras asociadas al término x después de la eliminación de términos con frecuencias extremales. La expansión del conjunto de términos índice T' es:

$$T'' = \bigcup_{x \in T'} \{y | (y, k) \in L'(x)\}.$$

T'' es, entonces, una manera alternativa de representar el texto T.

5. *Experimento*

Ya que nos propusimos averiguar el desempeño del PT en la selección de términos índice, elegimos dos métodos no supervisados para confrontar los resultados basados en el PT. Estos métodos fueron DF y TS.

5.1. Colección de prueba

Una manera de medir la calidad de los grupos generados es través del llamado *gold standard*, el cual consiste en el agrupamiento manual de textos completos. De esta manera podemos determinar la utilidad de los grupos generados.

Se utilizó una colección de prueba formada por 48 resúmenes de textos del dominio *Lingüística Computacional y Procesamiento de Textos*, correspondiente al evento *CiCLing 2002*. Los textos de la colección están repartidos en 4 clases:

1. Lingüística (semántica, sintaxis, morfología y *parsing*).
2. Ambigüedad (WSD, anáfora, etiquetamiento, y *spelling*).
3. Léxico (léxico, *corpus*, y generación de texto).
4. Procesamiento de texto (recuperación de información, resumen automático, y clasificación de textos).

Después de eliminar las palabras cerradas y aplicar un algoritmo de Porter para truncar el resto, el número total de términos de la colección fue 956, y cada texto contuvo 70.4 términos en promedio.

5.2. Método

Consideramos en nuestro experimento una colección de textos $D = \{T_1, \ldots, T_k\}$ con vocabulario V_D. Los textos se encuentran clasificados en m clases $C = \{C_1, \ldots, C_m\}$, formando una partición de D; $D = \cup_i C_i$ y $C_i \cap_{i \neq j} C_j = \emptyset$. Nuestro objetivo es obtener un agrupamiento de D; *i.e.* una partición, $G = \{G_1, \ldots, G_n\}$ lo "más parecida" a C. Así, es necesario conocer C, el *gold standard*, para evaluar los resultados.

Los términos índice de un texto se determinaron siguiendo los métodos presentados en la sección 4. Denotaremos con $Q_p(D)$ el conjunto formado con $p\,\%$ de términos índice determinados por el método Q sobre la colección D. Si nuestro método es DF, $DF_{10}(D)$ comprenderá el diez por ciento de los términos t con mayor valor df_t en la colección D. Cada texto será representado por sus términos índice filtrando su vocabulario con $Q_p(D)$; tomado T como conjunto de términos, sus índices son: $T' = T \cap Q_p(D)$.

Una vez representado cada texto por sus términos índice se aplica el algoritmo *star* (Shin y Han, 2003), el cual inicia construyendo la matriz de similitudes entre todas las instancias por agrupar. Utilizamos, en esta etapa, un umbral canónico definido como el promedio de las similitudes. En el siguiente paso se realiza una iteración, en tanto existan instancias que rebasen el umbral, se elige el par de textos con máxima similitud para formar el grupo en curso. Enseguida, se añaden al grupo en curso todas las instancias cuya similitud sea mayor que el umbral. Eliminadas las instancias agrupadas, se repite el proceso para formar otro grupo. En nuestro experimento usamos la función de similitud de Jaccard (Manning y Schütze, 1999).

[1] En este trabajo se descartaron las palabras cuya frecuencia fuera mayor o igual al PT_{df} calculado en la colección: *paper*, *present*, y *use*. Claramente vemos que estas palabras son iniciales comunes de los resúmenes. Además, los métodos de selección harán el trabajo de eliminar al menos un buen número de palabras de baja frecuencia.

5.3. Medidas de desempeño

Con el propósito de conocer cuál método, y en qué condiciones, realizaba un mejor agrupamiento, utilizamos la medida F (Rijsbergen, 1979), muy empleada en RI. Para un agrupamiento $\{G_1, \ldots, G_m\}$ y clases $\{C_1, \ldots, C_n\}$ se define, en primer lugar, F_{ij}, $1 \leq i \leq m$, $1 \leq j \leq n$, como:

$$F_{ij} = \frac{2 \cdot P_{ij} \cdot E_{ij}}{P_{ij} + E_{ij}}, \qquad (2)$$

donde P_{ij} (pureza), y E_{ij} (pureza inversa) se definen como

$$P_{ij} = \frac{\text{No. de textos del grupo } i \text{ en la clase } j}{\text{No. de textos en la clase } j},$$

y

$$E_{ij} = \frac{\text{No. de textos del grupo } i \text{ en la clase } j}{\text{No. de textos en el grupo } i}.$$

