reciamuc

Revista científica de investigación actualización del mundo de las ciencias

Gladys Patricia Guevara Albán [a]; Cristian Salomón Guevara Albán [b]; Daniel Elizondo Remache [c]

Tratamiento de la información en la web: Text Mining y Web Mining

Revista Científica de Investigación actualización del mundo de las Ciencias. Vol. 1 núm., 4, octubre, 2017, pp. 403-418

DOI: 10.26820/reciamuc/1.4.2017.403-418

Editorial Saberes del Conocimiento

a. Universidad Técnica de Babahoyo; Instituto Tecnológico Superior Babahoyo; gguevara@utb.edu.ec
b. Instituto Tecnológico Superior Babahoyo; cguevara@institutobabahoyo.edu.ec
c. Instituto Tecnológico Superior Babahoyo; delizondo@institutobabahoyo.edu.ec

RESUMEN

El descubrimiento de información útil en la web, la gestión de la información, de la documentación y del conocimiento son retos que enfrenta las empresas en el aumento de los datos no estructurados que se genera diariamente, aquellos datos en muchos casos no son tratado de manera adecuada tomando en cuenta que pueden ofrecer perspectivas únicas del comportamiento y actitudes de los usuarios que a utilizan. La minería de texto y la minería web son áreas con un gran crecimiento en los últimos años, por lo que han despertado el interés entre los investigadores y las empresas de ambas áreas. Dando la posibilidad a las empresas de explorar grandes cantidades de información, no organizados en forma de datos, establecer patrones y extraer conocimientos útiles a través de la Web. El tipo de investigación que fue aplicada es de tipo documental, descriptivo y bibliográfico. Entre las técnicas a utilizar es la recopilación de información y fichas electrónica. En este artículo se analizará los conceptos básicos, etapas y aplicaciones de la minería de texto y la minería Web, la mejora en el proceso de la toma de decisiones en las empresas y los servicios al cliente y un análisis comparativo entre ambas.

Palabras Claves: Minería de texto; minería web; toma de decisiones; extracción de información.

ABSTRACT

The discovery of useful information on the web, information management, documentation and knowledge are challenges facing companies in the increase of unstructured data that is generated daily, those data in many cases are not treated in a way Taking into account that they can offer unique perspectives of the behavior and attitudes of the users that use it. Text mining and web mining are areas of great growth in the last few years, which has sparked interest among researchers and companies in both areas. By giving enterprises the ability to explore large amounts of information, not organized in the form of data, to establish patterns and extract useful knowledge through the Web. The type of research that was applied is documentary, descriptive and bibliographic. Among the techniques to be used is the collection of information and electronic tokens. This article will analyze the basic concepts, stages and applications of text mining and web mining, improvement in the decision-making process in companies and customer services and a comparative analysis between both.

Keywords: Text mining; web mining; decision making; extraction of information.

Introducción.

A comienzos de los años ochenta empieza el boom del estudio del manejo de los datos no estructurados a través de la minería de textos, en ese entonces necesitaban una gran cantidad de esfuerzo humano para la recuperación de grandes volúmenes de información, pero en la última década los avances tecnológicos han permitido que esta área progrese considerablemente y que se haya consolidado como un objetivo importante y un problema a resolver la integración del usuario en el proceso de minería texto.

"La complejidad de las páginas Web excede la complejidad de cualquier colección de documentos de texto tradicional. Aunque la Web funciona como una enorme biblioteca digital, las páginas, en si mismas, carecen de una estructura uniforme y contienen muchos más estilos y contenidos variados que cualquier conjunto de libros o documentos basados en texto tradicional. Por otra parte, el gran número de documentos que esta biblioteca digital contiene, no pueden ser indexados, lo cual hace que la búsqueda de los datos en sus contenidos sea extremadamente complicada" (Aguilar, 2011).

Minería de texto es utilizada en muchos campos de la ciencia, a nivel financiero y bancario, en el análisis de mercados y comercios, en el área de la salud, a nivel educativo, en procesos industriales, en medicina, biología y bioingeniería así en las telecomunicaciones, bibliotecología y diferentes áreas. (Pérez López & Santín González, 2007).

