Una Nueva Ponderación para el Modelo de Espacio Vectorial de Recuperación de Información

[1]Rubí J. Cabrera, [2]David Pinto, [3]Darnes Vilariño, [4]Héctor Jiménez-Salazar

Facultad de Ciencias de la Computación,
Benemérita Universidad Autónoma de Puebla,
14 sur y Av. San Claudio, Ciudad Universitaria, Edif. 135
Puebla, Pue., México, C.P. 72570
Tel. (+52-222)2295500 Ext. 7227, Fax (+52-222)2295672
[1]rubicabrera@hotmail.com, {[2]dpinto, [3]darnes, [4]hjimenez}@cs.buap.mx

Resumen. El modelo planteado busca representar los documentos a través de los términos más importantes del *corpus*, usando un esquema de pesos en cierta manera similar al modelo de espacio vectorial (MEV). Se utilizan dos componentes para el peso de un término en un documento; el primer componente se calcula mediante la distancia inversa del término al punto de transición de su documento, mientras que el segundo componente se basa en la distancia al punto de transición global (considerando todo el *corpus*). Los resultados obtenidos muestran valores de precisión por niveles comparables con el peso clásico $tf * idf$, sin embargo, la dimensionalidad de los términos en nuestro modelo es de tan solo el 3% de la MEV clásico.
Palabras clave: Ponderación de términos, Sistemas de Recuperación de Información.

1 Introducción

1.1 Sistemas de Recuperación de información

La creación de los buscadores de información en Internet sigue impulsando la generación de nuevas técnicas para la representación de información y por consiguiente la recuperación de la misma. La forma más común de encontrar información en Internet es utilizando los llamados "motores" de búsqueda o buscadores. Algunos de los más populares son: Google, Yahoo, Altavista, Excite, InfoSeek, Web Crawler, entre otros. Existen buscadores en casi todos los idiomas del mundo y algunos de ellos tienen ciertas "especialidades". Los mismos han sufrido grandes mejoras con las diversas investigaciones acerca de la evolución de la información; un caso particular es el de Google el cual realizó mejoras en sus técnicas de indización automática para lograr y brindar un buen funcionamiento [6][7]. Ciertamente, resulta necesaria la creación de nuevas técnicas y mejoras a las ya existentes para proporcionar información de una forma eficaz ante una búsqueda dada, sobre todo en el lenguaje Español.

Los Sistemas de recuperación de información (SRI) consisten básicamente en un conjunto de procesos interrelacionados que permiten obtener información

de interés a partir de una determinada consulta [3]. Un SRI soporta una serie de operaciones sobre la colección de documentos almacenados, como son: introducción de nuevos documentos, reindización de los documentos almacenados y eliminación de los mismos. También debe contar con un método de localización de documentos, para presentarlos según la consulta del usuario. Los SRI implementan estas operaciones en modelos diversos, lo que provoca una amplia variedad en lo relacionado con la naturaleza de los mismos; regularmente es posible encontrar variaciones con respecto a los métodos de búsqueda y técnicas de representación [2].

En este artículo se presenta un nuevo modelo de ponderación de términos de los documentos, basado en la técnica llamada Punto de Transición [9] (que posteriormente abordaremos), así mismo, se realiza una comparación con la ponderación clásica $tf * idf$ [2] y se obtiene una evaluación. En la primera sección se muestra la problemática que implica la creación de nuevas técnicas para la representación de documentos; así también, se hace mención de los mecanismos de representación existentes. En la segunda sección se describe el modelo de ponderación propuesto y se define la técnica en la que se basa dicho modelo; se muestran también las fórmulas para el cálculo de pesos que definen al nuevo modelo de ponderación y la definición de la ponderación de la consulta. En la tercera sección se describen los resultados experimentales usando un *corpus* de prueba (TREC–5 [8]); también se muestra la evaluación del nuevo modelo propuesto, comparándolo con el modelo de espacio vectorial clásico. Por último se realizan una serie de conclusiones en base a los resultados obtenidos y se discuten algunos trabajos a futuro basados en el modelo propuesto.

1.2 Mecanismos de representación

En [6], se refiere a la proliferación considerable en estos últimos años de herramientas para buscar información en la Web; se estima que en la actualidad existen más de 2000 motores de búsqueda diferentes en la Web, mientras que en 1995 había tan solo una docena. Cada uno de ellos tiene sus propias características, utilidades e interfaces de usuario.

