Selección de términos no supervisada para agrupamiento de resúmenes*

Héctor Jiménez & David Pinto
Facultad de Ciencias de la Computación
B. Universidad Autónoma de Puebla
C.U. Puebla, México
hjimenez@aleteya.cs.buap.mx, 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 selección de términos es un problema que impacta varias tareas del Procesamiento del Lenguaje Natural (PLN), ya que no sólo se pretende reducir el tamaño de la representación de un texto, sino, además, mejorar la eficacia de las tareas del PLN. En este trabajo analizamos los resultados de algunos métodos de selección no supervisada de términos, sin la utilización de fuentes de información externa, en la tarea de agrupamiento de textos; particularmente, resúmenes de artículos científicos de un mismo dominio. Nuestro propósito es iniciar una exploración de los factores que influyen en la eficacia del agrupamiento: métodos de selección de términos, umbrales, fuentes de conocimiento externas, y algoritmos de agrupamiento. Así, presentamos los resultados de algunos experimentos de agrupación que nos arrojan datos para realizar dicha exploración con mayor profundidad.

1. Introducción

En numerosas tareas de procesamiento de textos, como Recuperación de Información (RI), Categorización de Textos (CT), Agrupamiento de Textos (AT), y Resumen Automático (RA), entre otras, es necesario representar los textos usando los términos contenidos en ellos. Pero además, suele hacerse una reducción de estos términos, debido a que gran parte de los términos vician el procedimiento correspondiente, y aumentan innecesariamente el empleo de recursos (almacenamiento en memoria y tiempo de procesamiento). Por ello, se usan variados métodos para elegir los términos que representarán 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 resolver un problema, según la colección de entrenamiento [12], mientras que los no supervisados miden la importancia de los términos con base en características que se supone influyen en la solución. 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; y Ganancia de Información (IG), cuya puntuación representa la cantidad de información que provee un término para predecir la clase de la instancia en la que ocurre.

Ciertamente, los métodos supervisados obtienen mejores resultados que los no supervisados, a cambio del requisito de información sobre la relación de los términos con la solución al problema que se trate; relación entre términos y categorías para CT, entre términos y resumen para RA, etc. Empero, algunos métodos no supervisados pueden llegar a tener eficacia cercana a los supervisados [6]. En este trabajo analizamos, especialmente, un método de selección no supervisada de términos basado en el punto de transición.

El *punto de transición* (PT) es una consecuencia de las observaciones de George Kinsley Zipf, quien 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 [16]. 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 [7].

Algunos autores, han llevado a cabo experimentos con

*Este trabajo fue parcialmente apoyado por BUAP-VIEP 3/G/ING/05, R2D2 (CICYT TIC2003-07158-C04-03), e ICT EU-India (ALA/95/23/2003/077-054).

las ideas anteriores; en la indización automática de textos, y la identificación de palabras clave de un texto [15] [10]. A partir de la ley de ocurrencia de palabras con baja frecuencia propuesta por Booth [3], 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. Esta fórmula es: $PT = (\sqrt{1 + 8 \times I_1} - 1)/2$, donde I_1 representa el número de palabras con frecuencia 1. Empíricamente, partiendo de la hipótesis de que las palabras muy frecuentes tienen rangos diferentes, también es posible identificar el PT. Así, recorriendo las frecuencias, ordenadas ascendentemente, de las palabras de un texto, podemos identificar como PT la primera frecuencia con rango que no se repita. Como puede verse, la determinación del PT es poco costosa; además, será visto que su uso, en la selección de términos, implica casi siempre una reducción de terminos mayor con respecto a los demás métodos.

En este trabajo analizaremos los resultados de un experimento de selección de términos para agrupamiento de resúmenes de artículos científicos, los cuales ofrecen una dificultad particular por ser textos cortos y *narrow domain*, es decir que pertenecen practicamente al mismo dominio (ej. subdominios de la "Lingüística Computacional") ; lo que complica la elección de términos.

Se exponen en las secciones subsecuentes algunos trabajos relacionados, la descripción de métodos de selección no supervisados para selección de términos, y los resultados de un experimento en AT donde se aplican los métodos para selección de términos, y, finalmente se presenta una discusión de los resultados.

2. Algunos resultados sobre selección de términos usando PT

El PT se ha empleado en diversas aplicaciones del PLN. Mencionaremos las referentes a CT, RI, y AT.

