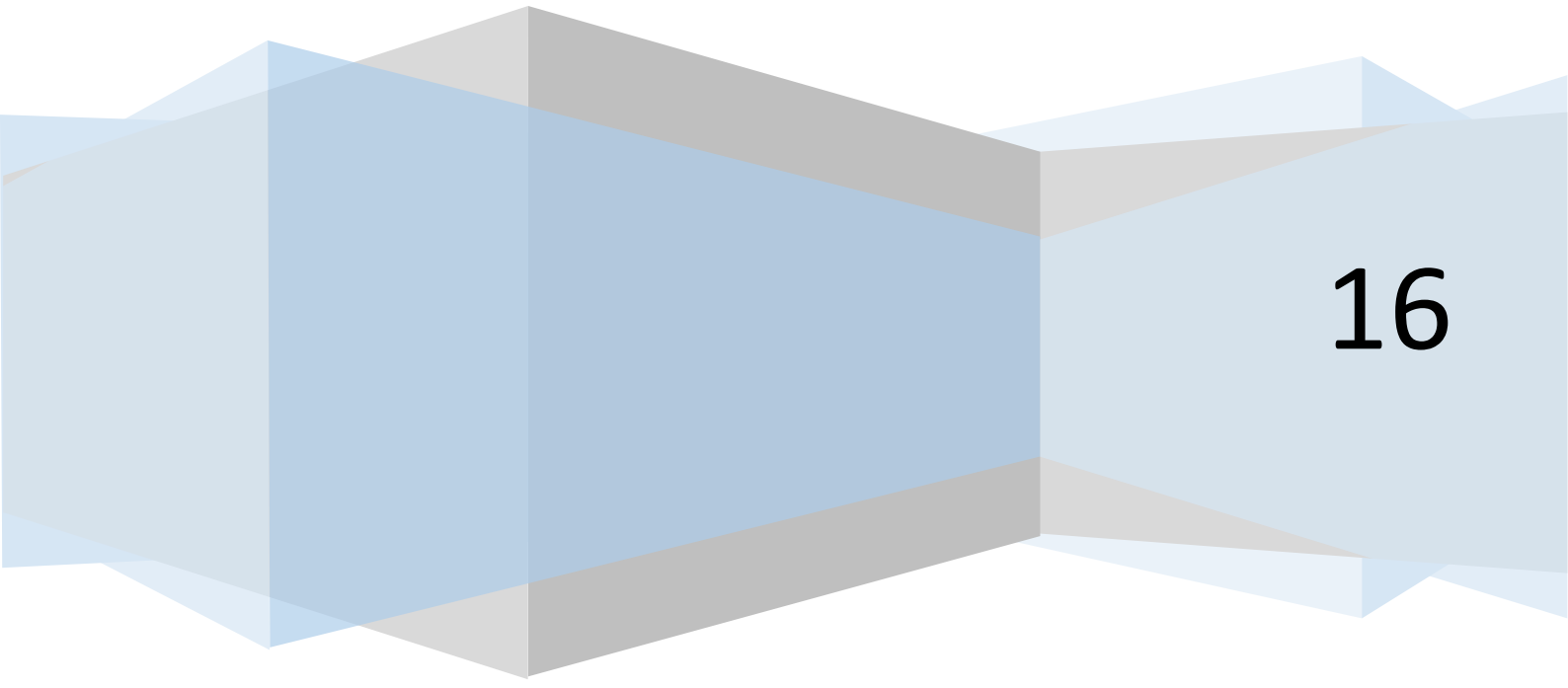
FUNDAMENTOS DE INGENIERIA ADMINISTRATIVA

MINERIAS (TEXTO, DATOS, SENTIMIENTOS)

KEILA YERITZE ROJAS GUTIERREZ



16

CONTENIDO

INTRODUCCIÓN

El aumento de los medios sociales tales como blogs y redes sociales ha traído consigo información, esta información generalmente no tiene un tratamiento previo ni tampoco se puede saber la razón o el estado de ánimo del escritor.

Conocer en general donde buscar información necesaria, como tratarla y cómo interpretarla es de vital importancia, no solo para los individuos sino igual para las organizaciones, el siguiente articulo busca ayudar al lector en su búsqueda constante de datos e información.

DESARROLLO

DEFINICIONES

La **minería de textos** se refiere al proceso de derivar información nueva de textos, consiste en descubrir, a partir de cantidades de texto grandes, el conocimiento que no está literalmente escrito en cualquiera de los documentos. Esto incluye buscar tendencias, promedios, desviaciones, dependencias, etc.