Con los valores F_{ij} se calcula el desempeño global del agrupamiento:

$$F = \sum_{1 \leq i \leq m} \frac{|G_i|}{|D|} \max_{1 \leq j \leq n} F_{ij}, \qquad (3)$$

5.4. Pruebas

Una prueba inicial fue necesaria para ajustar un factor que permite cambiar el umbral. Tomando 20 % de los términos de cada método de selección, y variando este factor entre 10^{-4} hasta 10, se eligió como mejor factor: 0.1; esto es, el umbral usado fue 0.1 veces el umbral canónico.

Se efectuó la prueba de elegir diferentes porcentajes de términos con cada método de selección: $PT_i(D)$, $DF_i(D)$ y $TS_i(D)$ $(i = 20, 30, 40, 50, 60)$, y la eficacia del agrupamiento se midió con F. Además de los valores F, en el cuadro 1 se presenta el número de grupos obtenidos con la selección de términos efectuada. Puede observarse que, en todos los porcentajes el método PT supera a los demás y alcanza su máximo con 40 % de los términos más cercanos al PT.

Al aplicar los métodos DF o TS, el porcentaje de términos se toma del vocabulario de toda la colección, mientras que el método PT elige los términos directamente de cada texto. Así, la cantidad de términos tomada por PT es diferente. En el cuadro 2 se muestra la cantidad de términos para cada uno de los porcentajes aludidos en el cuadro 1.

%	PT/#G	DF/#G	TS/#G
20	0,4267/13	0,4044/3	0,3716/13
30	0,4397/11	**0,4309**/4	0,4217/11
40	**0,6038**/7	0,4309/4	0,4353/12
50	0,5941/7	0,4309/4	**0,4701**/12
60	0,4948/3	0,4041/4	0,4071/10

Cuadro 1: Medidas F para diferentes porcentajes de términos ordenados por tres métodos de selección.

Claramente, la cantidad de términos seleccionados por PT es menor que los elegidos por los métodos DF y TS.

%	#Term.	#T. PT
20	191	133
30	286	181
40	382	263
50	478	274

Cuadro 2: Cantidad de términos elegidos por PT con respecto a porcentajes del vocabulario.

Una prueba más fue analizar el comportamiento de los métodos tratando de incluir términos relacionados con los términos de cada uno de los conjuntos elegidos por los métodos de selección. Con la misma estructura que el cuadro 1, en el cuadro 3 se presenta la medida F para esta prueba. Aunque con una base tan pobre para obtener términos relacionados, como fueron los mismos resúmenes, se observa, nuevamente, que el método basado en el punto de transición rebasa a los demás, excepto cuando se toma el 60 % de los términos.

%	PT/#G	DF/#G	TS/#G
20	**0,5805**/4	**0,4386**/4	0,3670/7
30	0,5805/4	0,4275/3	0,4590/5
40	0,5580/3	0,4309/3	0,3903/5
50	0,5580/3	0,3945/4	**0,5151**/3
60	0,4231/2	0,3945/4	0,4383/4

Cuadro 3: Medidas F para diferentes porcentajes de términos ordenados por tres métodos de selección usando lista de asociaciones.

El enriquecimiento provee una "suavización" de la representación; *i.e.*

al tomar más términos se sigue cumpliendo el objetivo de representar el texto. Sin embargo, cuando los términos no son adecuados se observa inestabilidad, variación no monotónica del índice F (ver F para DF y TS).

Se realizó, además, una evaluación con una clasificación standard diferente (tomada de la estructura que tiene la memoria del evento CICLing-2002) compuesta de dos clases: *Lingüística Computacional* y *Procesamiento de Textos*. Se reiteró la ventaja que tiene PT sobre los otros dos métodos. Adicionalmente, se observó un valor $F = 0{,}8725$ usando PT con lista de asociación de términos.

6. Conclusiones

Se confirmó, con una colección de textos cortos, que los términos con frecuencia de ocurrencia media obtenidos a través del punto de transición, representan mejor a los textos, específicamente en la tarea de agrupamiento. La sencillez para determinar el PT anima a continuar la experimentación, no sólo en AT sino, además, en las vastas aplicaciones del procesamiento automático de textos. Además, al enriquecer los términos con listas de asociación se observa mayor estabilidad con los términos seleccionados por PT.