En CT el PT se utilizó para "recortar" la selección de términos que determinan los métodos clásicos: IG, CHI, DF [8]. El recorte se realizó descartando los términos elegidos por un método siempre que su frecuencia no ocurra en una vecindad del PT. Con DF, PT recorta el 10% de los términos además mejora F_1. A partir de 20% de la selección con CHI, se obtiene una reducción del 10% de los términos seleccionados y el valor F_1 es el mismo.

Posteriormente, con el fin de conocer el alcance del PT, se definió una puntuación basada en esta frecuencia para seleccionar términos, también, en la tarea de CT [9], lo cual logró resultados comparables a los métodos clásicos. El desempeño del PT en esta tarea es equiparable a los métodos DF, IG, y CHI, a partir del 5% de términos, medido con *microaveraging* sobre la colección REUTERS 21578.

Asimismo, hay algunos resultados preliminares sobre el uso del PT en RI; esencialmente, se utiliza el Modelo de Espacio Vectorial (MEV) para representación de textos, y, también, modificando la ponderación de los términos con base en la cercanía de la frecuencia de un término al PT. La representación de textos se hace mediante su extracto [2], conseguido éste con las palabras cuya frecuencia está alrededor del PT. En una subcolección de TREC-5 compuesta de 884 documentos con el 30% de documentos relevantes, para dos consultas, F_1 fue 0.323 usando el modelo clásico de RI, y $F_1 = 0.286$ para la representación de textos por su extracto. La reducción de índices llega a ser mayor al 10% del total usado por el modelo clásico. Para la ponderación basada en el PT *vs.* la ponderación clásica $tf_{ij} \cdot idf_i$, en la colección antes citada, se tienen valores F_1 semejantes [4].

En AT el PT obtuvo un buen desempeño también en la etapa de selección de términos [5]. El uso del PT en la selección provee mayor estabilidad que los métodos clásicos no supervisados a partir de una selección de 20% de términos. Para varios porcentajes de términos con frecuencia alrededor del PT se obtiene una reducción alrededor del 30% de los seleccionados por otros métodos, y el máximo F lo obtiene PT con 0.6038. También, enriqueciendo los términos seleccionados con términos de asociación, el PT continúa manteniendo su nivel con respecto a los demás métodos, aunque su eficacia es menor que sin el empleo de dicho enriquecimiento ($F = 0.58$ para 20% de términos). Es importante decir que el enriquecimiento se hizo a partir de la misma colección utilizando un método apoyado en frecuencia de coocurrencia.

Debe señalarse que es necesario robustecer los anteriores hallazgos utilizando colecciones de texto diversas y variando los métodos suplementarios en cada aplicación.

Para la tarea de agrupamiento de resúmenes, partimos de los resultados presentados en [1] y el antes descrito[5] en los cuales se trabaja con la misma colección que aquí empleamos, pero con diferente enfoque. En [1], usan una fuente de conocimiento externa para enriquecer la selección no supervisada de términos, y varios algoritmos de agrupamiento, destacadamente *MajorClust*, obteniendo un índice muy alto, $F = 0.78$.

Nuestro propósito es iniciar una exploración de los factores que influyen en la eficacia del agrupamiento: métodos de selección, umbrales, fuentes de conocimiento externas, y algoritmos de agrupamiento. Así, presentaremos algunos experimentos en AT que nos arrojan datos para realizar dicha exploración con mayor profundidad.

3. Métodos de selección

A continuación se describen los métodos no supervisados que utilizaremos.

1. 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.

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

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

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 definición, 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.

3. 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 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, con él, se reportan resultados de AT cercanos a los métodos supervisados [6].

4. Agrupamiento de resúmenes

4.1. Colección de prueba

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

En nuestro experimento se utilizó una colección formada por 48 resúmenes de textos del dominio *Lingüística Computacional y Procesamiento de Textos*, correspondiente al evento *CiCLing 2002* (`http://www.cicling.org`). 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 el 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.

4.2. Método

Consideramos en nuestro experimento una colección de textos $D = \{T_1, \ldots, T_k\}$. 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 (*gold standard*).