(TRIPOD, 2015)

La **minería de datos** o **exploración de datos** es un campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. (MARCEL, 2015)

Mineria de sentimientos o también conocido como **minería de opinión**, se refiere al uso de procesamiento de lenguaje natural, análisis de texto y lingüística computacional para identificar y extraer información subjetiva de unos recursos, el análisis de sentimiento intenta determinar la actitud de un interlocutor o un escritor con respecto a algún tema

o la polaridad contextual general de un documento. La actitud puede ser su juicio o evaluación.

(BANNISTER, 2015)

COMENTARIO

Se suele confundir la minería de textos con la minería de datos, pero lo diferenciaremos porque en la minería de datos la información se obtiene normalmente de bases de datos, en la que la información está estructurada. Por este motivo es más sencilla la extracción de la información de una base de datos, que está pensada para que se pueda tratar su información de manera automática, al contrario, a lo que ocurre en la minería de textos.

MINERIA DE TEXTOS

La minería de textos es un área multidisciplinaria basada en la recuperación de información, minería de datos, aprendizaje automático, estadísticas y la lingüística computacional. Como la mayor parte de la información (más de un 80%) se encuentra actualmente almacenado como texto, se cree que la minería de textos tiene un gran valor comercial

Historia

Ya en 1977, el sistema THOMAS ilustró cómo las palabras o las frases clave podían utilizarse para guiar a los usuarios en el descubrimiento de documentos de referencia útiles. Las frases clave son un tipo especialmente útil de información abreviada. Sin embargo, tales frases se eligen con frecuencia manualmente, bien por los autores o por indizadores profesionales. Condensan documentos en unas pocas palabras y frases, ofreciendo una descripción breve y precisa de los contenidos de un documento.

A comienzos de los años ochenta surgieron los primeros esfuerzos de minería de textos que necesitaban una gran cantidad de esfuerzo humano, pero los avances tecnológicos han permitido que esta área progrese de manera rápida en la última década.

Actividades fundamentales

Recuperación de información, es decir, seleccionar los textos pertinentes.

Extracción de la información incluida en esos textos: hechos, acontecimientos, datos clave, relaciones entre ellos, etc.

Metodología

El ser la minería de textos una técnica relativamente nueva, cambiante y que puede adaptarse a diferentes situaciones y casos, por lo que no existe un método estricto a seguir siempre. Sin embargo, en términos generales se podría decir que estas son las cuatro etapas principales:

Primera	Determinación de los objetivos. Aclarar que es lo que se esta buscando con esta investigación, acotando hasta que punto se quiere profundizar en la misma y definiendo claramente los limites.
Segunda	Preprocesamiento de los datos, que sería la selección, análisis y reducción de los textos o documentos de los que se extraerá la información. Esta etapa consume la mayor parte del tiempo.
Tercera	Determinación del modelo. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse unas técnicas u otras.
Cuarta	Análisis de los resultados. A partir de los datos extraídos se tratara de ver su coherencia y se buscaran evidencias, similitudes, excepciones, etc, que puedan servir al especialista o al usuario que haya encargado el estudio para extraer conclusiones que pueda utilizar para mejorar algún aspecto de su empresa, compañía, administración u organización en general.

Aplicaciones académicas

El tema de la minería de textos es de importancia para publicadores que tengan grandes bancos de data que requieran de indexación. Esto es el caso en particular para disciplinas

científicas en las que hay una gran cantidad de información muy específica en forma de texto escrito. Es por ello que se han presentado iniciativas como el Open Text Mining Interface (OTMI) y el common Journal Publishing Document Type Definition (DTD) de la NIH, que ofrecerían datos semánticos para responder a preguntas muy específicas sin quitar las barreras del publicador al acceso público.

(GALEON, 2015)

MINERIA DE DATOS

La minería de datos es el proceso de detectar la información accionale de grandes conjuntos de datos. Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos.Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiado datos.

Los modelos de minería de datos se pueden aplicar en escenarios como los siguientes:

Previsión: calcular las ventas y predecir las cargas de servidor o el tiempo de inactividad del servidor.

Riesgo y probabilidad: elegir los mejores clientes para la distribución de correo directo, determinar el punto de equilibrio probable para los escenarios de riesgo, y asignat probabilidades a diagnósticos u otros resultados.

Recomendaciones: determinar los productos que se pueden vender juntos y generar recomendaciones.

