Biblioteca Universitaria
ISSN: 0187-750X
public@dgb.unam.mx
Universidad Nacional Autónoma de México
México

Contreras Barrera, Marcial
Minería de texto: una visión actual
Biblioteca Universitaria, vol. 17, núm. 2, julio-diciembre, 2014, pp. 129-138
Universidad Nacional Autónoma de México
Distrito Federal, México

Minería de texto: una visión actual

Text mining: a current view

Marcial Contreras Barrera*

Resumen

En la actualidad, debido a las actividades sociales, económicas y académicas, en conjunto con la utilización de las Tecnologías de Información y Comunicación (TIC), la producción de conocimiento y de información digital se genera cada vez con mayor rapidez y, en ocasiones, se acumula en enormes volúmenes de información en espera de un procesamiento adecuado. Áreas como la minería de datos y la minería de texto son utilizadas en el procesamiento automatizado de información, las cuales ofrecen la posibilidad de explorar grandes cantidades de datos o texto, estructurados y no estructurados, para buscar patrones, encontrar relaciones y extraer conocimiento; por lo que el propósito del presente artículo es mostrar las características y aplicación de la minería de texto como área emergente en el procesamiento de información digital.

Palabras clave: Minería de texto, información digital, procesamiento.

Abstract

Nowadays, due to the social, economic and academic activities, together with the use of information and communication technologies (ICT), knowledge production and digital information is generated increasingly faster and sometimes accumulates in huge volumes of information, pending for proper processing. Areas such as data mining and text mining are used in the automated processing of information, which provide the ability to scan large amounts of data or text, structured and unstructured, to look for patterns, find relationships and extract knowledge. The purpose of this article is to show the features and applications of text mining as an emerging area of digital information processing.

Keywords: Text mining, digital information, information processing

* Departamento de Producción, Dirección General de Bibliotecas (DGB). Universidad Nacional Autónoma de México (UNAM). Biblioteca Central, Ciudad Universitaria, 04510, México, D. F., México. Correo electrónico: marcial@dgb.unam.mx

Introducción

La producción e incremento del volumen de información digital en los últimos años ha sido de forma exponencial, de tal suerte que en la actualidad es necesario contar con equipos de cómputo de alto rendimiento con capacidad de almacenamiento, desde varios gigabytes hasta varios hexabytes[1] de información, y al mismo tiempo con la infraestructura de comunicaciones adecuada, anchos de banda del orden de los Mbits/s hasta los Gbits/s[2]; ejemplo de ello son los estudios realizados por la empresa EMC[3], en donde muestra que la información digital mundial se duplica cada dos años. Para el año 2011 la empresa tenía calculado un volumen de información mundial de 1,8 zettabytes[4] y estima que para 2020 el mundo va a generar 50 veces la cantidad de información y 75 veces el número de "contenedores de información", mientras que el personal de TI (tecnología de información) para gestionar crecerá menos de 1.5 veces.

La mayoría de los datos que recopilan, crean y gestionan las compañías hoy en día están desestructurados; repartidos en documentos de procesadores de texto, hojas de cálculo, imágenes y vídeos que no se pueden interpretar fácilmente[5] por lo que es necesario el procesamiento automatizado de éstos.

En lo referente a la producción de documentos científicos, según las estadísticas de MEDLINE se estima que existen alrededor de 13 000 a 14 000 títulos biomédicos publicados actualmente en todo el mundo de los cuales 5 300 títulos están indexados y se incluyen en la base de datos MEDLINE.[6]

Un área que ha experimentado este incremento exponencial ha sido el de la producción de información,[7] que junto con las tecnologías de información han dado lugar a complejos sistemas de gestión y provisión de información para diversas tareas que se desarrollan en la sociedad contemporánea. En especial, la educación superior precisa contar con eficaces herramientas de información para la enseñanza, el aprendizaje y la investigación.

A partir de la segunda mitad del siglo XX se inicia la automatización de información dando lugar a la creación de bases de datos de tipo referencial, que proporcionaban a los usuarios datos básicos de un documento como el título, autor, la fuente, palabras clave, entre otros. Con el desarrollo de la informática mejoran las técnicas de indexación, organización y clasificación y surgen nuevos servicios y documentos electrónicos tales como libros, revistas, periódicos, tesis, directorios, diccionarios electrónicos y bases de datos de texto completo que permiten a los usuarios la obtención automatizada del documento.

Durante más de una década, Don Swanson ha argumentado que es posible recuperar nueva información que se deriva de colecciones de textos, debido a que los expertos sólo leen una pequeña parte de lo que se publica en sus áreas de conocimiento y no son conscientes de la evolución de los campos relacionados.[8] Por lo tanto, debería ser posible encontrar los vínculos entre la información útil en la literatura de las diferentes áreas del conocimiento.