El manejo de información en la web da inicio a la "Minería Web es una herramienta útil para el hallazgo de nuevos conocimientos; para eso, emplea la información obtenida de los documentos y servicios Web (textos, imágenes, videos, hiperenlaces, ficheros Log, etc.)". (Reyes & Ruiz, 2007)

En los últimos años la minería de texto y la minería web han ocupado un lugar primordial en el mundo empresarial, en la mayoría de estas la problemática que se presenta, es el aumento de grandes volúmenes de datos no estructurados, donde estos datos pueden ofrecer perspectivas únicas del comportamiento del cliente y que deben ser aprovechados por el empresario para mejorar la gestión de la información y la toma de decisiones.

En la mayoría de investigaciones existentes en esta área se dedican al estudio de que técnicas y patrones para la búsqueda de conocimiento en grandes colecciones de documentos no estructurados, siendo pocos los que aborden las diferentes aplicaciones de estas herramientas, convirtiendo así en el objetivo principal de este trabajo de investigación, la cual permitirá mostrar las principales aplicaciones de la minería de texto y minería web que facilitan la recaudación y análisis de la información para la toma de decisiones en tiempo real.

Desarrollo

Minería de texto

La producción de la información digital en grandes cantidades, ha generado la necesidad de desarrollar métodos, patrones, técnicas y sistemas que permitan procesar datos no

estructurados, para su mayor comprensión en el tratamiento de información y en la toma de decisiones esto ha dado origen a la adopción de la tecnología de minería de textos.

"La minería de texto o text mining 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". (Contreras Barrera, 2014)

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. 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.

"Text Mining se refiere al examen de una colección de documentos y el descubrimiento de información no contenida en ningún documento individual de la colección; en otras palabras, trata de obtener información sin haber partido de algo". (Nasukawa y otros, 2001)

En muchas ocasiones confunde los términos minería de datos con la minería de textos, en la primera la información se obtiene normalmente de bases de datos, en la que la información está estructurada, la cual su manipulación, consulta y extraer información de la base de datos es más sencilla. Mientras que en la segunda su objetivo principal es la búsqueda de conocimiento en grandes colecciones de documentos no estructurados.

En un porcentaje alto las empresas almacenan información de manera textual no estructurada como: documentos, actas de reuniones, informes, catálogos, entre otros. Uno de los objetivos de la Minería de Texto es detectar patrones no triviales e incluso información sobre el conocimiento almacenado en las mismas.

En las empresas, las técnicas de minería de texto son sistemas encargados de gestionar las colecciones de información digitales, las cuales permiten mejorar la toma de decisiones. El análisis e interpretación de la información recopilada en las redes están sean positivas o negativas, ayudan a mejorar el desempeño de la empresa y el servicio al cliente.

Las organizaciones que pueden beneficiarse de este sistema de manejo de texto, por un lado son aquellas compañias que tienen un contacto directo con sus clientes, como aerolineas, agencias de viajes, hoteles, restaurantes, entre otras. Y por otro lado serian las empresas que brindan servicios on line, que poseen sistemas de actividades, encuestas, correos, entrevistas, preguntas, que muchas veces no se pueden analizar por extenso para la toma de decisiones, generando la mineria web.

Para Polo Ahumada (2016) el text mining comprende tres actividades fundamentales las cuales son:

- Recuperar la información : seleccionar los textos adecuados
- Extraer la información contenida en esos textos: datos claves, hechos y acontecimientos
- Utilizar la minería de datos para encontrar asociaciones entres esos textos claves (galeon.com, 2016)

En el proceso del text mining consiste de dos etapas principales: una etapa de pre-procesamiento y una etapa de descubrimiento (Tan, 1999).

