Cartas

La minería de texto: perspectiva metodológica para la realización de resúmenes documentales

Lic. Y. Mariela del Castillo Zayas[1] **y Lic. Amed Abel Leiva Mederos**[2]

Cita (Vancouver): Castillo Zayas YM, Leiva Mederos AA. La minería de texto: perspectiva metodológica para la realización de resúmenes documentales. Acimed 2007;15(5). Disponible en: **http://bvs.sld.cu/revistas/aci/vol15_5_07/aci14507.htm** [Consultado: día/mes/año].

La Ciencia de la Información ha incorporado a su quehacer algunos de los avances tecnológicos más importantes de las ciencias computacionales que, sin duda, permean las estructuras de representación de esta especialidad. Uno de los campos que más expresiones de cambio ofrece es la representación y organización de la información; y dentro de esta, la indización y el resumen es uno de los estratos que, desde el punto de vista metodológico, necesita más apropiarse de las herramientas lógicas que la inteligencia artificial desarrolla.

Por otra parte, los dominios diversos que interactúan con esta información no poseen el tiempo para la consulta y la asimilación de los grandes volúmenes de información existentes. Este fenómeno no sólo ataña a los usuarios como consultores de información, sino también a los especialistas de la información. Estos, evidentemente, han cambiado sus funciones y actitudes y son cada vez menos lo que se dedican a la labor de extracción o de realización de resúmenes. La minería de textos, en estos momentos, ofrecen perspectivas, desde un punto de vista elemental, que podrían explotarse por los servicios de las instituciones de información a la vez que sus presupuestos metodológicos pudieran relacionarse con el desarrollo de nuevos métodos para resumir información en el ciberespacio.

Los cambios que han ocurrido en el entorno de Internet y sobre todo en los documentos que circulan por esta red exigen buscar nuevas formas para resumir los grandes volúmenes de información que se generan diariamente y que incorporan nuevos elementos como voz, imagen, sonido y movimientos. En este entorno, los resúmenes y los servicios de resúmenes como instrumentos de condensación de la información relevante, adquieren un mayor valor.

Minería de texto

Muchos autores coinciden en que la minería de texto o *Text Mining* es una herramienta que proviene del área del procesamiento automático de textos y que permite localizar y extraer la información más significativa y esencial de los documentos, así como información y conocimiento implícito y oculto en grandes *corpus* textuales electrónicos, estructurados o no estructurados, como mensajes de correos electrónicos, discursos, artículos, entre otros. Debido a esto, en ocasiones se asocia con el espionaje.

Funciona a partir de una telaraña semántica, que tiene como objetivo construir toda una estructura de metadatos, información sobre la estructura y significado de los datos almacenados e incluirlos en los documentos de forma que sean navegables, identificables y entendibles por las máquinas, por lo que es una herramienta eficaz para gestionar el conocimiento. “S e enfoca en el descubrimiento de patrones interesantes y nuevos conocimientos en un conjunto de textos, es decir, su objetivo es descubrir tendencias, desviaciones y asociaciones en la gran cantidad de información textual disponible”,[1] e s decir, facilita realizar análisis y se erige como un área emergente de la minería de datos. Elimina la información duplicada y detecta información similar o relacionada con la existente. La minería de textos utilizada en las Ciencias de la Información pudiera explotarse como herramienta en los nuevos métodos de resumen porque permite la decodificación y análisis del lenguaje natural e interfaces en la lengua materna de cada dominio, traducción automática, procesamiento de voz, generación de texto, etcétera.[2] Todas estas cualidades de la minería de texto son la razón que fundamenta la propuesta de esta herramienta como perspectiva metodológica para la realización de resúmenes documentales.

Las perspectivas metodológicas de la minería de texto aplicables en las instituciones de información son disímiles, porque su rango de acción no sólo se desarrolla en el trabajo con el texto, sino que además explora otros sectores como el procesamiento de voz, decodificación de imágenes, construcción de *corpus* documentales, representación y graficación de términos mediante herramientas de ponderación asociadas, entre otros.

Algunos sistemas que se emplean para hacer minería de texto son: *SMART*, *ANES*, *SIM-SUM*, *KADS*, *Classifier*, *Parse* r, *Text Classifier*, *Text Recognizer*, la plataforma ILC, *NEURODOC*, *SDOC*, *HENOCH*, algunos basados en inteligencia artificial, entre otros. Todos estos sistemas permiten extraer la información relevante de un documento, agregan y comparan información automáticamente, clasifican y organizan los documentos según su contenido y organizan los depósitos para la búsqueda y recuperación de la información, pero la elección del sistema que permitirá hacer minería de texto estará determinada por la misión, visión y objetivos de la institución de información, así como las tecnologías disponibles para su implementación.

