Reducción de términos índice usando el punto de transición*

Berenice Reyes-Aguirre, Edgar S. Moyotl-Hernández & Héctor Jiménez-Salazar
Facultad de Ciencias de la Computación
B. Universidad Autónoma de Puebla
14 sur y Av. San Claudio. Edif. 135. Ciudad Universitaria
Puebla, Pue. 72570. México
Tel. (01222) 229 55 00 ext. 7212 Fax (01222) 229 56 72
reyesbere@mail.cs.buap.mx, emoyotl@mail.cs.buap.mx, hjimenez@fcfm.buap.mx

RESUMEN

Con el modelo vectorial usado en la representación de información se obtienen los índices que representan a los documentos de una colección. En este trabajo se identificaron las palabras más representativas de un texto con base en el punto de transición, i.e., la frontera entre las palabras de baja y alta frecuencia, derivado de la Ley de Zipf. Un texto se representa, así, por las palabras alrededor del punto de transición. Para conocer el desempeño de esta representación se calculó el centroide de una colección de textos de un dominio particular, y para cada texto de una muestra de dos dominios más, diferentes al de la colección, se determinó la similitud. Los resultados muestran que con un número reducido de palabras representativas se puede obtener un desempeño equiparable al método vectorial clásico.

I. INTRODUCCIÓN

Hoy en día, la información, cada vez más abundante, ha generado la necesidad de mejorar los métodos de acceso al caudal cotidiano de imágenes y textos. La categorización de texto, que consiste en la asignación automática de documentos a categorías predefinidas ha sido reconocida como una herramienta útil para ayudar a los usuarios a acceder a gran cantidad de textos disponibles en Internet y en sus organizaciones. Por otra parte, con el surgimiento del modelo empirista del tratamiento de la lengua, que considera los corpora (conjunto o muestra de textos que representa a un dominio, un uso lingüístico, etc.) como una fuente de conocimiento, también surge la necesidad de desarrollar nuevas técnicas para facilitar la adquisición y manejo de los mismos. Para realizar las tareas anteriores es necesaria una etapa de preprocesamiento para transformar los textos a algún tipo de representación que facilite su posterior análisis. En muchos sistemas que manejan textos, principalmente de recuperación de información, la representación se hace mediante un vector de pesos de los términos que el texto contiene. Entonces, las palabras usadas en la representación de los documentos deberán caracterizar a cada uno de estos. Se ha comprobado experimentalmente que las palabras con muy baja frecuencia de ocurrencia no contribuyen a la discriminación adecuada de los documentos, mientras que las palabras muy frecuentes no permiten distinguir entre diferentes documentos. Debido a la ley de Zipf si las palabras de un texto se ordenan descendentemente por su frecuencia de ocurrencia, para definir el rango (posición de una palabra en la ordenación), entonces el producto de la frecuencia por el rango es constante. Con regularidades como la que formuló Zipf es posible identificar palabras que son más representativas que otras dentro de un texto. En este trabajo usamos el punto de transición, una consecuencia de la ley de Zipf, para identificar palabras representantes de un texto.

En el año de 1999 Rubén Urbizagástegui [3], de la Universidad de California, empleó el así llamado *punto de transición*, dentro de la lista de términos que ocurren en un texto para identificar los términos de baja y alta frecuencia. El experimento realizado se basó en la Ley de Zipf de frecuencias bajas la cual se presenta en [1]. Para un artículo corto de una revista identificó los términos índice que describían adecuadamente su contenido. El éxito de este experimento radica en que a pesar de su sencillez, se observa, que con base en la frecuencia de ocurrencia de las palabras en un texto, la ley de Zipf es capaz de predecir con notable precisión ciertos parámetros como son la cantidad de palabras de cierto rango. En este experimento, los términos identificados fueron muy parecidos a las palabras clave del artículo, y por tanto es un buen camino a seguir para reconocer si el contenido de un texto es similar al de otro.

En este trabajo se toma como base [3] y los resultados derivados de la Ley de Zipf que aparecen en [1]. En

*Este trabajo es parte del proyecto CONACYT No. I39165A

la siguiente sección se fundamenta el punto de transición. Inmediatamente después, el modelo vectorial, que es el marco de trabajo para realizar la representación de textos, y, posteriormente, se describe el experimento llevado a cabo. Al final, se dan las conclusiones del presente trabajo.