"En la primera etapa, los textos se transforman a algún tipo de representación estructurada o semi-estructurada que facilite su posterior análisis, mientras que en la segunda etapa las representaciones intermedias se analizan con el objetivo de descubrir en ellas algunos patrones interesantes o nuevos conocimientos". (Montes, 2001)

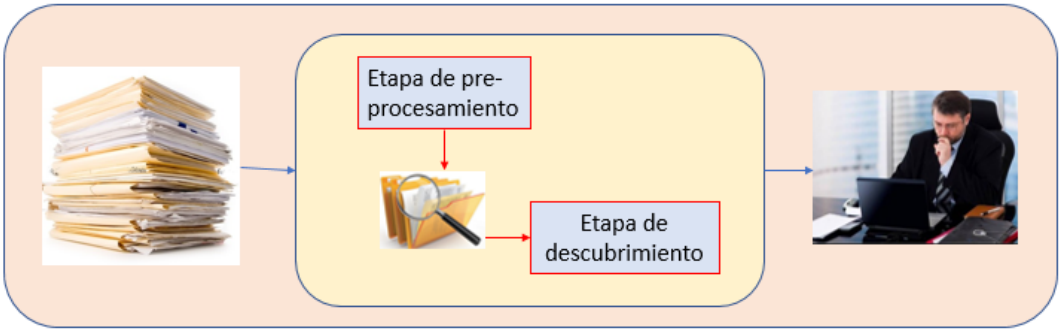

Figura 1. Etapas de la minería de datos, elaborado por Montes M.

Aplicación de la minería de textos

La aplicación de la minería de texto a través de revisión bibliográficas o simplemente "googleando" permite acceder a documentos en el área de medicina, empresarial, bibliográfica, educación, entre otros.

Según Rochina (2017) manifiesta que el text mining se podrá aplicar para:

- La extracción de información
- El análisis de sentimientos o minería de opiniones

- La clasificación documental
- La elaboración de resúmenes
- La extracción de conocimiento.

Mientras que Polo (2016) sostiene que el text mining se aplica en:

- **Extracción de información:** extrae información entre grandes volúmenes de datos que se encuentra en la web, para su mejor comprensión.
- **Clasificar documentos:** Cataloga y navega en grupos de documentos en la web, aplica patrones de búsqueda en estos grupos y extrae información descriptiva de cada grupo.
- **Elaboración de resúmenes**: Sobre una temática específica se puede obtener una descripción de manera general de la información recopilada.
- **Extracción de conocimiento**: En información extraída se puede aplicar modelos de conocimientos.

Para Badom (2014) la minería de texto se utiliza principalmente para:

- Extraer información relevante de un documento.
- Agregar y comparar información automáticamente.
- Clasificar y organizar documentos según su contenido.
- Organizar depósitos para búsqueda y recuperación

- Clasificar textos e indizarlos en el Web.

Minería web

Mineria web es otra fuente extensa de informacion valiosa, el descubrimiento de los patrones utilies y la informacion proveniente de la World Wide Web, la cual ayuda a descubrir conocimientos potencialmente útiles para las organizaciones

"El proceso de descubrir relaciones o patrones interesantes en un conjunto de datos en la web se llama minería web". (Baeza-Yates, 2004)

Según Kosala y sus colaboradores (2000): "Consiste en aplicar las técnicas de minería de datos a documentos y servicios del Web. Todos los que visitan un sitio en Internet dejan huellas digitales (direcciones de IP, navegador, etc.) que los servidores automáticamente almacenan en una bitácora de accesos (Log). Las herramientas de Web Mining analizan y procesan estos Log para producir información significativa. Debido a que los contenidos de Internet consisten en varios tipos de datos, como texto, imagen, vídeo, metadatos o hiperligas, se utiliza el término multimedia Data Mining (minería de datos multimedia) como una instancia del Web Mining". Po lo antes expuesto podemos decir que Minería Web es como la aplicación de técnicas de Minería de Datos en Internet para el descubrimiento y análisis de páginas Web, la generación de patrones para clasificar la información de las páginas Web, entre otras cosas. (Aguilar, 2000)

En la figura 2, se pueden distinguir tres tipos de minería de Web, (Fuentes & Ruiz, 2007):

Minería de utilización web. Descubre información a partir de los datos de utilización de la Web.

Minería de contenidos web. Extrae información a partir de los contenidos de los documentos Web.

Minería de estructura web. Descubrir información a partir de la estructura de la Web.

De los tipos de minería de Web antes mencionados, los más utilizados para la extracción de información en los sistemas empresariales basados en web es la minería de utilización Web o web usage mining y la Minería de contenidos web o Web content mining.