Hacer un resumen automático a partir de la extracción de palabras clave o frases significativas del texto produce como resultado un resumen de muy baja calidad, con dificultades desde el punto de vista lingüístico (sinonimia, polisemia, anáfora, etc.). Tradicionalmente, su producción se ha basado en métodos estadísticos y técnicas de probabilidades, las cuales no aportan ningún nivel de entendimiento de los conceptos y términos. La capacidad de entender el lenguaje humano está en terreno de la lingüística. Sus principales dificultades estriban en las técnicas léxico-sintácticas de selección, en las actividades lógico-semánticas de interpretación y en las tareas pragmático-documentales de producción. Algunos especialistas en la materia ven la solución de este problema en los sistemas expertos de inteligencia artificial, porque con solo analizar las

dificultades que presentan se hace evidente que los sistemas actuales en general aún no están preparados para el reto que implica la producción de resúmenes automáticos de alta calidad.

CONSIDERACIONES FINALES

A pesar de todos los intentos que se han realizado en esta área, aún faltan esfuerzos en pos de lograr que el estudio de estos sistemas esté soportado desde una óptica lingüista, es decir, que se orienten a entender la forma de pensamiento humano que es su principal aspiración.

Se han estudiado poco las características físicas, intelectuales y operativas de estas nuevas formas de representación, así como sus complementos: sonido, imágenes fijas y en movimiento, etcétera.

Los métodos automáticos no logran proporcionar resúmenes con igual calidad que los tradicionales, pero sí son eficaces para determinados contextos.

La información de origen connotativo que esté presente en los documentos, podrá interpretarse por el hombre, pero no por un sistema, lo que limita en gran medida la recuperación de información.

La minería de textos es una forma más de enfrentar el problema de la representación y por ende de la recuperación de información relevante y pertinente un ángulo diferente, pero tampoco ofrece la solución definitiva.

Se pretende que la minería de textos se base no sólo en la detección de palabras clave, sino que además emplee representaciones que consideren más tipos de elementos textuales, como grafos conceptuales para representar el contenido de los textos y llegar a un nivel más descriptivo.

REFERENCIAS BIBLIOGRÁFICAS

1. Montes y Gómez M. Minería de texto: un nuevo reto computacional. Disponible en: **http://ccc.inaoep.mx/~mmontesg/publicaciones/2001/MineriaTexto-md01.pdf** [Consultado: 3 de marzo del 2007].
2. Gelbukh A, Bolshakov I. Avances y perspectivas de procesamiento automático de lenguaje natural: cuento de una máquina parlante. Disponible en: **http://www.gelbukh.com/CV/Publications/2000/IPN-Proc-Leng-Nat.htm** [Consultado: 3 de marzo del 2007].

Recibido: 30 de marzo del 2007. Aprobado: 12 de abril del 2007.

Lic. *Y. Mariela del Castillo Zayas*. Centro de Información y Documentación. Escuela de Hotelería y Turismo Playas del Este. Calle 462 e/ 5ta y 7ma, Guanabo. CP 19120. La Habana, Cuba. Correo electrónico: **mariela@ehtpe.co.cu**

[1]Licenciada en Bibliotecología y Ciencia de la Información. Centro de Información y Documentación. Escuela de Hotelería y Turismo Playas del Este. Cuba.
[2]Licenciado en Bibliotecología y Ciencia de la Información. Universidad Central de Las Villas. Cuba.

Ficha de procesamiento

Términos sugeridos para la indización

Según DeCS[1]

RESUMEN E INDIZACIÓN; PROCESAMIENTO AUTOMATIZADO DE LA INFORMACIÓN.

ABSTRACTING AND INDEXING; AUTOMATICA DATA PROCESSING.

Según DeCI[2]

PROCESAMIENTO AUTOMATIZADO DE LA INFORMACIÓN; PROCESAMIENTO DE LA INFORMACIÓN, RESÚMENES.

AUTOMATICA DATA PROCESSING; INFORMATION PROCESSING; ABSTRACTS.

[1]BIREME. Descriptores en Ciencias de la Salud (DeCS). Sao Paulo: BIREME, 2004.

Disponible en: **http://decs.bvs.br/E/homepagee.htm**

[2]Díaz del Campo S. Propuesta de términos para la indización en Ciencias de la Información. Descriptores en Ciencias de la Información (DeCI). Disponible en: **http://cis.sld.cu/E/tesauro.pdf** }