II. PUNTO DE TRANSICIÓN

Referimos en esta sección principalmente al artículo de A. Booth [1]. Consideremos para un texto T la tabla de las palabras que ocurren en T dispuesta en orden descendente respecto a su frecuencia de ocurrencia en el texto T. El rango de una palabra v será la posición en la que ocurra v dentro de la tabla. Así, la palabra más frecuente tendrá rango 1, de las restantes la más frecuente tendrá rango 2 (si su frecuencia de ocurrencia es diferente a la más frecuente de todas, en caso contrario su rango será también 1), y así sucesivamente. Denotemos con $p(r)$ la probabilidad de que una palabra tenga rango r. Si el texto contiene N palabras (incluyendo repetidas) entonces, para una palabra v con rango r y frecuencia f, teóricamente $Np(r) = f$. Zipf [4] observó que si una palabra cumple:

$$1{,}5 > Np(r) \geq 0{,}5, \tag{1}$$

entonces tal palabra ocurre con frecuencia 1 en los textos. La ley de Zipf sugiere que $p(r) = k/r$, donde k es una constante asociada al texto. De esta forma, substituyendo en la Ec. (1) tenemos

$$1{,}5 > Nk/r \geq 0{,}5. \tag{2}$$

A partir de la Ec. (2) podría decirse entonces que hay dos valores para r, uno mínimo y otro máximo:

$$r_{min} = \frac{kN}{1{,}5}, \qquad r_{max} = \frac{kN}{0{,}5}, \tag{3}$$

esto es, hay varias palabras con frecuencia 1 para las cuales se debe satisfacer que su rango está entre estos valores. Más concretamente, se considera, al igual que en la Ec. (1), que los valores de los rangos para palabras de frecuencia 1 son los dados por la Ec. (3). Si I_1 representa el número de palabras con frecuencia 1, entonces $I_1 = r_{max} - r_{min}$, lo cual conduce a que

$$I_1 = \frac{4}{3}kN. \tag{4}$$

Como en (1), algo semejante puede hacerse para el número de palabras con frecuencia n. Una palabra ocurre n veces en un texto si cumple:

$$n + \frac{1}{2} > Np(r) \geq n - \frac{1}{2}. \tag{5}$$

Entonces, el número de palabras con frecuencia n, sería:

$$I_n = \frac{kN}{n^2 - \frac{1}{4}}. \tag{6}$$

Con las Ecs. (1) y (6) se llega a la siguiente proporción, independiente de las constantes del texto k y N, es decir válida para cualquier texto:

$$I_n/I_1 = \frac{3}{4n^2 - 1}. \tag{7}$$

Notemos que la observación de Zipf referente a que las palabras con frecuencia 1 está dada por la Ec. (1), puede reformularse suponiendo que

$$2 > Np(r) \geq 1, \tag{8}$$

y algo semejante con la ecuación (5). Ahora con (5), desarrollamos los calculos anteriores para llegar a que:

$$I_n/I_1 = \frac{2}{n(n+1)}. \tag{9}$$

En [3] se supone que $I_n = 1, (n > 1)$, para un valor fijo de n, lo cual refiere a una de las palabras cuya frecuencia de ocurrencia en el texto es n. Notemos que tales palabras no son raras en los textos pues su frecuencia ya no es baja. Este valor de n señala precisamente la posición, en la tabla de palabras ordenadas por su frecuencia, del punto llamado de transición pues abajo de él habrá palabras poco usadas y arriba palabras muy usadas. Se supone, como se observa en [3] y [2] que las palabras con frecuencia media tienen el mayor contenido semántico del texto donde aparecen. Es por ello que interesa conocer el valor de n. Al aplicar la condición $I_n = 1$ en (9) podemos despejar n y obtener:

$$n = \frac{\sqrt{1 + 8I_1} - 1}{2}. \tag{10}$$

Obtenido el punto de transición, deberán seleccionarse términos alrededor de él para conformar el conjunto de palabras que representan al texto. Los experimentos reportados en [3] indican que es posible tomar una banda de frecuencias del 25 % alrededor del punto de transición.

III. CÁLCULO DE ÍNDICES

En la sección anterior hemos dicho cómo se calcula el punto de transición, con lo cual es posible extraer una banda de frecuencias de palabras del texto que tendrán mayor contenido semántico. Una forma alternativa de

definir los términos que representan a un texto lo expresa el modelo vectorial [2].

III.1 Modelo vectorial

En este modelo un documento D_i puede considerarse como un vector representado por sus términos índice. Los términos pueden estar ponderados de acuerdo con su importancia, i.e.,

$$D_i = (d_{i1}, d_{i2}, \ldots, d_{im}), \tag{11}$$

donde d_{ik} representa el peso del término k-ésimo en el documento i.