Los términos índice de un texto se determinaron siguiendo los métodos presentados en la sección 3. 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* [13], una variante de *NN*, el cual en nuestro experimento apoyamos en la función de similitud de Jaccard. Alternativamente se aplicó el algoritmo *Majority-Clust* [14].

4.3. Medidas de desempeño

Con el propósito de conocer cuál método, y en qué condiciones, realizaba un mejor agrupamiento, utilizamos la medida F, muy empleada en RI [11]. 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:

$$F_{ij} = \frac{2 \cdot P_{ij} \cdot E_{ij}}{P_{ij} + E_{ij}}, \quad (1)$$

donde P_{ij}, y E_{ij} se definen como

$$P_{ij} = \frac{\text{No. de textos del grupo } i \text{ en la clase } j}{\text{No. de textos en el grupo } i}, \quad (2)$$

y

$$E_{ij} = \frac{\text{No. de textos del grupo } i \text{ en la clase } j}{\text{No. de textos en la clase } j}. \quad (3)$$

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

$$F = \sum_{1 \le i \le m} \frac{|G_i|}{|D|} \max_{1 \le j \le n} F_{ij}. \quad (4)$$

4.4. Pruebas

Una prueba inicial fue necesaria para ajustar un factor que permite cambiar el umbral; a partir del umbral canónico: el promedio de las similitudes entre todas las parejas de textos. 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 $TS_i(D)$ ($i = 1, 3, 5, 7, 9, 11, 13, 15, 20, 30, 40, 50, 60$), y la eficacia del agrupamiento se midió con F. Además de los valores F, en las tres columnas finales de la tabla 1, se presenta, separado por una diagonal, el número de grupos obtenidos con la selección de términos efectuada. En la segunda y tercera columna, de dicha tabla aparece el número de términos que usaron los métodos DF y TS, y el que usó PT, respectivamente; los cuales difieren debido a que los primeros toman un porcentaje del total, y, los últimos, un porcentaje del valor de PT por documento.

%	#Ter.	#Ter.PT	PT/#G	DF/#G	TS/#G
1	10	57	0,3340/14	0,4122/4	0,1058/2
3	29	57	0,3340/14	**0,5467**/3	0,2096/4
5	48	57	0,3340/14	0,5263/4	0,2555/7
7	67	57	0,3340/14	0,5263/4	0,2401/8
9	86	59	0,3268/14	0,4044/3	0,3201/11
11	104	71	0,3082/13	0,4044/3	0,3732/10
13	123	108	0,4013/14	0,4044/3	0,3367/10
15	142	108	0,4013/14	0,4044/3	0,3464/12
20	191	133	0,4267/13	0,4044/3	0,3716/13
30	286	181	0,4397/11	0,4309/4	0,4217/11
40	382	263	**0,6038**/7	0,4309/4	0,4353/12
50	478	274	0,5941/7	0,4309/4	**0,4701**/12
60	571	485	0,4948/3	0,4041/4	0,4071/10

Tabla 1. Medidas F para diferentes porcentajes de términos obtenidos por tres métodos de selección.

Como un mecanismo de compararción indicativa se aplicó el algoritmo *MajorClust* al 20% de los términos seleccionados por PT usando dos funciones de similitud, una basada en el coseno y otra en la distancia euclideana. Los valores obtenidos fueron $F = 0.4208$ para el coseno y $F = 0.3983$ para la euclideana.

5. Discusión

Se han presentado los resultados de la eficacia que tienen tres métodos de selección de términos en la tarea de agrupamiento de resúmenes de artículos científicos. Cada uno de ellos fue probado con un amplio rango de umbrales de selección. Todos, también, alimentaron un método variante de *NN*. Los resultados deben ubicarse en el contexto de textos cortos de un mismo dominio, sin supervisión ni el empleo de fuentes de información externas, lo cual hace al problema especialmente difícil.

El método de selección basado en el PT supera los demás y alcanza su máximo en 40% de los términos con frecuencia más cercana al PT y $F_1 = 0.603$ (263 términos)

Es notable que el método DF obtenga $F = 0.54$ con solamente 29 términos. Analizando la selección hecha por DF se encuentran términos espurios de alta frecuencia (como *paper*, y *present*; muy usados en los resúmenes), que, coincidentemente, agruparon en forma correcta. El PT es sensible a las frecuencias bajas, y es una causa del bajo rendimiento con porcentajes menores al 20%. Este aspecto es tratado de manera particular en [1], donde se usa un escalamiento de frecuencias para trabajar con ellas más convenientemente.

