NUEVO MODELO DE REPRESENTACIÓN DE TEXTOS: GRAFOS DE ASOCIACIÓN

Resumen / *Abstract*

En este artículo se presenta un nuevo modelo, denominado grafo de asociación, para mejorar la representación de documentos, facilitando su dimensión ontológica. Se explica cómo crear y usar este tipo de grafo. También se analiza una medida de similitud de documento basada en esta representación. El modelo clásico de vector espacial fue usado para evaluar el modelo y la medida de similitud, investigando sus fortalezas y debilidades. El modelo propuesto ha proporcionado resultados prometedores.

In this paper we present a new model, designated as association graph, to improve document representation, facilitating the ontological dimension. We explain how to create and use this kind of graph. Also, we analyze a document similarity measure based on this representation. A classical vector space model was used to evaluate this model and the similarity measure, investigating their strengths and weaknesses. The proposed model was found to give promising results.

Palabras clave *Key words*

Minería de texto, recuperación de información, inteligencia artificial

Text mining, information retrieval, artificial intelligence

Ernesto Guevara Martínez, *Ingeniero Informático, Centro de Estudios de Ingeniería de Sistemas (CEIS), Instituto Superior Politécnico José Antonio Echeverría, Cujae, Ciudad de La Habana, Cuba*
e-mail:eguevara@ceis.cujae.edu.cu

José E. Medina Pagola, *Licenciado en Cibernética Matemática, Doctor en Ciencias Técnicas, Investigador Auxiliar, Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV), Ciudad de La Habana, Cuba*
e-mail:jmedina@cenatav.co.cu

José Hernández Palancar, *Licenciado en Cibernética Matemática, Doctor en Ciencias Técnicas, Investigador Titular,CENATAV, Ciudad de La Habana, Cuba*
e-mail:jpalancar@cenatav.co.cu

Recibido:mayo del 2006
Aprobado:junio del 2006

INTRODUCCIÓN

En la sociedad moderna, cuando la información comienza a ser un poderoso recurso en la vida diaria, el análisis y reconocimiento de su esencia representan un requerimiento indispensable. Por esta razón, es necesario desarrollar técnicas para descubrir patrones interesantes y comprensibles en los textos, comparar diferentes recursos y resumir grandes cantidades de información.

Un tipo de sistema que requiere estas técnicas, y que ha recibido cierta atención recientemente, es el señalado por Yi Yu Yao como **Sistema de apoyo a la investigación** (RSS - Research Support Systems) y **WRSS** (Web-based RSS),[1] los cuales mejoran las herramientas de búsqueda actuales, ayudando a los científicos a acceder, explorar, evaluar y usar la información almacenada en librerías digitales o en la Web, aumentando la productividad y la calidad.[2]

La **minería de texto**, junto a otras técnicas tales como la **creación de perfiles** (Profiling), el **filtraje cooperativo** (Collaborative Filtering), los **agentes inteligentes**, etc., pudieran ser consideradas para desarrollar estos sistemas. La **minería de texto**, así como muchas otras tareas de procesamiento de textos, es usualmente realizada con representaciones simples del contenido del texto. Sin embargo, la **creación de perfiles,** el **filtraje cooperativo** y los **WRSS** requieren tener en cuenta relaciones semánticas más complejas usualmente expresadas como grafos semánticos.[3]

En este artículo se presenta una propuesta usando **grafos de asociación** como una representación alternativa de documentos, facilitando su dimensión ontológica y mejorando el desarrollo de los **WRSS**.

MINERÍA DE TEXTO

La **minería de texto** pudiera ser definida como un proceso de descubrimiento de patrones interesantes y de nuevo conocimiento en una colección de textos.[4] Por lo tanto, su objetivo es descubrir características tales como regularidades, tendencias, desviaciones, y asociaciones en grandes volúmenes de datos en formas textuales.[5,6]

Durante el preprocesamiento de los documentos, como parte del proceso de **minería de texto**, se obtiene una secuencia de términos distinguidos. Estos suelen considerarse como grupos (**Bags**) de términos, usualmente estructurados mediante modelos vectoriales.[7] En esta representación la secuencia de los términos, o su relación sintáctica, no se analiza, suponiéndose su independencia mutua.