Minería de texto

La proliferación del uso de dispositivos computacionales y de comunicación para la producción de información

---

1 10^{18} bytes = 1 millón de millones de bytes

2 Transmisión de bits por segundo, 106 bit = 1 000 000 bit/s = 1 Mbit/s (un megabit o un millón de bits)

3 *Digital Universe* [en línea]. EMC2. <http://www.emc.com/leadership/programs/digital-universe.htm>

4 Un *zettabyte* es una unidad de almacenamiento de información cuyo símbolo es el ZB, equivale a 10^{21} bytes.

5 BEATH, Cynthia, Becerra-Fernández, Irma, Ross, Jean, Short, James. El valor de la explosión de la información [en línea]. *IQ Intelligence quarterly: diario de análisis avanzado,* 2013, p. 8-11. <http://www.sas.com/offices/europe/spain/IQ/Q113/IQBigAnalytics_GrandesOportunidades.pdf>

6 *Journal Selection for medline Indexing at NLM* [en línea]. U.S. National Library of Medicine. <http://www.nlm.nih.gov/pubs/factsheets/j_sel_faq.html#a15>

7 Documentos, reportes, e-mails y páginas web.

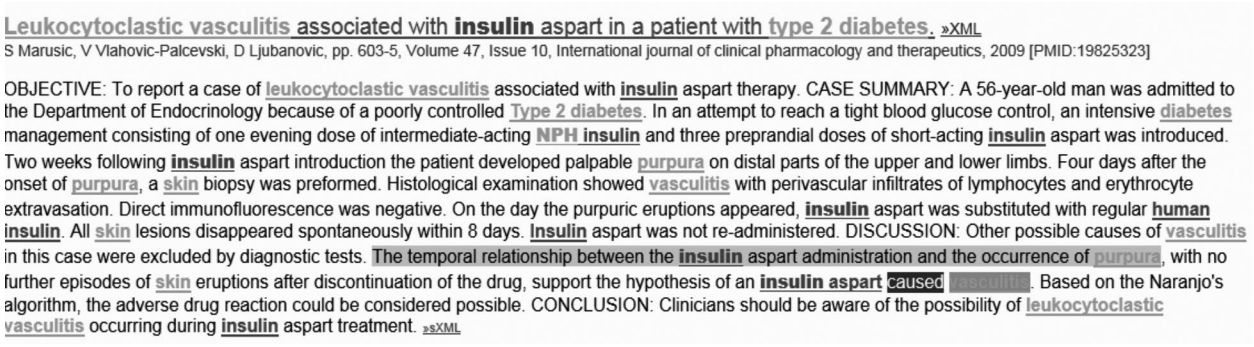
8 SWANSON, Don R. Complementary structures in disjoint science literatures.

digital, y en particular en la producción de documentos textuales, ha generado la necesidad de desarrollar métodos, algoritmos y sistemas capaces de realizar el procesamiento automatizado de datos textuales estructurados, semi-estructurados y no estructurados para su organización y consulta, y con ello el surgimiento de áreas de estudio de la información como la minería de texto.

La minería de texto es un área de investigación del procesamiento automático de la información. Se define como el proceso de descubrimiento de patrones interesantes y nuevos conocimientos en una colección de textos, es decir, es el proceso encargado del descubrimiento de conocimientos que no existían explícitamente en ningún texto de la colección, pero que surgen de relacionar el contenido de varios de ellos.[9] Para Witten, la minería de texto es el proceso de analizar escritos o conjuntos de enunciados para extraer información que resulta útil para propósitos particulares.[10] Según Sukanya, la minería de texto es un campo interdisciplinario joven el cual se basa en la recuperación de información, minería de datos, aprendizaje de máquina, estadística y lingüística computacional.[11]

Para Sánchez[12] la minería de texto hace referencia al descubrimiento no trivial potencialmente útil de conocimiento partiendo de una colección de documentos de texto no estructurado. Y puede realizarse una analogía con la minería de datos, encargada de descubrir conocimiento en bases de datos.

Finalmente, la minería de texto se puede ver como un área que se encarga del estudio de la información digital y en particular de los documentos textuales, con el objetivo de descubrir tendencias, patrones, desviaciones y asociaciones de una colección de textos, para –en última instancia– pasar al descubrimiento de conocimiento en considerables cantidades de información no estructurada. Como ejemplo de lo anterior, la figura 1 muestra la relación que existe entre la insulina y la vasculitis



Leukocytoclastic vasculitis associated with insulin aspart in a patient with type 2 diabetes. »XML

S Marusic, V Vlahovic-Palcevski, D Ljubanovic, pp. 603-5, Volume 47, Issue 10, International journal of clinical pharmacology and therapeutics, 2009 [PMID:19825323]

OBJECTIVE: To report a case of leukocytoclastic vasculitis associated with insulin aspart therapy. CASE SUMMARY: A 56-year-old man was admitted to the Department of Endocrinology because of a poorly controlled Type 2 diabetes. In an attempt to reach a tight blood glucose control, an intensive diabetes management consisting of one evening dose of intermediate-acting NPH insulin and three preprandial doses of short-acting insulin aspart was introduced. Two weeks following insulin aspart introduction the patient developed palpable purpura on distal parts of the upper and lower limbs. Four days after the onset of purpura, a skin biopsy was preformed. Histological examination showed vasculitis with perivascular infiltrates of lymphocytes and erythrocyte extravasation. Direct immunofluorescence was negative. On the day the purpuric eruptions appeared, insulin aspart was substituted with regular human insulin. All skin lesions disappeared spontaneously within 8 days. Insulin aspart was not re-administered. DISCUSSION: Other possible causes of vasculitis in this case were excluded by diagnostic tests. The temporal relationship between the insulin aspart administration and the occurrence of purpura, with no further episodes of skin eruptions after discontinuation of the drug, support the hypothesis of an insulin aspart caused vasculitis. Based on the Naranjo's algorithm, the adverse drug reaction could be considered possible. CONCLUSION: Clinicians should be aware of the possibility of leukocytoclastic vasculitis occurring during insulin aspart treatment. »sXML

Figura 1: Relación entre insulina y vasculitis.