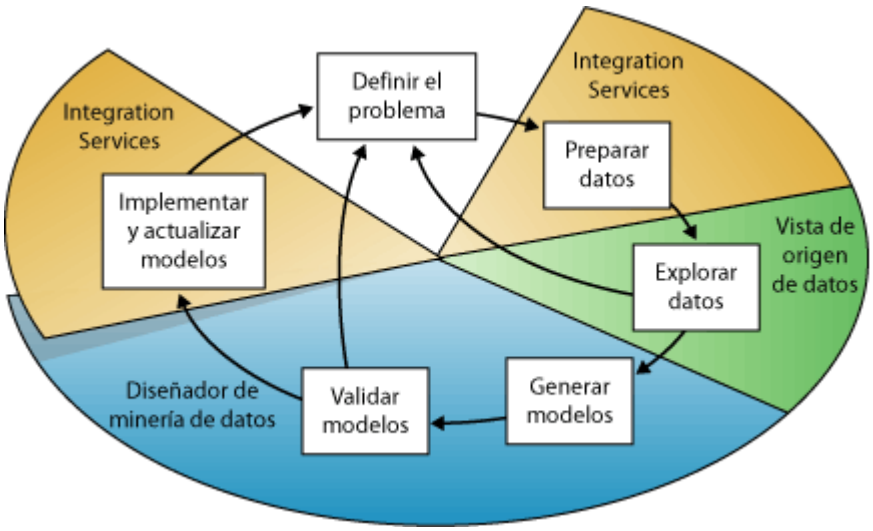
Buscar secuencias: analizar los artículos que los clientes han introducido en el carrito de compra y predecir los posibles eventos.

Agrupación: separar los clientes o los eventos en clústeres de elementos relacionados, y analizar y predecir afinidades.

La generación de un modelo de minería de datos forma parte de un proceso mayor que incluye desde la formulación de preguntas acerca de los datos y la creación de un modelo para responderlas, hasta la implementación del modelo en un entorno de trabajo. Este proceso se puede definir mediante los seis pasos básicos siguientes:

- Definir el problema
- Preparar los datos
- Explorar los datos
- Generar modelos
- Explorar y validar los modelos
- Implementar y actualizar los modelos

El siguiente diagrama describe las relaciones existentes entre cada paso del proceso

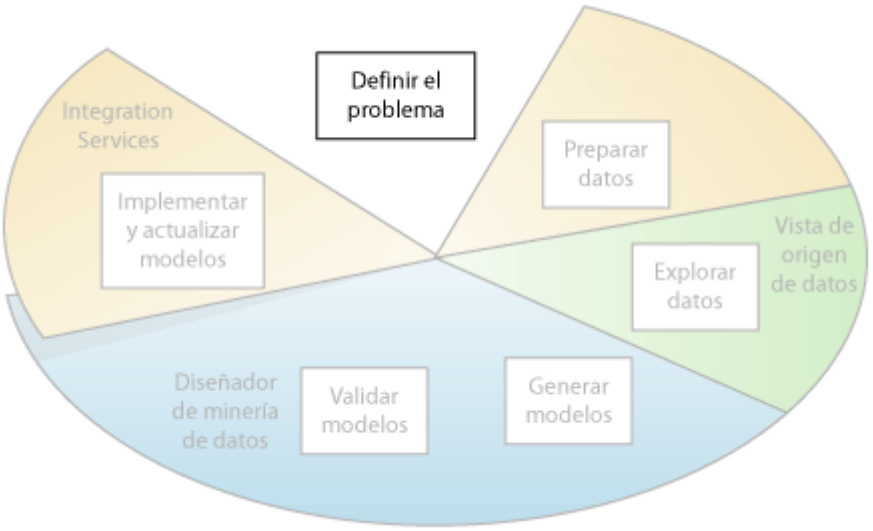


El proceso que se ilustra en el diagrama es cíclico, lo que significa que la creación de un modelo de minería de datos es un proceso dinámico e iterativo. Una vez que ha explorado los datos, puede que descubra que resultan insuficientes para crear los modelos de minería de datos adecuados y que, por tanto, debe buscar más datos. O bien, puede generar varios modelos y descubrir entonces que no responden adecuadamente al problema planteado cuando los definió y que, por tanto, debe volver a definir el problema. Es posible que deba actualizar los modelos una vez

implementados debido a que haya más datos disponibles. Puede que haya que repetir cada paso del proceso muchas veces para crear un modelo adecuado.

Definir el problema

El primer paso del proceso de minería de datos, tal como se resalta en el siguiente diagrama, consiste en definir claramente el problema y considerar formas de usar los datos para proporcionar una respuesta para el mismo.



Este paso incluye analizar los requisitos empresariales, definir el ámbito del problema, definir las métricas por las que se evaluará el modelo y definir los objetivos concretos del proyecto de minería de datos.Estas tareas se traducen en preguntas como las siguientes:

¿Qué está buscando?

¿Qué tipos de relaciones intenta buscar?