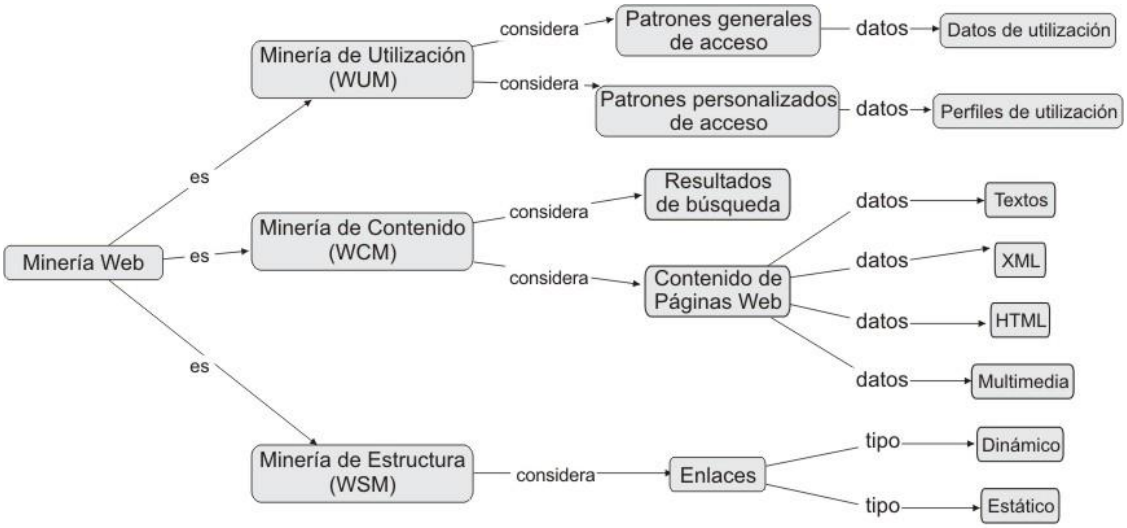

Figura 2. Mapa conceptual de la clasificación minería Web, según Juan Carlos Dürsteler

Guiar al usuario en la búsqueda, recuperación y utilización de la información en la web es un reto que se presenta hoy en día, se debe desarrollar sistemas web que faciliten la formación de destrezas y habilidades en el acceso de la información en la web. Los motores de búsqueda

pueden desempeñar un papel esencial en la viabilidad de los sistemas de información basados en Internet, siempre que existan aplicaciones que puedan analizar y evaluar la relevancia de la información para el usuario. (Bordón & D'Avanzo, 2004)

Algunas técnicas de minería web

Para Aguilar (2011), entre algunas técnicas de minería web tenemos:

- Agrupamiento
- Clasificación
- Detección de reglas de asociación.
- Análisis de caminos.
- Detección de patrones secuenciales, entre otras

Agrupamiento. La técnica de agrupamiento o clustering agrupan características y comportamientos similares en grupos homogéneos.

Clasificación- Los clasifica de acuerdo al tipo de agrupamiento que previamente tuvieron, los organiza de acuerdo al perfil para clientes/usuarios que acceden a archivos concretos del servidor, en función de sus patrones de acceso o de la información extraída.

Detección de reglas de asociación. Descubre todas las asociaciones y correlaciones entre los accesos y usos de la Web por parte de los usuarios.

Análisis de caminos. Esta técnica realiza el análisis mediante grafos orientados que representan relaciones entre páginas Web.

Detección de patrones secuenciales. Esta técnica localiza las secuencias de patrones de un conjunto de elementos seguidos por otro elemento en un conjunto de transacciones o visitas ordenadas en el tiempo.

Discusión y resultados.

Al revisar el estado del arte de varias investigaciones acerca de la temática de minería de texto y minería web, se observa que en un 70% de los visitantes que navegan en la web cumplen con patrones de costumbre de visitar el sitio web varias veces. Esto genera que los patrones de asociación, clasificación y agrupamiento de clientes o de información, facilitara la gestión de la información para actividades futuras dentro de las empresas. Además, esto ayuda a validar que tipos de visitantes navegan por el sitio web dado, que tipo de usuarios prefieren un contenido específico, los rasgos o características tiene en común, si son fieles al sitio web. También se observó las tendencias que los usuarios prefieren de acuerdo al tipo de contenido y estructura, todo lo expuesto anteriormente estos datos depende del tipo de aplicación dado por las herramientas text mining o web mining.