Para el logro de los objetivos la minería de texto utiliza diversos métodos y tecnologías como: la recuperación de información, métodos estadísticos y matemáticos, procesamiento de lenguaje natural, métodos de clasificación y agrupamiento de datos, entre otros.

En las instituciones bibliotecarias o unidades de información, las técnicas de minería de texto tienen mucho que ofrecer a las bibliotecas digitales y a sus usuarios, por esta razón existen trabajos desarrollados para integrar las técnicas de minería de texto a los sistemas encargados de gestionar las colecciones de las bibliotecas digitales, por ejemplo el sistema de cómputo Greenstone utilizado en la gestión de bibliotecas digitales desarrollado en Nueva Zelanda.

Greenstone organiza los documentos y los pone disponibles en Internet; además de permitir la búsqueda en texto completo cuenta con módulos para realizar la

[9] Swanson, Don R., *idem.*

[10] Text mining in a digital library. Ian H. Witten, Katherine J. Don, Michael Dewsnip, Valentin Tablan.

[11] Sukanya, M., Biruntha, S. Techniques on Text Mining.

[12] Sánchez, D., Martín-Bautista, M. *Un enfoque deductivo para la minería de texto* [en línea]. <http://www.softcomputing.es/estylf08/es/2006-XIII%20Congreso/articulos/40.pdf>

minería de texto a través de la extracción de acrónimos y su definición a partir del texto completo y de la colección, agregando esta información a su base de datos como metadatos. El sistema Greenstone contiene un módulo de minería de texto el cual realiza la extracción de frases clave y las adiciona como metadatos.[13]

Otro ejemplo sobre la aplicación de minería de texto es el software que utiliza métodos de agrupamiento o *clustering* para la organización de documentos en la biblioteca digital o en colecciones individuales, el cual agrupa los documentos en diferentes grupos temáticos basados en su contenido[14]; el resultado del agrupamiento se puede visualizar con el empleo del software libViewer[15] como se muestra en la figura 2.

Figura 2. Visualización de la biblioteca digital de la sección de economía y deportes.

Etapas de la minería de texto

La minería de texto, como proceso, establece una serie de pasos necesarios para el procesamiento de texto y su consecuente extracción de datos/conocimiento;[16] dependiendo del objetivo de la minería de texto los pasos a seguir varían. A continuación se describen de manera general los más relevantes.

El primer paso es determinar el propósito de estudio de la minería de texto; como ejemplo, en el caso de textos de biología se puede identificar y etiquetar entidades biológicas, extracción y normalización de sinónimos, homónimos y abreviaturas, identificación de entidades biológicas, generación de hipótesis, etcétera.

El segundo paso es recolectar, identificar, recoger y validar información; en esta fase se realiza la recuperación de información (IR) en la cual se buscan e identifican las fuentes más relevantes para el objetivo de estudio de la minería de texto. Luego, se deben recopilar los documentos detectados en el mejor formato, se seleccionan, se evalúa su relevancia y se realizan las anotaciones necesarias.[17] En esta etapa se cuenta con el conjunto de documentos necesarios para la realización de la minería de texto.

En el tercer paso se realiza el procesamiento de texto con la finalidad de eliminar información que no ayuda al propósito de la minería de texto, realizándose algunas de las siguientes acciones: análisis léxico, tratamiento y separación de palabras vacías (artículos, preposiciones, conjunciones), tratamiento de términos flexionados (términos relacionados morfológicamente, variaciones de género, número o tiempo verbal), tratamiento de palabras compuestas, normalización de palabras, obtención de las raíces de las palabras y etiquetado de palabras, además de corregir algunos problemas que presenten los documentos como: los problemas de formato, polisemia, homonimia, sinonimia. Esta fase exploratoria se basa principalmente en lingüística computacional (análisis morfológico y sintáctico), además de algoritmos informáticos.[18] El objetivo de esta fase es facilitar la selección de características deseadas para identificar palabras clave,

[13] Text mining in a digital library, *op. cit.*

[14] Gupta, Vishal, Lehal, Gurpreet S. A Survey of Text Mining Techniques and Applications.

[15] Rauber, Andreas, Merkl, Dieter. Text Mining in the somlib Digital Library System: The Representation of Topics and Genres.

[16] Sukanya, M., Biruntha, S., *op. cit.*

[17] *Extraction des informations et des connaissances* [en línea]. <http://touriaelouahabi.wordpress.com/text-mining/principes-et-concepts-text-mining/>

[18] Abbott, Dean. *Introduction to Text Mining* [en línea]: *Virtual Data Intensive Summer School,* July 10, 2013. <http://www.vscse.org/summerschool/2013/Abbott.pdf>

identificación de entidades biológicas, individuos, organizaciones, lugares, oraciones, conceptos, etcétera.

En el cuarto paso se realiza la extracción y análisis de clases, relaciones, asociaciones o secuencias, con el fin de encontrar evidencias de conceptos y de estructuras existentes. En esta etapa los documentos se pueden representar a través del modelo del espacio vectorial,[19] en donde cada documento es modelado como un vector de dimensión *n* y es representado de la siguiente manera $D_i = (d_{i1}, d_{i2}, ..., d_{it})$ donde cada *dij* representa el número de repeticiones de la palabra en el documento, también los datos obtenidos en esta etapa son representados en alguna estructura informática que facilita su análisis; las estructuras representan las relaciones entre las entidades de un mismo tipo de datos, palabras o conceptos clave, documento-términos, términos-autores, interacción de proteínas, etcétera.