Los valores de esos vectores son considerados como pesos. Una de las formas más comunes de interpretar esos pesos es mediante la **frecuencia de términos** (TF - Term Frequency), absoluta o normalizada, o variaciones de esta.[8]

Diferentes medidas de semejanza entre esos vectores de términos han sido aplicadas. Una de las más conocidas es la del **coseno**, definida como:

$$\cos(d_i, d_j) = (d_i \cdot d_j) / (\|d_i\| \, \|d_j\|) = \sum w_{ir} * w_{jr} / (\sqrt{} w_{ir}^2 \; w_{jr}^2)$$

Donde d_i, d_j son los vectores de los documentos i, j, $\|d_i\|$, $\|d_j\|$ las normas de los vectores, y w_{ir}, w_{jr} los pesos en los vectores d_i, d_j.

DIMENSIÓN ONTOLÓGICA

Aunque generalmente, los términos que aparecen en un documento están interrelacionados y el modelo vectorial ha sido la forma dominante para representar y medir la semejanza entre los documentos, algunos autores consideran este tratamiento como una forma simple de la dimensión ontológica de la información.

Propuestas alternativas al modelo vectorial son los modelos de lenguajes. Estos modelos consideran la probabilidad de ocurrencia de una frase *S* en un lenguaje *M*, indicado por *P*(*S*/*M*). No obstante, las frases son usualmente reducidas a un término, asumidos como unigramas independientes. Ejemplo de este modelo es la divergencia de Kullback-Leibler, definido como:

$$D(d_i \| d_j) = \sum P(t/d_i) \log (P(t/d_i)/P(t/d_j))$$

Otra propuesta es la realizada por Kou y Gardarin.[9] En ella se propone un tipo de modelo de lenguaje, considerando la semejanza entre dos documentos como:

$$\text{sem}(d_i, d_j) = d_i \cdot d_j = \sum_r wir*wjr + \sum_r \sum_{s \neq r} w_{ir} * w_{js} * (t_r \cdot t_s)$$

Donde w_{ir}, w_{js} son los pesos de los términos en los vectores de documento d_i, d_j, y $(t_r \cdot t_s)$ es una correlación a priori entre los términos tr y ts. Estas correlaciones expresan las probabilidades $P(t_r,t_s/M)$ de frases que contienen los términos t_r, t_s en un lenguaje *M*. Además, esa expresión pudiera reducirse a la medida del coseno (normalizada por la longitud de los vectores) si se considera la independencia de los términos y, por esta razón, la correlación $(t_r \cdot t_s)$ es cero. Aunque la propuesta de Kou-Gardarin mejora la limitación de independencia del modelo vectorial, esta considera que dos términos están correlacionados como una tendencia e independiente de los documentos analizados. Tal suposición subvalora la dimensión ontológica intrínseca a todo documento.

En general, puede asumirse que una mejor representación ontológica de la información recuperada y discriminada, mejora la minería realizada sobre los documentos. Además, se espera que una mejor representación deba superar la pobre capacidad de descripción del modelo vectorial.

GRAFOS DE ASOCIACIÓN

Para modelar la relación entre dos términos en un documento será considerada la menor distancia física entre esos términos. Por tanto, dos documentos debieran estar más cercanos entre sí cuanto mayor sea el número de términos comunes y más semejantes sean las distancias físicas más cortas entre ellos. La distancia física entre dos términos t_r y t_s en un documento i, designado por D_{rsi}, pudiera ser definida de formas diferentes. Por ejemplo, pudiera considerarse el número de palabras entre ellas. Sin embargo, en este caso se ignora la relación semántica entre los términos.