Con respecto a la representación de la información almacenada, en [5] se hace referencia a la indización, o indexación, como la operación destinada a representar los resultados del análisis de contenido de un documento o de una parte del mismo, mediante elementos (denominados genéricamente 'términos de indización') de un lenguaje documental o natural, generalmente para facilitar la recuperación.

Un modelo de representación tiene como objetivo satisfacer las necesidades reales y potenciales de información de todos los usuarios, proporcionándoles la información veraz pertinente, justo a tiempo y al menor coste. En particular, contempla una serie de etapas con las que debe cumplir para ser considerado un modelo de representación de RI óptimo [1]:

1. Obtener representación de los documentos. Generalmente los documentos se presentan utilizando un conjunto más o menos grande de términos índice.

2. Identificar la necesidad informativa del usuario. Se trata de obtener la representación de esa necesidad, y plasmarla formalmente en una consulta acorde con el sistema de recuperación.
3. Búsqueda de documentos que satisfagan la consulta. Consiste en comparar las representaciones de documentos y la representación de la necesidad informativa para seleccionar los documentos pertinentes.
4. Obtención de resultados y presentación al usuario.
5. Evaluación de los resultados por parte del usuario.

En [7] se hace referencia al gran desarrollo y crecimiento que ha tenido el motor de búsqueda Google como mecanismo de representación de documentos, debido a su extenso uso de la estructura en hipertexto y a su diseño de arrastre e indexación eficiente del Web.

En [6] se describe la paginación de la Web como un método para clasificar las páginas Web objetivamente y mecánicamente con una gran efectividad para el intéres humano.

En particular, en la Facultad de Ciencias de la Computación, BUAP, se están desarrollando un conjunto de herramientas destinadas al proceso de RI. Una técnica particular que se encuentra en investigación en este grupo, es precisamente la del punto de transición. Un trabajo derivado del estudio de esta técnica se puede ver en [4], en donde se presenta un mecanismo para reducir los términos de representación de un documento mediante el PT.

En este artículo se hace uso de la técnica PT para la obtención de un nuevo mecanismo de ponderación comparándolo con el peso clásico del modelo de espacio vectorial propuesto por Salton [2].

2 Modelo Propuesto

2.1 Punto de Transición

El Punto de Transición (PT) refiere básicamente a un término en el vocabulario del texto que divide al mismo vocabulario en términos de alta y baja frecuencia. Urbizagástegui [9], por ejemplo, se refiere a este concepto a través de la ley de Zipf [10], y presenta un ejercicio en donde argumenta el hecho de que, existe una vecindad de términos alrededor del punto de transición que describen de manera general el contenido del mismo texto. Este concepto es sumamente importante, ya que dichos términos podrían utilizarse para representar el documento. La fórmula para la obtención del valor de frecuencia del PT se muestra en la ecuación (1).

$$PT = \frac{\sqrt{1 + 8 * I_1} - 1}{2} \tag{1}$$

Donde I_1 es el número de términos que tienen frecuencia 1.

En nuestro caso, la representación de cada documento se realiza por medio de un conjunto de términos pertenecientes al vocabulario [1] del mismo documento y que poseen un valor de frecuencia tan cercano al punto de transición como PT*.25; es decir, se escoge una vecindad del 25% alrededor del PT, la cual resultó la mejor opción después de experimentar con diversos valores de vecindad.

Obtenido el punto de transición con una banda de frecuencia del 25%, deberán seleccionarse términos alrededor de él para conformar el conjunto de palabras que representarán al documento.

2.2 Modelo de Ponderación Propuesto

Una forma de representar los documentos es por medio del cálculo de pesos, en donde se asigna un valor numérico a cada término del documento. En este caso, se propone establecer un peso sobre los términos alrededor del punto de transición, de acuerdo a la distancia de los mismos hacia el PT. Lo anterior, evaluado por las fórmulas mostradas en (2), (3) y (4).