En el último paso se presentan los resultados a través de resúmenes, marcados de texto, relaciones, taxonomías y visualización para su interpretación. En esta etapa también se puede almacenar la información procesada en bases de datos para su recuperación posterior. La figura 3 muestra las etapas del proceso de minería de texto en el área biomédica para el reconocimiento de entidades.

Existe una diferencia entre recuperación de información y minería de texto. La recuperación de información es el primer paso en la minería de texto, es decir, una vez que ya se localizó un conjunto de documentos éstos sirven de base en los procesos de la minería de texto para extraer información e identificar patrones, lo que lleva al descubrimiento de nueva información con el objetivo de encontrar información que antes sólo pudo haber sido descubierta por la lectura de un gran número de documentos.

Algunos de los métodos utilizados en esta área emergente para el procesamiento de documentos textuales son: métodos de aprendizaje de máquina, procesamiento del lenguaje natural, redes neuronales artificiales,

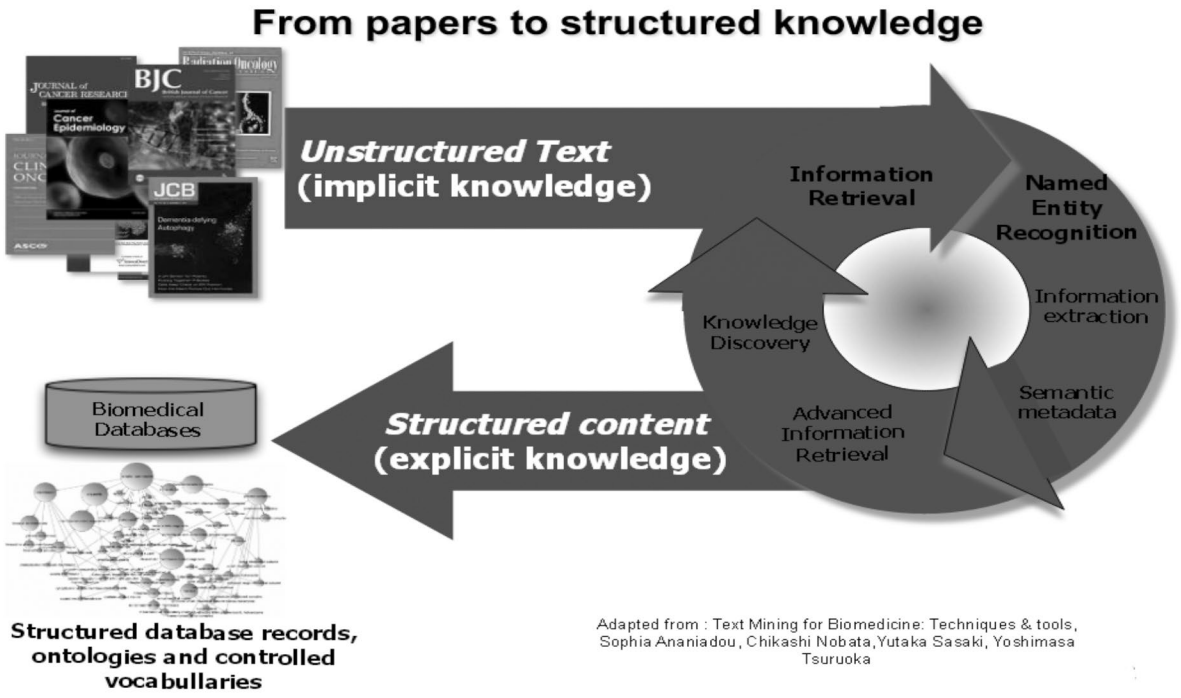


Figura 3: Etapas del proceso de minería de texto.[20,21]

[19] Salton, Gerard. *Automatic text processing: The transformation, analysis, and retrieval of information by computer.*

[20] *Value and benefits of text mining* [en línea]. Jisc. <http://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining>

[21] Krallinger, Martin. *Current trends in biomedical text mining: chemicals, drugs and mutations* [video]

extracción de la información, métodos matemáticos, modelo Bayesiano, modelos probabilísticos, modelo del espacio vectorial, Indexación Latente Semántica (LSI), entre otros. Algunos de los algoritmos de minería de texto empleados para clasificación, categorización y agrupación de información son los siguientes: cuantización vectorial, K-medias y Redes Neuronales Artificiales (RNA). Dentro de este último grupo se encuentran las RNA-Kohenen, Self-Organizing Map (SOM), Learning Vector Quantization (LVQ), entre otras.[22] Todos utilizan un método competitivo, pero unos son no supervisados y otros son supervisados.

Aplicaciones de la minería de texto

Las técnicas de minería de texto se pueden utilizar para procesar diferentes tipos de documentos textuales digitales, por ejemplo en biomedicina, donde los volúmenes de información sobre temas específicos hacen que sea imposible leer todos los artículos por cualquier investigador y mucho menos estudiar los artículos relacionados. La figura 4 muestra algunas de las relaciones que se pueden establecer en el área de la biología.