En muchas áreas las aplicaciones de minería de texto y minería web son muy utilizadas con el objetivo de mejorar la gestión de la información en la www. En el área empresarial estas aplicaciones son de gran ayuda ya que permiten direccionar información específica a departamentos que lo soliciten. Como por ejemplo cuando llega un email enviado por un proveedor este lo redirecciona al departamento de compras, dependiendo del patrón de búsqueda

que se haya aplicado para identificar el contenido y validarlo. Otra área en la que está siendo utilizada esta herramienta es la de mercados en la Web, extrae el conocimiento sobre estadísticas de manejo de determinados conceptos y temas en la web. En el área Bibliotecaria extrae la información más relevante del documento, compara la información automáticamente en los grupos de documentos con temas a fines, mostrándolos en la web de manera indexada.

Al realizar un cuadro comparativo sobre la aplicación del text mining en la recuperación de la información en la web de acuerdo a las diversas posiciones de los autores anteriormente mencionados, se observa que coinciden en varios procesos como: Extracción de información, clasificar la información o documentos, extracción de conocimientos, elaboración de resúmenes.

Conclusión.

La minería de textos y la minería web brinda herramientas muy útiles para el análisis de los datos, textos, documentos, imágenes, hiperenlaces, entre otros dentro de la web, una vez identificada los patrones de comportamiento que ayuden a la toma de decisiones permitirá establecer en cada empresa, el tipo de técnica que va a utilizar de acuerdo a las características, especificaciones, y problemas de ellas.

La minería de la web y minería de texto en los últimos años ha tenido una amplia aceptación en el mercado empresarial, diseño de sitios web, e-commerce y la publicidad en internet.

Gracias a la disposición la minería de texto y la minería web, se puede conocer, entender y predecir las conductas, rasgos, necesidades de los usuarios en la web, a través de esto se puede

personalizar los procesos de búsqueda, agrupación de contenidos, aplicación de algoritmos, de acuerdo a los requerimientos.

Bibliografía.

Aguilar, José; Altamiranda, Junior (2004). Conceptos sobre minería web. Gerencia Tecnología Informática. 3(7): 5-15.

Aguilar J., Altamiranda, J. (2003) "Minería de Datos en la Web usando Computación Evolutiva", (Ed. N. Brisaboa). Barcelona: AECI RISTOS2: 153-168.

Baeza-Yates, R. (2004). *Excavando la web.* Obtenido de El profesional de la informació: http://personales.dcc.uchile.cl/~rbaeza/inf/EPIexcavando.pdf

Bordón L, D'Avanzo E. (2004). Perspectivas para la integración de la minería de textos y la gestión del conocimiento. The IPTS Report. 85(1).

Contreras Barrera, M. (2014). Text mining: a current view. *Biblioteca Universitaria Mexico*, 129-138.

Fuentes, S., & Ruiz, M. (2007). Minería Web: un recurso insoslayable para el profesional de la información. *ACIMED*.

Montes, M. (2001). *Minería de texto: Un nuevo reto computacional.* Obtenido de https://ccc.inaoep.mx/~mmontesg/publicaciones/2001/MineriaTexto-md01.pdf

Pérez López, C., & Santín González, D. (2007). Minería de datos: Técnicas y herramientas. Madrid: Internacional Thomson Ediciones Paraninfo

Polo Ahumada Ana. (2016). Minería de datos, de textos y de sentimientos. Recuperado de https://www.gestiopolis.com/mineria-datos-textos-sentimientos-2/

Reyes SC, Ruiz Lobaina M. Minería Web: un recurso insoslayable para el profesional de la información. Acimed 2007; 16(4).

Rochina, P. (25 de 04 de 2017). *¿Qué es y cuáles son las aplicaciones del Text Mining?* Obtenido de Revista digital - INESEM

Sukanya M., Biruntha S. (2012). Techniques on Text Mining. Madrid: Conference: Advanced Communication Control and Computing Technologies (ICACCCT)

Witten Ian, Don Katherine, Dewsnip Michael, Tablan Valentin. (2004). Text mining in a digital library. International Journal on Digital Libraries. 4(1): 56-59.