Así, *Wij* es el peso que le corresponde al término i en el documento j.

$$W_{ij} = IDPT_{ij} * DPTC_i \tag{2}$$

Donde, *IDPTij* es la distancia inversa del término i-ésimo al punto de transición del documento j. Se eleva al cuadrado el denominador para asignar un valor de importancia cuadrático a los términos, en función de su cercanía al PT.

$$IDPT_{ij} = \left| \frac{1}{\left| PT_j - F(t_{ij}) \right|^2} \right| \tag{3}$$

DPTCi, por su parte, es la distancia de la frecuencia del término i del vocabulario del *corpus* al punto de transición evaluado sobre todo el *corpus* (PT global).

$$DPTC_i = \sqrt{(PT - F(t_i))^2} \tag{4}$$

2.3 Ponderación de la Consulta

Los usuarios que consultan a través del sistema de recuperación (SRI) para buscar información, deben traducir su necesidad informativa en una consulta adecuada al SRI. Esto supone utilizar un conjunto de términos que expresen semánticamente su necesidad. En sistemas tradicionales es habitual utilizar un valor de peso asociado a cada término de la consulta. En este caso se asigna el peso usando el cálculo definido en la sección 2.2, a excepción de que *IDPTij* se asume con un valor de 1, debido principalmente a la cantidad y calidad de términos usualmente contenidos en una consulta.

[1] El vocabulario de un texto es el conjunto de palabras no repetidas del mismo documento.

3 Resultados Experimentales

3.1 *Corpus* de prueba

TREC (Text Retrieval Conference) constituye uno de los esfuerzos más significativos de investigación experimental en recuperación de información (RI). El patrocinio de elaboración de estas conferencias se encuentra a cargo de la National Institute of Standards and Technology (NIST) y de la Defense Advanced Research Projects Agency (DARPA). Dichas conferencias comenzaron en 1992 (TREC–1) y vienen celebrándose con periodicidad anual hasta la fecha. De manera particular, en 1996 se celebró TREC–5, en cual se utilizó una colección de aproximadamente 250,000 noticias en español. El TREC–5 posee alrededor de 50 consultas supervisadas y las respuestas son indicadas en el mismo *corpus* [8]. La idea es establecer comparaciones fiables entre los distintos sistemas empleados por los investigadores en TREC–5, dado que todos operan con las mismas colecciones y las mismas consultas, y presentan sus resultados en la misma forma; obviamente, utilizan sistemas y técnicas diferentes. En base a lo anterior se creó un *corpus* basado en el TREC–5 para la realización de las pruebas, comprendido de 884 noticias correspondientes al Diario el Norte de Guadalajara, de las cuales el 33% consiste de las noticias relevantes de las consultas 26 y 28; el 67% restante consiste de noticias no relevantes para estas mismas consultas. Los resultados obtenidos se muestran en la siguiente sección.

3.2 Consultas Utilizadas

El SRI fue evaluado con las siguientes dos consultas tomadas del TREC–5.

1. *Indicaciones de las relaciones económicas y comerciales de México con los paises europeos.* (Consulta 26)
2. *Indicaciones de las relaciones económicas y comerciales de México con los paises asiáticos, por ejemplo Japón, China y Corea.* (Consulta 28)

3.3 Evaluación con el MEV

Se utiliza una gráfica de precisión por niveles estándar para comparar el nuevo modelo de ponderación de términos en documentos contra el modelo de espacio vectorial propuesto por Salton [2].

Las gráficas muestran que ambos sistemas obtienen resultados comparables, sin embargo, es importante remarcar que el modelo de ponderación propuesto usa únicamente el 3% de términos que usa el modelo de espacio vectorial clásico. Esto implica incluso que los tiempos de cálculo de representación de documentos se reducen drásticamente.

En la figura 1, se presenta la evaluación sobre la consulta 26 del TREC–5, mientras que en la figura 2 se evalua sobre la consulta 28.