¿Refleja el problema que está intentando resolver las directivas o procesos de la empresa?

¿Desea realizar predicciones a partir del modelo de minería de datos o solamente buscar asociaciones y patrones interesantes?

¿Qué resultado o atributo desea predecir?

¿Qué tipo de datos tiene y qué tipo de información hay en cada columna?En caso de que haya varias tablas, ¿cómo se relacionan?¿Necesita limpiar, agregar o procesar los datos antes de poder usarlos?

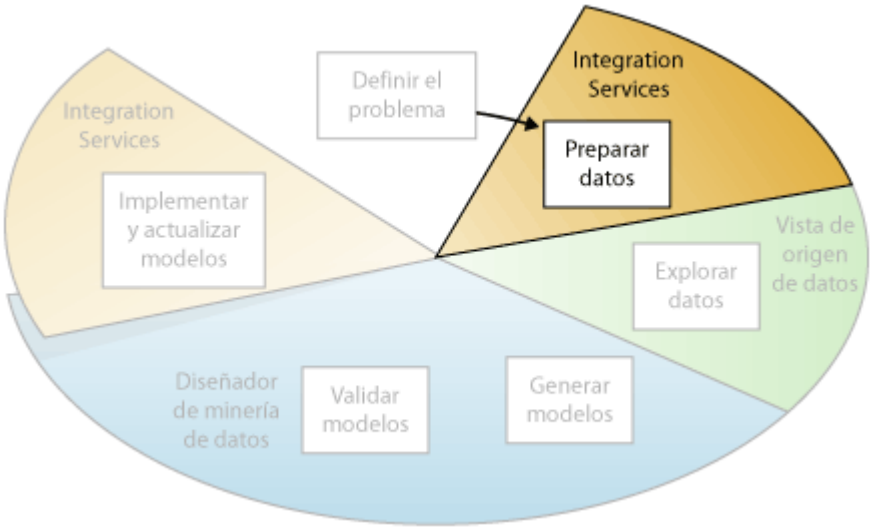
¿Cómo se distribuyen los datos?¿Los datos son estacionales?¿Los datos representan con precisión los procesos de la empresa?

Para responder a estas preguntas, puede que deba dirigir un estudio de disponibilidad de datos para investigar las necesidades de los usuarios de la empresa con respecto a los datos disponibles. Si los datos no abarcan las necesidades de los usuarios, podría tener que volver a definir el proyecto.

También debe considerar las maneras en las que los resultados del modelo se pueden incorporar en los indicadores de rendimiento clave que se utilizan para medir el progreso comercial.

Preparar los datos

El segundo paso del proceso de minería de datos, como se indica en el siguiente diagrama, consiste en consolidar y limpiar los datos identificados en el paso Definir el problema.



Los datos pueden estar dispersos en la empresa y almacenados en formatos distintos; también pueden contener incoherencias como entradas que faltan o incorrectas. Por ejemplo, los datos pueden mostrar que un cliente adquirió un producto incluso antes que se ofreciera en el mercado o que el cliente compra regularmente en una tienda situada a 2.000 kilómetros de su casa.

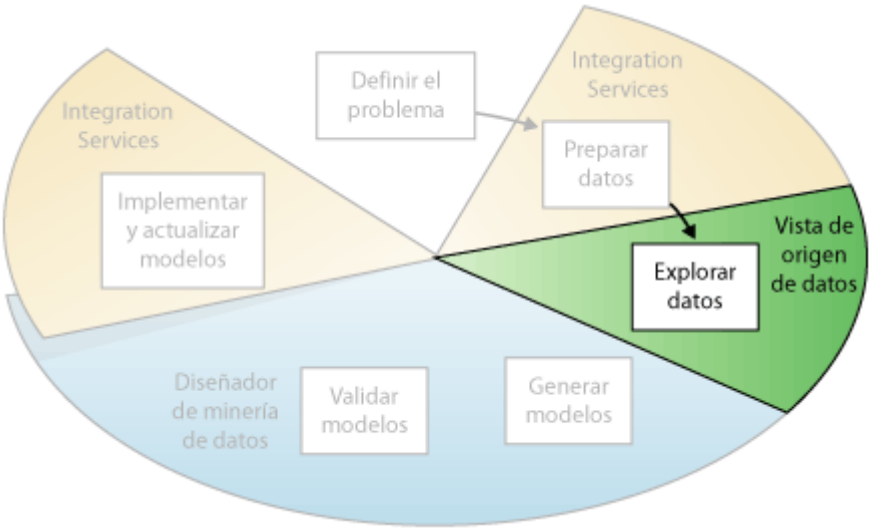
La limpieza de datos no solamente implica quitar los datos no válidos o interpolar valores que faltan, sino también buscar las correlaciones ocultas en los datos, identificar los orígenes de datos que son más precisos y determinar qué columnas son las más adecuadas para el análisis. Por ejemplo, ¿debería utilizar la fecha de envío o la fecha de pedido?¿Qué influye más en las ventas: la cantidad, el precio total o un precio con descuento? Los datos incompletos, los datos incorrectos y las entradas que parecen independientes, pero que de hecho están estrechamente correlacionadas, pueden influir en los resultados del modelo de maneras que no espera.