Es necesario, por supuesto, reforzar las hipótesis sobre el PT con una variedad heterogénea de colecciones, y continuar estudiando las propiedades del PT, particularmente, el contenido semántico de los términos en una vecindad de esta frecuencia.

Bibliografía

Booth, A. D. 1967. A Law of Occurrences for Words of Low Frequency. *Information and control*, 10(4):386–393.

Bueno, C., D. Pinto, y H. Jiménez-Salazar. 2005. El párrafo virtual en la generación de extractos. En H. Calvo, editor, *Research on Computing Science*. Instituto Politécnico Nacional.

Dhillon, I. S., Y. Guan, y J. Kogan. 2002. Refining clusters in high dimensional text data. En *Text Data Mining and Applications*.

Gelbukh, A. F., editor. 2005. *Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005, Proceedings*, volumen 3406 de *Lecture Notes in Computer Science*. Springer.

Hindle, D. 1990. Noun classification from predicate-argument structures. En *28th Annual Meeting of the Association for Computational Linguistics*, páginas 268–275.

Hynek, J. K. y J. O. Rohlikm. 2000. Short Document Categorization Itemsets Method. En Jan M. Zytkow Djamel A. Zighed, Henryk Jan Komorowski, editor, *Principles of Data Mining and Knowledge Discovery*, volumen 1910 de *Lecture Notes in Computer Science*, páginas 9–14, Lyon, France. Springer-Verlag.

Kerner, Y. H., Z. Gross, y A. Masa. 2005. Automatic extraction and learning of keyphrases from scientific articles. En Gelbukh (Gelbukh, 2005), páginas 657–669.

Liu, T., S. Liu, Z. Chen, y W. Ma. 2003. An evaluation on feature selection for text clustering. En T. Fawcett y N. Mishra, editores, *ICML*, páginas 488–495. AAAI Press.

Luhn, H. P. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*, 2(2):159–165.

Makagonov, P., M. Alexandrov, y A. Gelbukh. 2004. Clustering Abstracts instead of Full Texts. En *Proceedings of the Seventh International Conference on Text, Speech and Dialogue (TSD 2004)*, volumen 3206 de *Lecture Notes in Artificial Intelligence*, páginas 129–135, Brno, Czech Republic. Springer-Verlag.

Makagonov, P., M. Alexandrov, y K. Sboychakov. 2000. Keyword-based technology for clustering short documents. *Selected Papers. Computing Research*, páginas 105–114.

Manning, D. C. y H. Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press.

Mikhail, A., A. Gelbukh, y P. Rosso. 2005. An Approach to Clustering Abstracts. En *Proceedings of the 10th International Conference NLDB-05*, Lecture Notes in Computer Science, páginas 8–13, Alicante, Spain. Springer-Verlag. To be published.

Moyotl, E. y H. Jiménez. 2004. An analysis on frequency of terms for text categorization. En SEPLN, editor, *Memorias del XX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, páginas 141–146. SEPLN.

Moyotl-Hernández, E. y H. Jiménez-Salazar. 2005. Enhancement of dtp feature selection method for text categorization. En Gelbukh (Gelbukh, 2005), páginas 719–722.

Pinto, D. y F. Pérez. 2004. Una técnica para la identificación de términos multipalabra. En L. Sandoval, editor, *Proceedings of the 2nd National Conference on Computer Science*, páginas 257–259. BUAP Press.

Rijsbergen, C. J. Van. 1979. *Information Retrieval, 2nd edition.* Dept. of Computer Science, University of Glasgow.

Salton, G. y C. Buckley. 1988. Term-weighted approaches in autometic retrieval. *Information Processing in Management*, 24(5):513–523.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Shin, K. y S. Y. Han. 2003. Fast clustering algorithm for information organization. En A. F. Gelbukh, editor, *CICLing*, volumen 2588 de *Lecture Notes in Computer Science*, páginas 619–622. Springer.

Urbizagástegui, A. R. 1999. Las posibilidades de la ley de zipf en la indización automática. Informe técnico, BUniversidad de California, Riverside.

Voorhees, E. M. 1994. Query expansion using lexical-semantic relations. En W. B. Croft y C. J. Van Rijsbergen, editores, *SIGIR*, páginas 61–69. ACM/Springer.

Zipf, G. K. 1949. *Human behaviour and the principle of least effort.* Addison-Wesley.

Zizka, J. y A. Bourek. 2002. Automated Selection of Interesting Medical Text Documents by the TEA Text Analyzer. En A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing (CICLing-2002)*, volumen 2276 de *Lecture Notes in Computer Science*, páginas 402–404, Mexico DF Mexico. Springer-Verlag.