Un ejemplo particular de la aplicación de los sistemas de minería de texto en el análisis de un documento se muestra en la figura 5, en donde el texto sombreado representa los posibles términos representativos del documento (los términos pueden estar formados por más de una palabra).

Otra de las aplicaciones es en el área médica, donde Swanson analizó artículos de la base MEDLINE para encontrar la relación existente entre la migraña y la deficiencia del magnesio; estos datos fueron validados

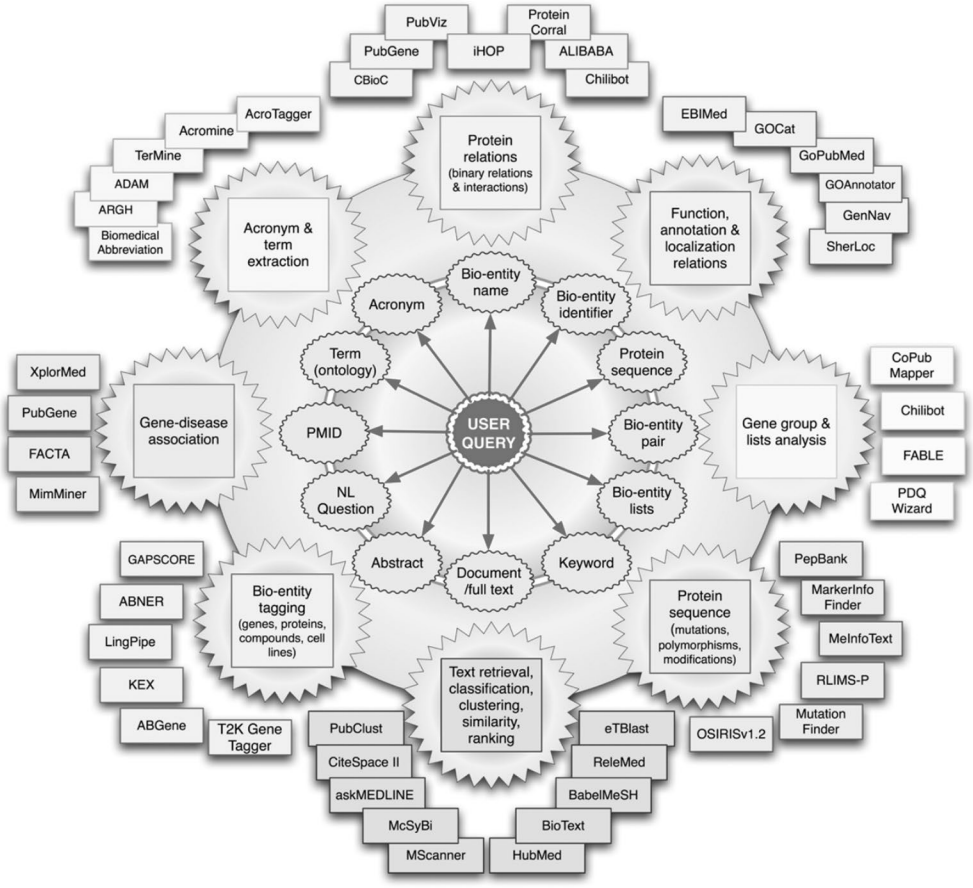
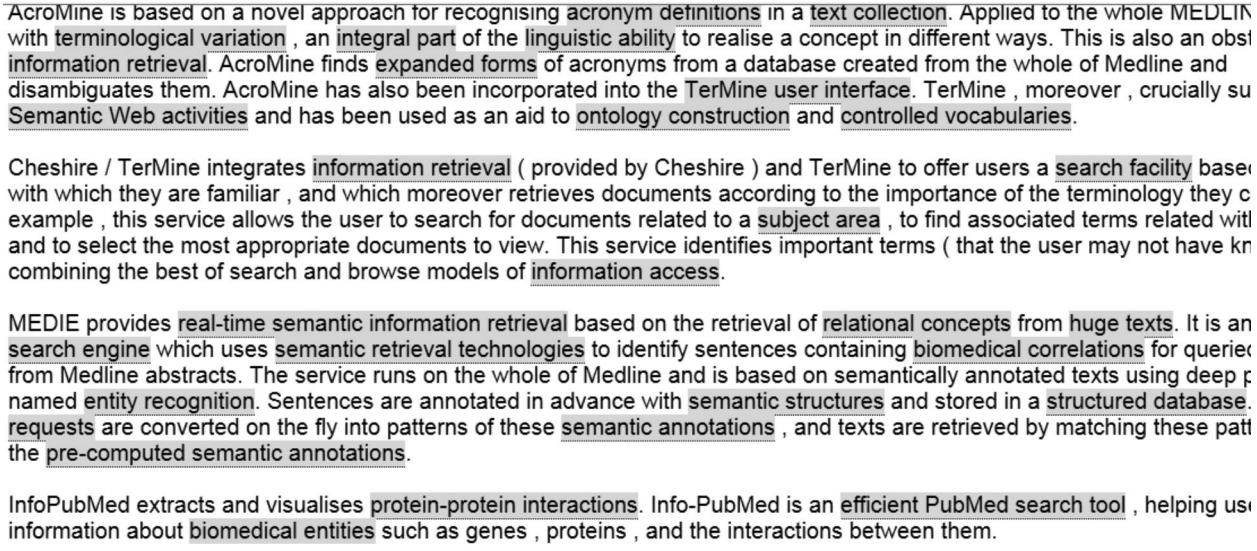


Figura 4: Minería de texto en biología. [23]

[22] HAYKIN, Simon. *Neuronal Networks: A Comprehensive Foundation.*

[23] KRALLINGER, Martin, *op. cit.*


AcroMine is based on a novel approach for recognising acronym definitions in a text collection. Applied to the whole MEDLIN with terminological variation , an integral part of the linguistic ability to realise a concept in different ways. This is also an obst information retrieval. AcroMine finds expanded forms of acronyms from a database created from the whole of Medline and disambiguates them. AcroMine has also been incorporated into the TerMine user interface. TerMine , moreover , crucially su Semantic Web activities and has been used as an aid to ontology construction and controlled vocabularies.

