Estudio de la eficiencia del LSA en la agrupación de documentos de texto en español de una colección, en función de la dimensión del espacio reducido.

JUAN CARLOS SOLORIO LEYVA[1], María Estela Romero Fuentes[1], Margarita Torres Figueroa[1], Rosa Isela Ponce del Campo[1] y Luis Ángel Govea Magaña[1]

1 Instituto Tecnológico de La Piedad. juancsol@hotmail.com

El Análisis Semántico Latente (LSA por sus siglas en Inglés) es una técnica del procesamiento de lenguaje natural. Este método ofrece buenos resultados en la recuperación automática de información y la clasificación automática de documentos de texto. En esta técnica, se construye la matriz de frecuencias de términos y documentos de la colección. Luego, se realiza la descomposición de valores singulares de dicha matriz y se reduce la dimensión del espacio, manteniendo un número determinado de valores propios. Con esto se reduce el rango de la matriz. Se ha observado que la eficiencia de los resultados del uso de esta técnica para agrupar documentos de una colección de textos depende de la dimensión del espacio reducido. En este trabajo, se presentan los resultados de las pruebas realizadas para medir la eficiencia del método LSA para una colección de documentos de texto en español en función de la dimensión del espacio reducido, es decir, del número de valores propios que se conservan al hacer la descomposición de valores singulares de la matriz de términos y documentos de la colección. Hemos introducido una mejora al método mediante la creación de documentos guía, que se agregan a la colección. El prototipo que se utilizó para la agrupación de los documentos de la colección fue desarrollado por el equipo de trabajo. Se encontró que el método mejorado tiene muy alta eficiencia al conservar aproximadamente un 22 por ciento de los valores propios.