Por otro lado tendremos la frecuencia de un término k por documento i, denotada por tf_{ik}, y la frecuencia de un término en la colección de documentos, denotada por df_k, esto es el número de documentos que hacen referencia al término k. Para asignar el peso de los términos utilizamos:

$$idf_k = \log_2(M) - \log_2(df_k) + 1,$$

donde k es el término, M el número de documentos y df_k es la frecuencia del término k en el corpus. Finalmente, el peso estará dado por la siguiente fórmula:

$$d_{ik} = tf_{ik} \cdot idf_k. \tag{12}$$

Dados los vectores índice para dos documentos, D_i y D_j, es posible medir el coeficiente de similitud entre ellos, $sim(D_i, D_j)$. El modo más común de determinar la similitud, entre dos documentos representados mediante el modelo vectorial (Ecs. (11) y (12)), es calculando el coseno del ángulo formado por ambos vectores:

$$sim(D_i, Dj) = \frac{\sum_{k=1}^{m} d_{ik} d_{jk}}{\sqrt{\sum_{k=1}^{m} d_{ik}^2 \cdot \sum_{k=1}^{m} d_{jk}^2}}. \tag{13}$$

En esta ecuación se incluye la normalización de los vectores, para no privilegiar documentos largos frente a otros documentos menos extensos.

En la representación vectorial se discriminan términos considerando que los que mejor diferencian a los documentos tienen una frecuencia de referencia por documento df_k de acuerdo con:

$$df_k \in \left[\frac{M}{100}, \frac{M}{10}\right], \tag{14}$$

donde M es el número de documentos [2].

Para representar un conjunto de documentos se emplea el centroide que es un vector característico en el cual cada componente está definida como el promedio de los pesos de los términos de todos los documentos del conjunto. Dado un conjunto K de documentos $\{D_1, \ldots, D_m\}$, el centroide es:

$$C_K = \frac{1}{m} \sum_{j=1}^{m} D_j. \tag{15}$$

A través de la discriminación de textos para un dominio confrontamos las dos formas de elección de índices, lo cual se expone en la sección siguiente.

IV. PRUEBAS REALIZADAS

Para reconocer la efectividad de nuestra propuesta sobre palabras representativas, se tomó un corpus de un dominio determinado y se calculó su centroide. Posteriormente, se tomó una muestra de textos de tres dominios, dos de ellos diferentes al del corpus. Cada uno de estos documentos fue representado en forma vectorial, y usando los índices provistos por el punto de transición. Finalmente, en cada caso se calculó la similitud con el centroide, y usando un umbral se decidió si el documento pertenecía al dominio del corpus. En las siguientes subsecciones damos detalles de este proceso.

IV.1 Corpus

El corpus contiene un total de 63,083 términos provistos por 80 noticias sobre narcotráfico; las cuales fueron publicadas en los años de 1996 a 1999 en el periódico La Jornada. El número total de palabras diferentes en el corpus original (sin preprocesamiento) fue de 7,584, con un promedio de 94.8 palabras por documento. Se trata de documentos sin uniformidad en cuanto a su tamaño, ya que varían entre 3Kb y 9Kb. No hemos empleado ningún sistema de lematización, que nos permitiera trabajar con términos normalizados en lugar de palabras crudas.

IV.2 Procedimiento

En el proceso de entrenamiento se construyo el vector centroide para el corpus mediante (15), utilizando un sistema de pesos, calculados a partir de la fórmula (12). Para probar el sistema se empleó una colección de 25 noticias publicadas en el año 2002 del periódico Milenio. La distribución de estas noticias en tres dominios fue como sigue: diez sobre narcotráfico, diez sobre guerra en Irak, y cinco sobre seguridad informática. Para estimar el grado de similitud entre los documentos

a seleccionar y el centroide se empleó el coeficiente del coseno Ec. (13). El criterio para decidir si un nuevo texto es del dominio del corpus es con base en un umbral. Éste se define calculando la similitud entre cada uno de los documentos del corpus y el centroide, y tomando el menor índice de similitud obtenido como umbral. De manera que un documento situado por encima de este valor indicará que el documento es similar al corpus.