Cheshire / TerMine integrates information retrieval (provided by Cheshire) and TerMine to offer users a search facility base with which they are familiar , and which moreover retrieves documents according to the importance of the terminology they c example , this service allows the user to search for documents related to a subject area , to find associated terms related witl and to select the most appropriate documents to view. This service identifies important terms (that the user may not have kn combining the best of search and browse models of information access.

MEDIE provides real-time semantic information retrieval based on the retrieval of relational concepts from huge texts. It is an search engine which uses semantic retrieval technologies to identify sentences containing biomedical correlations for queried from Medline abstracts. The service runs on the whole of Medline and is based on semantically annotated texts using deep p named entity recognition. Sentences are annotated in advance with semantic structures and stored in a structured database. requests are converted on the fly into patterns of these semantic annotations , and texts are retrieved by matching these patt the pre-computed semantic annotations.

InfoPubMed extracts and visualises protein-protein interactions. Info-PubMed is an efficient PubMed search tool , helping us information about biomedical entities such as genes , proteins , and the interactions between them.

Figura 5: Extracción de términos por métodos de minería de texto.[24]

posteriormente de manera experimental.[25,26] Además, la minería de texto se ha utilizado para el descubrimiento de fármacos, la toxicología predictiva, la inteligencia competitiva, la búsqueda de patentes, etcétera.

Max Haeussier, un investigador biólogo de la Universidad de California, utilizó documentos científicos de la base de datos de Elsevier para extraer datos sobre ADN en cerca de 3 millones de artículos con el fin de realizar el mapa del genoma humano.

En el área biomédica la minería de texto es aplicada en las siguientes tareas: reconocimiento de compuestos químicos y medicamentos, aplicación en la extracción de drogas y sus efectos, para determinar reacciones enzimáticas, en la clasificación de las mutaciones y las proteínas del cáncer, en la detección de especies u organismos.[27]

Otros proyectos, como las investigaciones sobre el cerebro humano desarrolladas por el consorcio europeo integrado por investigadores que trabajan en el proyecto cerebro humano, están utilizando la interfaz de Elsevier para realizar los trabajos de minería de texto. Un ejemplo más del uso de los artículos de Elsevier en la aplicación de la minería de texto es el proyecto del estudio de la fisiología de la neurona.[28]

También se puede emplear para: la inteligencia del gobierno y las agencias de seguridad tratando de reconstruir las advertencias terroristas y otras amenazas de seguridad, el monitoreo en redes sociales, la inteligencia en los negocios, el análisis sentimental; este último se refiere a la exploración de texto para identificar y extraer actitudes del escritor de un texto, identificar el estado emocional del autor o la intención emocional de comunicación. La tarea del análisis sentimental es clasificar un documento o una oración para identificar características positivas, negativas o neutras y también es utilizado para determinar estados emocionales como el enojo, la tristeza y la felicidad. El análisis sentimental puede ser aplicado en la Web, las redes sociales y los blogs.[29]

Debido a la importancia que ha adquirido la minería de texto algunas instituciones académicas hacen un esfuerzo para que esta área se desarrolle, como es el caso del Centro Nacional de Minería de Texto (NaCTeM, por sus siglas en inglés), el cual es operado por la Universidad de Manchester en estrecha colaboración con la Universidad

[24] Ejemplo generado con el software Termine: sitio http://www.nactem.ac.uk/software/termine/#form

[25] Swanson, Don R., *op. cit.*

[26] Smalhaiser, Neil, Swanson, D. Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease.

[27] Krallinger, Martin, *op. cit.*

[28] Noorden, Richard Van. Elsevier opens its papers to text-mining.

[29] *Sentiment analysis* [en línea]. Wikipedia, The Free Encyclopedia. <http://en.wikipedia.org/w/index.php?title=Sentiment_analysis&oldid=625073725>

de Tokio, apoyado con fondos públicos. Este centro tiene como objetivo la promoción de la investigación de las tecnologías de minería de texto y su uso; ofrece herramientas de software para algunas tareas de minería de texto, seminarios, conferencias, tutoriales, demostraciones y publicaciones.[30] En el sector privado también se están realizando investigaciones, muestra de ello son las compañías de fármacos que utilizan la minería de texto para reducir los costos de sus investigaciones.[31]