Por consiguiente, antes de empezar a generar los modelos de minería de datos, debería identificar estos problemas y determinar cómo los corregirá. En la minería de datos, por lo general se trabaja con un conjunto de datos de gran tamaño y no se puede examinar la calidad de los datos de cada transacción; por tanto, es posible que necesite usar herramientas de generación de perfiles de datos, y de limpieza y filtrado automático de datos.

Es importante tener en cuenta que los datos que se usan para la minería de datos no necesitan almacenarse en un cubo de procesamiento analítico en línea (OLAP), ni siquiera en una base de datos relacional, aunque puede usar ambos como orígenes de datos. Puede realizar minería de datos mediante cualquier origen de datos definido como origen de datos de Analysis Services. Por ejemplo, archivos de texto, libros de Excel o datos de otros proveedores externos.

Explorar los datos

El tercer paso es explorar los daros preparados



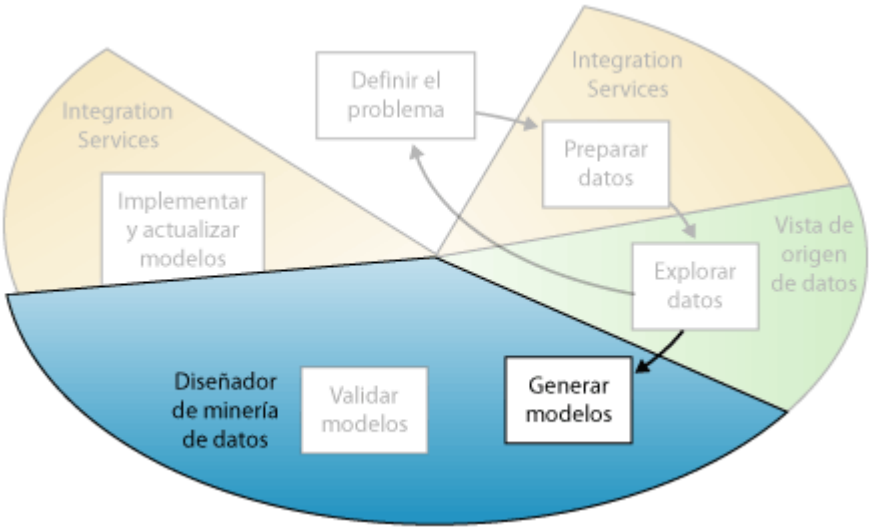
Se debe conocer los datos para tomar las decisiones adecuadas al crear los modelos de minería de datos. Entre las técnicas de exploración se incluyen calcular los valores mínimos y máximos, calcular la media y las desviaciones estándar, y examinar la distribución de los datos. Por ejemplo, al revisar el máximo, el mínimo y los valores de la media se podría determinar que los datos no son representativos de los clientes o procesos de negocio, y que por consiguiente debe obtener más datos equilibrados o revisar las suposiciones que son la base de sus expectativas. Las desviaciones estándar y otros valores de distribución pueden proporcionar información útil sobre la estabilidad y exactitud de los resultados. Una desviación estándar grande puede indicar que agregar más datos podría ayudarle a mejorar el modelo. Los datos que se desvían mucho de una distribución estándar se podrían sesgar o podrían representar una imagen precisa de un problema de la vida real, pero dificultan el ajustar un modelo a los datos.

Al explorar los datos para conocer el problema empresarial, puede decidir si el conjunto de datos contiene datos defectuosos y, a continuación, puede inventar una estrategia para corregir los problemas u obtener una descripción más profunda de los comportamientos que son típicos de su negocio.

Tenga en cuenta que cuando se crea un modelo, Analysis Services crea automáticamente resúmenes estadísticos de los datos contenidos en él, que puede consultar para su uso en informes o análisis.

Generar modelos

El cuarto paso del proceso de minería de datos, como se resalta en el siguiente diagrama, consiste en generar el modelo o modelos de minería de datos. Usará los conocimientos adquiridos en el paso Explorar los datos para definir y crear los modelos.