En este artículo se asume que una mejor medida de tales relaciones debiera considerar la natural coocurrencia en oraciones y párrafos. Para modelar esta suposición se considerará como el número del párrafo y la oración del término t_r, y análogamente para el término t_s, con lo que se define la distancia física D_{rs}^{i} como:

$$D^{i}_{rs} = \begin{cases} 1 & (t_r = t_s) \vee (n_r = n_s) \\ 2 & (n_r \neq n_s) \wedge (p_r = p_s) \\ |p_r - p_s| + 2 & (n_r \neq n_s) \wedge (p_r \neq p_s) \end{cases}$$

Se considerará que todo término tiene distancia uno consigo mismo.

De acuerdo con la suposición previa, un documento es modelado por un conjunto de relación de términos, los cuales conforman un grafo, donde los nodos son los términos distinguidos y los arcos son sus relaciones, pesados por las distancias. Aunque la relación física, junto a los términos comunes, puede emplearse para evaluar la semejanza entre los documentos, los pesos de los términos distinguidos no deben ignorase en una

medida de semejanza. Para incluir estos valores, el grafo del documento pudiera ampliarse con los pesos de los nodos.

Por lo tanto, un grafo de asociación de un documento puede definirse como el grafo pesado por los nodos, considerando los pesos de los términos distinguidos, y por los arcos, considerando la distancia física entre los términos adyacentes.

Para modelar la fuerza de los arcos se propone el vector A_{rs}^{i} como el peso del arco de los términos relacionados t_r, t_s en un documento i, definido como:

$$A_{rs}^{i} = (P_{ri}, P_{si}) = (w_{ir} / \sqrt{D_{rs}^{i}}, w_{is} / \sqrt{D_{rs}})$$

Donde w_{ir}, w_{is} son los pesos de los términos t_r y t_s en el documento i.

Con lo definido anteriormente, considerando los **grafos de asociación** de los documentos i, j, la semejanza entre estos puede expresarse como:

$$sem(d_i, d_j) = \tfrac{1}{2} \; \|A_{rs}^{i}\| / \sum_{tr,ts \in Tij} \|A_{rs}^{i}\| + \tfrac{1}{2} \sum_{tr,ts \in Ti} \|A_{rs}^{i}\| / \sum_{tr,ts \in Tij} \|A_{rs}^{j}\| + \sum_{tr,ts \in Tj} \|A_{rs}^{j}\|$$

Donde T_i, T_j representan los conjuntos de términos en los **grafos de asociación de los documentos** i, j, respectivamente, y T_{ij} el conjunto de términos comunes ($T_i \cap T_j$). Si T_i o T_j son los conjuntos vacíos, la expresión se define como cero.

La fracción ½ en la fórmula garantiza que tal medida se tenga definida en el intervalo [0,1].

EXPERIMENTO Y ANÁLISIS

Con el propósito de evaluar el modelo propuesto, se ha utilizado la base en español TREC-5 (http://trec.nist.gov). De esta base se tomaron 676 noticias publicadas por AFP durante 1994 y clasificadas en 22 tópicos.

El preprocesamiento de los documentos se realizó con la biblioteca del sistema JERARTOP,[8] la cual utilizó el analizador morfológico MACO+ desarrollado por el Grupo de Procesamiento de Lenguaje Natural de la Universidad de Catalunya, basado en la extensión del modelo estocástico ESGI.[10] Una descripción detallada de tal analizador puede encontrarse en http://www.lsi.upc.es/~nlp.

Se empleó el modelo vectorial clásico para evaluar la propuesta, aplicando la medida del coseno. El peso del término fue calculado como TF (frecuencia del término), normalizado por la frecuencia máxima. Se aplicó el clasificador K-NN tomando como valores de K a 5, 10, 15 y 20, respectivamente, sobre 169 documentos de prueba y 507 documentos de aprendizaje seleccionados de forma aleatoria.

Como puede observarse en la tabla 1, el modelo del **grafo de asociación** supera al del coseno para diferentes valores de K, incluso cuando los desempeños son similares. Esto prueba que el uso de las asociaciones físicas de los términos realmente mejora la efectividad de la categorización. Aunque estos resultados son solo preliminares, estos muestran que el **grafo de asociación** es un buen modelo, pudiendo mejorar el del coseno.