Las operaciones de preprocesado efectuadas han sido:

- Separación de palabras utilizando el carácter espacio y los caracteres de puntuación. Por ejemplo, un texto como *¿Por qué?, preguntó* quedará separado en las siguientes palabras "Por qué pregunto".
- Detección de nombres propios simples y formados con la inclusión "de/de la", suponiendo que el texto está en minúscula, salvo nombres propios o inicio de oración. Por ejemplo, en un texto como "Procuraduría General de la República" se toma "Procuraduría_General_República" como nombre propio.
- Eliminación de palabras cerradas (palabras casi vacías de contenido semántico, como artículos, preposiciones, conjunciones, etc.).

Se realizaron 2 experimentos, con el objetivo de obtener resultados sobre el efecto del método de representación. En el primer experimento los índices se determinan de la manera tradicional, esto es con base a (14). Los índices fueron aquellos cuya df_k tuviera un valor entre 1 y 8, obteniendo con esto un total de 6,853 índices. En el segundo experimento se seleccionaron los términos utilizando términos elegidos en una banda de frecuencias determinada por $n/1{,}25 \leq tf_k \leq n1{,}25$, donde tf_k es el número de ocurrencias de la palabra k en el corpus, y $n = 86$ es el punto de transición obtenido a partir de (10). Aquí tf_k quedó entre 69 y 107 lo cual arrojó como resultado 23 índices. Los umbrales empleados para decidir el reconocimiento de un documento a la clase del corpus de entrenamiento fueron: 0.10 para el método vectorial clásico, y 0.33 usando la reducción de Zipf.

En la Tabla 1 se muestra para cada método empleado en la selección de índices, para cada muestra de textos de los tres dominio ($\mathbf{N}$, narcotráfico; $\mathbf{G}$, guerra y $\mathbf{S}$, seguridad) empleados en la prueba, la cantidad de documentos que fueron reconocidos como textos del dominio de narcotráfico y en las dos últimas columnas la precisión $\mathbf{P}$ y la evocación $\mathbf{R}$ calculadas como: el número de documentos del total que fueron correctamente identificados dividido por el total de documentos identificados, y el número de documentos del total que fueron correctamente identificados dividido por el número total de documentos que realmente pertenecen a la clase, respectivamente. El índice $\mathbf{F}$ concentra las anteriores medidas y fue calculado con la fórmula $\mathrm{F} = 2\mathbf{PR}/(\mathbf{P}+\mathbf{R})$.

Método	**N**	**G**	**S**	**Total**	**P**	**R**	**F**
Clásico	8	1	2	11	0.72	0.8	0.7578
Zipf	10	4	2	16	0.62	1.0	0.7654

Tabla 1: Resultados de la identificación de documentos pertenecientes a un dominio.

V. Conclusiones y trabajo futuro

Hemos presentado una aplicación del punto de transición, el valor que divide las palabras de un texto en las de alta y baja frecuencia. La aplicación se dirigió a determinar si un texto pertenece a un dominio particular. Este procedimiento puede verse como la solución a un subproblema de la categorización de textos. Ciertamente, hay muchos métodos para el efecto de categorización, sin embargo, lo notorio es que el método empleado utiliza pocos rasgos extraídos del texto[1]. Además, como se muestra en los resultados (tabla 1) el desempeño es equiparable al método vectorial clásico.

Situamos las fallas de la aplicación realizada en varias direcciones. En primer lugar, los temas ($\mathbf{N}$, $\mathbf{G}$ y $\mathbf{D}$) son "cercanos", esto lleva a resolver un problema no trivial. Segundo, el corpus es pequeño, lo cual favorece despreciar rasgos importantes. Por último, el procesamiento no fue realizado como hubiéramos deseado, por ejemplo no se llevó a cabo la lematización de los términos contenidos en los documentos.

Entre los trabajos futuros, debe incluirse una prueba más completa para definir con precisión los alcances del método presentado, esto tendrá que considerar un corpus mayor, y probar en varios dominios, así como un número de textos mayor. Las aplicaciones inmediatas que se tienen son a la categorización de texto y a la compilación automática de corpora.

REFERENCIAS

[1] Booth, A.D.: "A law of occurrences for words of low frequency", *Information and control*, 10(4) pp 386-93, 1967.

[2] Salton, G., Wong, A. & Yang, C.S.: "A vector space model for automatic indexing", *Information Retrieval and Language Processing*, pp 613-620, 1975.

[1] En otros experimentos que estamos realizando el número de índices encontrados por el punto de transición no rebasa el 5 % del total del enfoque clásico.

[3] Urbizagástegui-Alvarado, R.: "Las posibilidades de la ley de Zipf en la indización automática", *Reporte de la Universidad de California Riverside*, 1999.

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