En este paso se deberán definir qué columnas de datos desea que se usen; para ello, creará una estructura de minería de datos. La estructura de minería de datos se vincula al origen de datos, pero en realidad no contiene ningún dato hasta que se procesa. Al procesar la estructura de minería de datos, Analysis Services genera agregados y otra información estadística que se puede usar para el análisis. Cualquier modelo de minería de datos que esté basado en la estructura puede utilizar esta información.

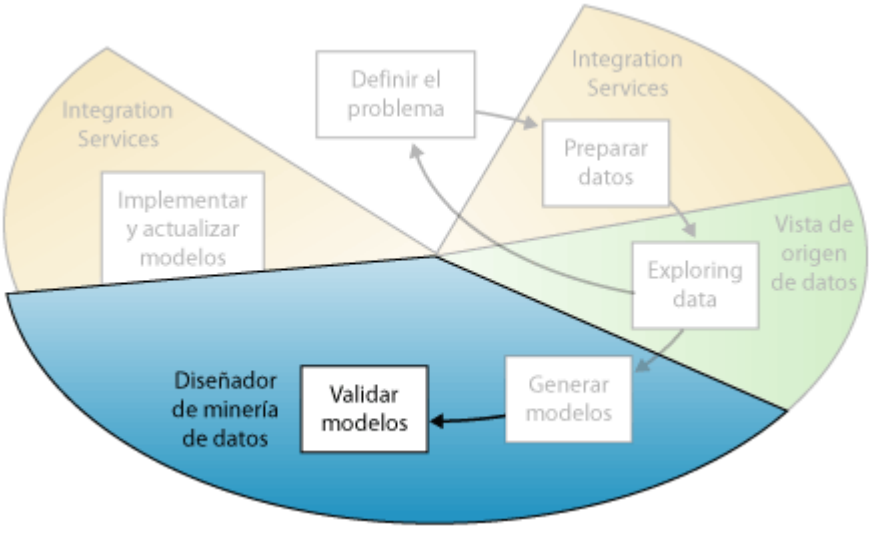
Antes de procesar la estructura y el modelo, un modelo de minería de datos simplemente es un contenedor que especifica las columnas que se usan para la entrada, el atributo que está prediciendo y parámetros que indican al algoritmo cómo procesar los datos. El procesamiento de un modelo a menudo se denomina *entrenamiento*. El entrenamiento hace referencia al proceso de aplicar un algoritmo matemático concreto a los datos de la estructura para extraer patrones. Los patrones que encuentre en el proceso de entrenamiento dependerán de la selección de los datos de entrenamiento, el algoritmo que elija y cómo se haya configurado el algoritmo.

También puede utilizar los parámetros para ajustar cada algoritmo y puede aplicar filtros a los datos de entrenamiento para utilizar un subconjunto de los datos, creando resultados diferentes. Después de pasar los datos a través del modelo, el objeto de modelo de minería de datos contiene los resúmenes y modelos que se pueden consultar o utilizar para la predicción.

Es importante recordar que siempre que los datos cambian, debe actualizar la estructura y el modelo de minería de datos.Al actualizar una estructura de minería de datos volviéndola a procesar, Analysis Services recupera los datos del origen, incluido cualquier dato nuevo si el origen se actualiza dinámicamente, y vuelve a rellenar la estructura de minería de datos.Si tiene modelos que están basados en la estructura, puede elegir actualizar estos, lo que significa que se vuelven a entrenar con los nuevos datos, o puede dejar los modelos tal cual.

Explorar y validar los modelos

El quinto paso del proceso de minería de datos, como se resalta en el siguiente diagrama, consiste en explorar los modelos de minería de datos que ha generado y comprobar su eficacia.



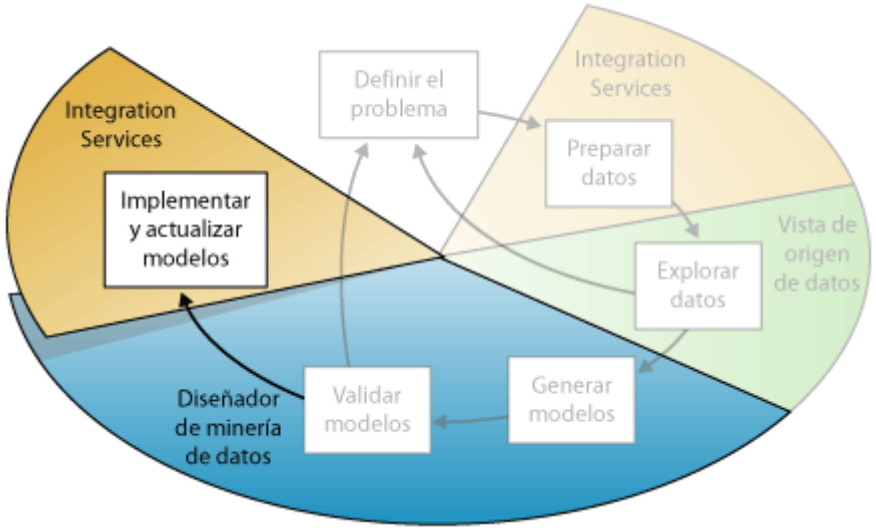
Antes de implementar un modelo en un entorno de producción, es aconsejable probar si funciona correctamente. Además, al generar un modelo, normalmente se crean varios con configuraciones diferentes y se prueban todos para ver cuál ofrece los resultados mejores para su problema y sus datos.