TABLA 1
Resultados del macropromedio

K	Precisión		Recall		F1	
	semC	semG	semC	semG	semC	semG
5	0.773 3	0.791 9	0.632 7	0.673 7	0.637 6	0.644 8
10	0.788 6	0.807 1	0.675 0	0.682 5	0.634 6	0.652 6
15	0.803 9	0.807 7	0.628 0	0.638 1	0.605 8	0.606 0
20	0.827 0	0.828 4	0.553 7	0.642 8	0.545 1	0.618 0
Media	0.798 2	0.808 8	0.622 4	0.659 3	0.605 8	0.630 4

CONCLUSIONES

La mayoría de los métodos actuales de minería de texto utilizan simples representaciones del contenido de los textos, especialmente como vectores de frecuencias de términos. Estas representaciones son relativamente fáciles de construir a partir de los textos, pero no pueden expresar varios detalles de sus significados, presentando una pobre capacidad de descripción. Para lograr una mejor representación del conocimiento contenido en los textos, conveniente en los WRSS, se han propuesto formatos más complejos. En correspondencia con esta idea, este artículo propone el uso de los **grafos de asociación** para representar el conocimiento contenido en los documentos, así como una medida de semejanza, permitiendo la comparación y discriminación de documentos, y su aplicación en diferentes técnicas, tales como los algoritmos de agrupamiento.

Esta propuesta puede ser mejorada en varias direcciones. Una de ellas es aplicándola a otros repositorios de textos, e incluyéndola en otras técnicas de agrupamiento y clasificación.

No obstante, el experimento ha arrojado resultados interesantes. Aunque otras experimentaciones deben ser realizadas; el modelo propuesto muestra resultados prometedores.

REFERENCIAS

1. **YAO, J. T. AND Y. Y. YAO:** *Web-based Information Retrieval Support Systems: Building Research Tools for Scientists in the New Information Age*, Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence (WI 2003), Halifax, Canada, 2003.
2. **XU, J.; Y. HU AND G. MADEY:** *A Research Support System Framework For Web Data Mining*, Proceedings of WI/IAT 2003 Workshop on Applications, Products and Services of Web-based Support Systems, WSS 2003, Halifax, Canada, 2003.
3. **ROJO, A.:** "RA: un agente recomendador de recursos digitales de la Web", Tesis de maestría, Universidad de las Américas, Puebla, México, 2002. URL: http://www.pue.udlap.mx/~tesis/msp/rojo_g_a/.
4. **MONTES GÓMEZ, M.:** "Minería de texto empleando la semejanza entre estructuras semánticas", Tesis Doctoral, CIC-IPN, DF, México, 2002.
5. **MOLINA, F. Y C. LUIS:** *Data Mining: torturando a los datos hasta que confiesen*, UOC, 2002. URL: http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html.
6. **LÓPEZ, V. Y R. AGUILAR:** *Minería de textos y aprendizaje automático en el procesamiento del lenguaje natural*, Departamento de Informática y Automática, Universidad Salamanca, España, 2002.
7. **RAGHAVAN, V. AND S. WONG:** "A Critical Analysis of Vector Space Model for Information Retrieval", *Journal of the American Society on Information Science*, Vol. 37, No. 5, pp. 279-287, 1986.
8. **PONS, A.** "Desarrollo de algoritmos para la estructuración dinámica de información y su aplicación a la detección de sucesos", Tesis Doctoral, University Jaume I, Spain, 2004.
9. **KOU, H. AND G. GARDARIN:** *Similarity Model and Term Association for Document Categorization*, NLDB 2002, LNCS 2553, pp. 223-229, Berlin Heidelberg, Springer-Verlag, 2002.
10. **YANG, Y.:** "An Evaluation of Statistical Approaches to Text Categorization", *Journal of Information Retrieval*, Vol. 1, No. 1/2, pp. 67-88, 1999.