Además, empresas como Elsevier han puesto al servicio de los científicos más de 11 millones de artículos de investigación disponibles en línea, para ser utilizados en tareas de minería de texto, para extraer hechos y relaciones de éstos. En tiempos pasados la empresa no permitía hacer una recolección de los artículos publicados en su sitio, impidiendo así la aplicación de la minería de texto, muestra de ello es el caso de Max Haeussier, de la Universidad de California, que invirtió más de tres años para conseguir el permiso para extraer datos sobre el ADN. Es por eso que el pasado 26 de enero de 2014, en la conferencia de la American Library Association en Philadephia, Pennsylvania, Elsevier anunció que instituciones académicas y de investigación pueden utilizar la interfaz de Elsevier (API) para descargar documentos en formato XML de forma masiva, con un máximo de 10 000 artículos por semana, y pueden ser utilizados libremente para la minería de texto siempre y cuando las instituciones académicas y de investigación establezcan un acuerdo legal. Los acuerdos incluyen que los investigadores pueden publicar sus productos de minería de texto mientras no sea para uso comercial y pueden incluir máximo 200 caracteres del texto original y deben hacer referencia al contenido original.[32]

Software de minería de texto

En la actualidad existe una amplia gama de software para realizar minería de texto, tanto de tipo comercial como de acceso libre; dentro del tipo de acceso libre se mencionan los siguientes: Gate, con capacidad de resolver problemas de procesamiento de texto; Carrot, utilizado para organizar de forma automatizada documentos en categorías temáticas; RapidMiner, utilizado para el aprendizaje de máquina, minería de datos y texto. De tipo comercial son: SAS Text Miner, para extraer información de datos no estructurados; Clearforest, utilizado para el análisis de texto y procesamiento de lenguaje natural; Autonomy, realiza minería de texto, categorización y clasificación de documentos; SPSS, dentro de sus funciones realiza minería de texto y procesamiento de lenguajes natural; Lexalytics, utilizado para procesamiento de texto, procesamiento de lenguaje natural y extracción de entidades. El URL de los desarrolladores de minería de texto se muestra en la tabla 1.

Tabla 1: Software comercial para minería de texto[33]

Vendedor	Sitio Web	Producto
Autonomy	http://www.autonomy.com/products/	IDOL Server, Retina
Clearforest	http://www.clearforest.com/solutions.html	ClearForest Text Analysis Suite
SAS	http://www.sas.com/en_us/software/analytics/text-miner.html	SAS Text Miner
IBM Text Analytics	http://www-ibm.com/software/ebusiness/jstart/textanalytics	IBM
SPSS	http://www-01.ibm.com/software/analytics/spss/	LexiQuest, Clementine
Lexalytics	http://www.lexalytics.com/web-demo	Text Analytics Software
Gate	http://gate.ac.uk/download/	Gate
Carrot	http://project.carrot2.org/index.html	Carrot Search
RapidMiner	http://sourceforge.net/projects/rapidminer/files/2. Extensions/Text Processing	RapidMiner

[30] *The National Centre for Text Mining* [en línea]. <http://www.nactem.ac.uk/>

[31] BELSKY, Gary. Why Text Mining May Be The Next Big Thing [en línea]. *Technology & Media,* 20 de Marzo de 2012. <http://business.time.com/2012/03/20/why-text-mining-may-be-the-next-big-thing/>

[32] NOORDEN, Richard Van, *op. cit.*

[33] Tapping the Power of Text Mining. Weiguo Fan, Linda Wallace, Stephanie Rich, Zhongju Zhang.

La diferencia en esta gama de software es que algunos realizan extracción de información, seguimiento de temas, resúmenes, categorización, agrupamiento, respuesta a preguntas, extracción de términos.[34] Por otra parte, algunos pueden ser utilizados a través de la Web y otros requieren instalación a nivel local. También existen desarrollos de sistemas de cómputo en las universidades cuya finalidad es poder programar los algoritmos de minería de texto, de acuerdo al tipo de información y a la necesidad particular de procesamiento de información.

Como podemos ver, la aplicación de la minería de texto es una herramienta de importancia relevante en la actualidad. Independientemente de la disciplina en la cual se aplique, nos permite localizar y descubrir información oculta en grandes volúmenes de información de manera automatizada, permitiendo analizar documentos con la finalidad de obtener nueva información para generar nuevo conocimiento o generar hipótesis de investigación.

Comentarios finales

El desarrollo tecnológico alcanzado en los últimos años ha traído como consecuencia una producción acelerada de la información en formato digital, lo que genera la necesidad de resolver problemas y retos para el procesamiento de la misma. Por tal razón, es preciso contar con los métodos y las tecnologías necesarios para su procesamiento, organización y clasificación, entre otras tareas.

La minería de texto ayuda a explorar y explotar los contenidos de los documentos textuales digitales, que están en continuo incremento, con la meta de obtener información relevante, sirviendo de base en el desarrollo de investigaciones académicas y de negocios que debido a su potencial puede ser aplicada en todas las disciplinas; el área en donde más aplicaciones y desarrollo tienen es en las ciencias biomédicas y campos relacionados, sin embargo, la minería de texto en otros campos del conocimiento no ha tenido gran aplicación. Las tecnologías de la minería de texto pueden ser utilizadas también en las bibliotecas en el procesamiento y análisis de contenidos, con el objetivo de obtener relaciones entre temas y autores de los documentos académicos. En los ambientes académicos existen algunos obstáculos que deben de superarse, como los derechos de copia y las restricciones de acceso, en beneficio del desarrollo de la minería de texto.

Obras consultadas

Abbott, Dean. *Introduction to Text Mining* [en línea]*: Virtual Data Intensive Summer School*, July 10, 2013. <http://www.vscse.org/summerschool/2013/Abbott.pdf> [Consulta: 17 junio 2014].