El conjunto de datos de entrenamiento se utiliza para generar el modelo y el conjunto de datos de prueba para comprobar la precisión del modelo mediante la creación de consultas de predicción.

Si ninguno de los modelos que ha creado en el paso Generar modelos funciona correctamente, puede que deba volver a un paso anterior del proceso y volver a definir el problema o volver a investigar los datos del conjunto de datos original.

Implementar y actualizar los modelos

El último paso del proceso de minería de datos, como se resalta en el siguiente diagrama, consiste en implementar los modelos que funcionan mejor en un entorno de producción.



Una vez que los modelos de minería de datos se encuentran en el entorno de producción, puede llevar acabo diferentes tareas, dependiendo de sus necesidades. Las siguientes son algunas de las tareas que puede realizar:

Use los modelos para crear predicciones que luego podrá usar para tomar decisiones comerciales.

Actualizar dinámicamente los modelos, cuando entren más datos en la organización, y realizar modificaciones constantes para mejorar la efectividad de la solución debería ser parte de la estrategia de implementación.

Técnicas de minería de datos

Como ya se ha comentado, las técnicas de la minería de datos provienen de la inteligencia artificial y de la estadística, dichas técnicas, no son más que algoritmos, más o menos sofisticados que se aplican sobre un conjunto de datos para obtener unos resultados.

Las técnicas más representativas son:

Redes neuronales.- Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida. Algunos ejemplos de red neuronal son:

El perceptrón.

El perceptrón multicapa.

Los mapas autoorganizados, también conocidos como redes de Kohonen.

Regresión lineal.- Es la más utilizada para formar relaciones entre datos. Rápida y eficaz pero insuficiente en espacios multidimensionales donde puedan relacionarse más de 2 variables.

Árboles de decisión.- Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial y el análisis predictivo, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema. Ejemplos:

Algoritmo ID3.

Algoritmo C4.5.

Modelos estadísticos.- Es una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales y en la regresión para indicar los diferentes factores que modifican la variable de respuesta.

Agrupamiento o *Clustering*.- Es un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia; se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes. Ejemplos:

Algoritmo K-means.

Algoritmo K-medoids.

Reglas de asociación.- Se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos.

Según el objetivo del análisis de los datos, los algoritmos utilizados se clasifican en supervisados y no supervisados

- Algoritmos supervisados (o predictivos): predicen un dato (o un conjunto de ellos) desconocido a priori, a partir de otros conocidos.
- Algoritmos no supervisados (o del descubrimiento del conocimiento): se descubren patrones y tendencias en los datos.

(MICROSOFT, 2016)

MINERIA DE SENTIMIENTOS

Una tarea básica de la minera de sentimientos es clasificar la polaridad de un texto dado a nivel de documento, oración, o rasgo/característica — si la opinión expresada en un documento, una oración o un rasgo/característica de una entidad es positiva, negativa, o neutral.

La clasificación de sentimiento más avanzada, "más allá de la polaridad" busca, por ejemplo, estados emocionales tales como "enfadado", "triste", y "feliz"

Dos enfoques

El problema se ha abordado principalmente desde dos enfoques diferentes; técnicas de aprendizaje computacional y aproximaciones semánticas. (AROMI, 2016).

Los **enfoques semánticos** se caracterizan por el uso de diccionarios de términos con orientación semántica de polaridad u opinión. Típicamente los sistemas preprocesan el texto y lo dividen en palabras, con la apropiada eliminación de las palabras de parada y una normalización lingüística por stemming o lematización, y luego comprueban la aparición de los términos del lexicon para asignar el valor de polaridad del texto mediante la suma del valores de polaridad de los términos.

Los sistemas además incluyen un tratamiento más o menos avanzado de:

a) términos modificadores (como muy, poco, demasiado) que aumentan o reducen la polaridad del o los términos a los que acompañan

b) términos inversores o negadores (como no,tampoco), que invierten la polaridad de los términos a los que afectan.