En el caso de la aplicación del algoritmo *MajorClust* con 20% de términos elegidos por PT se obtiene un valor casi idéntico al del algoritmo *Star*. Al parecer, en este caso, no es decisivo. En cambio la sensibilidad a la función de similitud habrá que considerarla en futuros experimentos, así como reforzar la hipótesis sobre los variados algoritmos de agrupamiento.

Es notable, también, que sin fuentes de conocimiento externas se logre un índice $F = 0.603$, comparando éste con el obtenido en [1] donde obtuvieron 0.78. Para el enriquecimiento de términos con la propia colección los resultados son pobres (no superan la prueba realizada sin enriquecimiento) por la imposibilidad de aplicar filtros como la información mutua. Lo anterior indica que el enriquecimiento de los términos con algún tipo de diccionario, por ejemplo WordNet, podría ser decisivo en el aumento de F.

En suma, podemos concluir que: a) deben realizarse más pruebas con algoritmos variados de agrupamiento; b) es conveniente el uso de diccionarios para enriquecer los términos; c) la combinación de técnicas de escalamiento de frecuencias con el PT podría superar las dificultades de bajas frecuencias; y d) la utilización de otras colecciones, más grandes y variadas del mismo tipo (igual dominio y textos cortos) podrán afianzar los resultados obtenidos.

6. Agradecimientos

Deseamos agradecer la colaboración del Dr. Mikhail Alexandrov, investigador del Centro de Investigación en Computación-IPN, por los resultados proporcionados con respecto al agrupamiento con el algoritmo *MajorClust*.

Referencias

[1] M. Alexandrov, A. Gelbukh & P. Rosso: An Approach to Clustering Abstracts, *NLDB*, 8-13, Springer, 2005.

[2] C. Bueno-Tecpanécatl, D. Pinto & H. Jiménez-Salazar: El párrafo virtual en la generación de extractos, en *Advances in Computer Science in México*, A. Gelbukh & H. Calvo (Eds.), p. 83-90, 2005.

[3] A. Booth: A Law of Occurrences for Words of Low Frequency, *Information and control*, 10(4) pp 386-93, 1967.

[4] R. J. Cabrera, D. Pinto, D. Vilariño & H. Jiménez-Salazar: Una nueva ponderación para el modelo de espacio vectorial para recuperación de información, en *Advances in Computer Science in México*, A. Gelbukh & H. Calvo (Eds.), p. 75-82, 2005.

[5] H. Jiménez-Salazar, D. Pinto & P. Rosso: El uso del punto de transición en la selección de términos índice para agrupamiento de textos cortos, *XXI Congreso de la SEPLN*, en prensa, 2005.

[6] T. Liu, S. Liu, Z. Chen & W. Ma: An Evaluation on Feature Selection for Text Clustering, *ICML*, p. 488-495, 2003.

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

[8] E. Moyotl & H. Jiménez: An Analysis on Frecuency of Terms for Text Categorization, *Proc. of SEPLN-04, XX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, pp 141-146, 2004.

[9] E. Moyotl & H. Jiménez: Enhancement of DPT Feature Selection Method for Text Categorization, *Proc. of CICLing-2005, Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, pp 706-709, 2005.

[10] M. L. Pao: Automatic indexing based on Goffman's transition of word occurrences, *American Society for Information Science*, 1977.

[11] C. J. van Rijsbergen: *Information Retrieval*. London, Butterworths, 1999.

[12] F. Sebastiani: Machine Learning in Automated Text Categorization, *ACM Computing Surveys*, Vol. 34(1), pp 1-47, 2002.

[13] K. Shin & S.Y. Han: Fast Clustering Algorithm for Information Organization, *CICLing*, p 619-622, Springer, 2003.

[14] B. Stein: Document Categoriztion with MajorClust, *Proc. 12th Workshop on Information Technology and Systems*, 2002.

[15] A. R. Urbizagástegui: Las Posibilidades de la Ley de Zipf en la Indización Automática, `http://www.geocities.com/ResearchTriangle/2851/RUBEN2.htm`, 1999.

[16] G. K. Zipf: *Human Behaviour and the Principle of Least Effort*, Addison-Wesley, 1949.