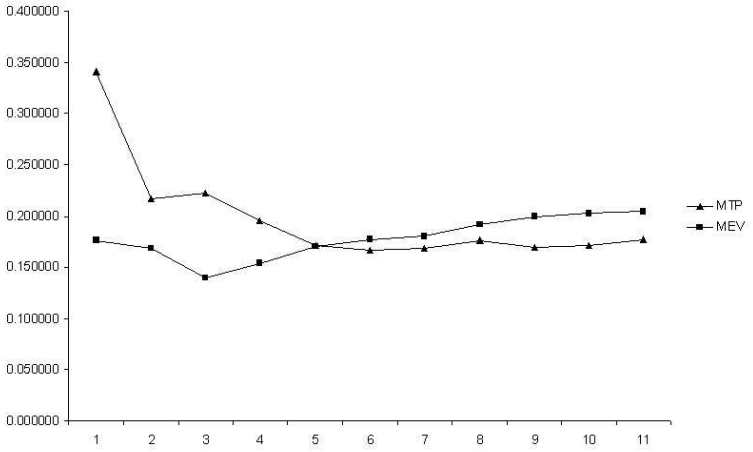

Fig. 1. Consulta 26

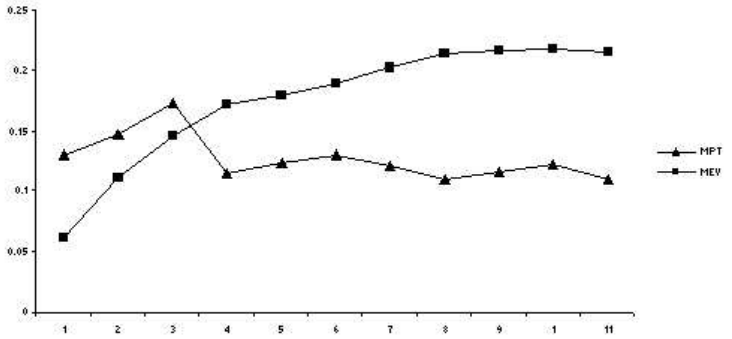

Fig. 2. Consulta 28

4 Conclusiones

Se ha presentado un nuevo modelo de ponderación de términos de documentos que obtiene una reducción de la dimensión del espacio vectorial bastante amplia. El efecto de reducción de términos proviene de la utilización de una técnica llamada punto de transición. El modelo usa una fórmula que asigna pesos a los términos utilizando dos componentes para medir la importancia de un término en un *corpus*. El primer componente mide la importancia de un término en un documento, obteniendo la distancia inversa de la frecuencia del término al PT del mismo documento. Por otro lado, el segundo componente mide la importancia del término en todo el corpus, y esto lo hace a través de la distancia hacia el PT global (i.e., considerando todo el *corpus*). Los resultados obtenidos muestran que el nuevo modelo de ponderación es comparable con el modelo clásico $tf * idf$ en cuanto a la precisión por niveles estándar, sin embargo, el nuevo modelo posee una cardinalidad sumamente menor que el MEV clásico. En este caso, el MEV clásico tiene vectores de representación con 30,202 términos, mientras que el nuevo modelo únicamente usa 1,185.
Es necesario hacer más pruebas que revelen los alcances de esta propuesta de ponderación. Por ejemplo, la determinación del umbral alrededor del punto de transición, considerar pruebas con todos los tópicos del TREC–5 y ajustar la fórmula de pesos para ganar eficiencia en el cálculo de índices.

Referencias

1. Angel F., Rodríguez Zazo, Figuerola G., Alonso J.L. and Gómez R., *Recuperación de Información utilizando el Modelo Vectorial*, Departamento de informática y automática, Universidad de Salamanca, 2002, Mayo.
2. G. Salton,*Automatic Text Processing*,Addison-Wesley,(1989).
3. Jiménez H. and Pinto D.,*Notas de Academia Recuperación de Información*,Octubre,(2003).
4. Moyotl E., Reyes B. and Jiménez H., *Reducción de términos índice usando el Punto de Transición*, (2003).
5. Normas Fundamentales, *Norma UNE 50-113-92 Documentación e información*, AENOR,23-73, vol. 2,(1997).
6. *The Pagerank citation Ranking Bringing Order to Web*, January,(1998).
7. Sergey Brin and Lawrence Page,*The Anatomy of a Large-Scale Hypertextual Web Search Engine*,Computer Science Department, Stanford University, Stanford, CA 94305, USA.
8. http://trec.nist.gov/Text Retrieval Conference (TREC) última revisión 13/sep/04.
9. Urbizagátegui A. R., *Las implicaciones de la ley de Zipf en la indización automática*,Universidad de California Riverside,(1999).
10. Zipf George K., *Humand Behavior and the Principle of Least-Effort*, Addison-Wesley, Cambridge MA,(1949).