Los enfoques basados en **aprendizaje computacional** consisten en entrenar un clasificador usando un algoritmo de aprendizaje supervisado a partir de una colección de textos anotados, donde cada texto habitualmente se representa con un vector de palabras (bag of words), n-gramas o skip-grams, en combinación con otro tipo de características semánticas que intentan modelar la estructura sintáctica de las frases, la intensificación, la negación, la subjetividad o la ironía.

Pros y Contras

La ventaja principal de los enfoques semánticos es que los errores son relativamente sencillos de corregir, añadiendo cuantos términos fuera necesario, y se podría obtener una precisión tan alta como se quisiera, simplemente invirtiendo más tiempo en la construcción del lexicon. En este sentido, los enfoques basados en aprendizaje automático suelen ser una caja negra en la que corregir errores o añadir nuevo conocimiento es más complicado, y muchas veces sólo es posible ampliando la colección de textos y volviendo a entrenar el modelo.

Por otra parte, la ventaja de los enfoques basados en aprendizaje automático es que cuesta muy poco construir un analizador de sentimientos a partir de la colección de textos etiquetados, ya que la tarea de modelado reside en el algoritmo. Por ello es relativamente fácil construir clasificadores adaptados a un dominio determinado. En

contraposición, el esfuerzo para construir un lexicon para un cierto dominio, empezando de cero, es muy elevado, porque se basa en mucho trabajo manual, así que en general son menos adaptables.

(VILLENA, 2015)

Comprensión del contexto y del tono

El lenguaje humano es complejo. Enseñar a una máquina a analizar los diferentes matices gramaticales, variaciones culturales, jergas y faltas de ortografía de las menciones online es un proceso difícil. Y enseñar a una máquina a entender cómo el contexto puede afectar al tono, es aún más difícil.

Los humanos son bastante intuitivos al interpretar el tono de cualquier escrito.

Observa la siguiente frase: “Mi vuelo se retrasa. ¡Genial!”

La mayoría de los humanos podrían interpretar rápidamente que la persona está siendo sarcástica. Sabemos que para la mayoría de la gente un retraso en un vuelo no es una experiencia grata (a no ser de que haya barra libre como recompensa). Al aplicar este entendimiento contextual a la frase, podemos identificarla fácilmente como negativa. Sin este entendimiento contextual, una máquina que procesara esta frase vería la palabra “genial” y la categorizaría como positiva.

CONCLUSIÓN

En la lectura se observó que existen herramientas que nos pueden ayudar en el tratamiento de datos y su interpretación, sin embargo, cada empresa decidirá usar métodos y en específicamente el análisis de sentimiento para lo que mejor se ajuste a sus objetivos de negocio.

AGRADECIMIENTOS

Agradecida con Dios por todas sus bendiciones, igualmente por la oportunidad de trabajar en el proceso de mejorarme a misma.

A mi “alma mater” el Instituto Tecnológico de Orizaba por su esmero en la formación de profesionistas de calidad, a mi Profesor M.A.E Fernando Aguirre y Hernández por su dedicación, esmero y compromiso al compartir sus conocimientos.

¡A Dios por la vida y por la ciencia!

PROPUESTA DE TESIS

DISEÑO DE UNA METODOLOGIA PARA ANALISIS DE SENTIMIENTO

Objetivo: diseñar una estructura de análisis de sentimiento útil para la toma de decisiones estratégicas de las empresas.

BIBILIOGRAFIA

AROMI, D. (2016). *CONEXION INTAL*. Obtenido de http://www19.iadb.org/intal/conexionintal/2016/03/02/linguistica-computacional-y-subjetividad-en-la-teoria-economica-moderna/

BANNISTER, K. (02 de 2015). *BRANDWATCH*. Obtenido de https://www.brandwatch.com/es/2015/02/analisis-de-sentimiento/

GALEON. (2015). *TEXTMINING*. Obtenido de http://textmining.galeon.com/

MARCEL. (2015). *DATAMASHUP*. Obtenido de http://www.datamashup.info/the-driving-need-for-analytics-in-a-big-data-world/

MICROSOFT. (2016). *MSDN*. Obtenido de https://msdn.microsoft.com/es-mx/library/ms174949(v=sql.120).aspx

TRIPOD. (2015). Obtenido de TRIPOD: http://mineriadetextos.tripod.com/

VILLENA, J. (OCTUBRE de 2015). *MEANINGCLOUD*. Obtenido de https://www.meaningcloud.com/es/blog/introduccion-al-analisis-de-sentimientos-mineria-de-opinion