Beath, Cynthia, Becerra-Fernández, Irma, Ross, Jean, Short, James. El valor de la explosión de la información [en línea]. *IQ Intelligence quarterly:diario de análisis avanzado*, 2013, p. 8-11. <http://www.sas.com/offices/europe/spain/IQ/Q113/IQBigAnalytics_GrandesOportunidades.pdf>

Belsky, Gary. Why Text Mining May Be The Next Big Thing [en línea]. *Technology & Media*, 20 de Marzo de 2012. <http://business.time.com/2012/03/20/why-text-mining-may-be-the-next-big-thing/>

Digital Universe [en línea]. emc2. <http://www.emc.com/leadership/programs/digital-universe.htm> [Consulta: 22 octubre 2013].

Extraction des informations et des connaissances [en línea]. <http://touriaelouahabi.wordpress.com/text-mining/principes-et-concepts-text-mining/> [Consulta: 17 junio 2014].

Fung, Benjamin C.M., Wang, Ke, Ester, Martin. B. *Hierarchical Document Clustering Using Frequent Itemsets* [en línea]. <http://epubs.siam.org/doi/pdf/10.1137/1.9781611972733.6> [Consulta: 14 julio 2010].

Fuzzy cluster analysis. Frank Höppner y otros. New York: J. Wiley, c1999. 289 p.

[34] *Idem.*

Gupta, Vishal, Lehal, Gurpreet S. A Survey of Text Mining Techniques and Applications. *Journal of emerging technologies in Web intelligence*, August 2009, vol. 1, no. 1, p. 60-76.

Haykin, Simon. *Neuronal Networks: A Comprehensive Foundation*. New York: Macmillan, 1994.

Journal Selection for MEDLINE Indexing at NLM [en línea]. U.S. National Library of Medicine. <http://www.nlm.nih.gov/pubs/factsheets/j_sel_faq.html#a15> [Consulta: 22 octubre 2013].

Krallinger, Martin. *Current trends in biomedical text mining: chemicals, drugs and mutations* [video]. Madrid, España: S. N. Centre, 2013. Jornadas Mavir, (19 de noviembre de 2013).

Marín, Maria José, Rico, Eva, Juli, Pascual. *Data mining en relación a la documentación periodística* [en línea]. <http://personales.upv.es/ccarrasc/doc/2003-2004/DMPeriodistico/TREBALLSRP.htm> [Consulta: 2011].

The National Centre for Text Mining [en línea]. <http://www.nactem.ac.uk/> [Consulta: 20 marzo 2014].

Neural Networks Design. Martin T. Hagan, Howard B. Demuth, Mark H. Beale, Orlando de Jesús. E.U.A: PWS Publishing, 1996.

Noorden, Richard Van. Elsevier opens its papers to text-mining. *Nature,* 3 febrero 2014, vol. 506, no. 7486.

Rauber, Andreas, Merkl, Dieter. Text Mining in the SOMLib Digital Library System: The Representation of Topics and Genres. *Applied Intelligence*, 2003, vol. 18, p. 271-293.

Salton, Gerard. *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Reading, Massachusetts: Addison Wesley, c1989. 530 p.

Sánchez, D., Martín-Bautista, M. *Un enfoque deductivo para la minería de texto* [en línea]. <http://www.softcomputing.es/estylf08/es/2006-XIII%20Congreso/articulos/40.pdf> [06 febrero 2014].

Sentiment analysis [en línea]. Wikipedia, The Free Encyclopedia. <http://en.wikipedia.org/w/index.php?title=Sentiment_analysis&oldid=625073725> [Consulta: 18 septiembre, 2014].

Smalhaiser, Neil, Swanson, D. Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease. *Neuroscience research communications*, 1994, vol. 15, p. 1-9.

Sukanya, M., Biruntha, S. Techniques on Text Mining. En: *IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), 23-25 Aug., 2012*. Ramanathapuram: IEEE, 2012.

Survey of Text Mining II: Clustering, Classification and Retrieval. Michael W. Berry, Malu Castellanos, editors. London: Springer, 2008. 240 p.

Swanson, Don R. Complementary structures in disjoint science literatures. En: *Proceeding SIGIR '91, Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and development in information retrieval*. New York: ACM, c1991, p. 280-289.

Tapping the Power of Text Mining. Weiguo Fan, Linda Wallace, Stephanie Rich, Zhongju Zhang. *Communication of the ACM*, September 2006, vol. 49, no. 9, p. 76-82.

Text mining: A new frontier for lossless compression. Ian H. Witten, Zane Bray, Malika Mahoui, Bill Teahan. En: *Data Compression Conference, 1999. Proceedings. DCC '99*. Snowbird: IEEE, 1999, p. 198-207.

Text mining in a digital library. Ian H. Witten, Katherine J. Don, Michael Dewsnip, Valentin Tablan. *International Journal on Digital Libraries*, 2003, vol. 4, no. 1, p. 56-59.

Value and benefits of text mining [en línea]. Jisc. <http://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining>

Yanliang Qi, Zhang, Yang, Song, Min. Text Mining for Bioinformatics: State of the Art Review. En: *Conference on Computer Science and Information Technology, 2009, ICCSIT 2009*, 2nd IEEE International Conference on Computer Science and Information Technology, 8-11 Aug., 2009. Beijing: IEEE, 2009.