INSTITUTO TECNOLÓGICO DE ORIZABA

DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

• • •

MAESTRÍA EN INGENIERÍA ADMINISTRATIVA

• • •

FUNDAMENTOS DE INGENIERÍA ADMINISTRATIVA

• • •

TEMA:

“MINERÍAS DE DATOS, TEXTOS Y OPINIONES”

• • •

PRESENTA:

ING. JOSUÉ PACHECO ORTIZ

ORIZABA, VER. **SEPTIEMBRE 2016**

Minerías de Datos, Textos y Opiniones

Introducción

El desarrollo de la tecnología ha permito facilitar hasta cierto grado el trabajo de las personas en los diferentes sectores. Por ejemplo, cuando la gente cultivaba sus productos, todo era manual, desde la preparación del terreno, la siembra, riego, abono y cosecha. Hoy en día, todo ese trabajo es realizado por maquinas que se encargan de sustituir y ayudar a las personas y hacer un trabajo más rápido.

En el sector de la medicina, los robots han logrado grandes avances, llegando a operar en lugares incluso inaccesibles para el ser humano. Lo mismo sucede en las grandes fábricas, donde la tecnología ha llegado a reemplazar hasta cierto punto a la mano de obra, logrando una producción continua, sin cansancio, sin horas extras.

En el área de la administración, la tecnología apareció para la ayuda en la toma de decisiones, a través de un análisis de información, se pueden lograr hacer predicciones, tal y como se verá en este artículo.

El crecimiento explosivo de las bases de datos, de Internet y el empleo de técnicas y herramientas, que en forma automática y eficiente, generan información a partir de los datos almacenados, permiten descubrir patrones, relaciones y formular modelos. En particular, estas técnicas han adquirido enorme importancia en áreas tales como estrategias de marketing, soporte de decisiones, planeamiento financiero, análisis de datos científicos, bioinformática, análisis de textos y de datos de la web.

La tecnología llegó para quedarse y día a día, tratar de complementar y facilitar el trabajo a las personas.

Minería de Datos – Data Mining

Definición

La minería de datos, es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

Básicamente, el datamining surge para intentar ayudar a comprender el contenido de un repositorio de datos. Con este fin, hace uso de prácticas estadísticas y, en algunos casos, de algoritmos de búsqueda próximos a la inteligencia artificial y a las redes neuronales.

De forma general, los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación que surge entre la información y ese modelo represente un valor agregado, entonces nos referimos al conocimiento.

El datamining se presenta como una tecnología emergente, con varias ventajas: por un lado, resulta un buen punto de encuentro entre los investigadores y las personas de negocios; por otro, ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios. Además, no hay duda de que trabajar con esta tecnología implica cuidar un sinnúmero de detalles debido a que el producto final involucra "toma de decisiones".

Ventajas

1. Resulta un buen punto de encuentro entre los investigadores y las personas de negocios.

Este punto hace referencia a que aparece nueva tecnología la cual muchas veces es adquirida por empresas grandes las cuales financian estos proyectos.

2. Ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios.
   Prácticamente esto apoya al punto anterior ya que cuando un proyecto es bueno este es financiado por una empresa la cual adquiere más dinero del que invirtió y gracias a esta tecnología una empresa puede abrir otras oportunidades en el mercado.

3. Trabajar con esta tecnología implica cuidar un sin número de detalles debido a que el producto final involucra "toma de decisiones".
   Tienes la tecnología y esta te abrió paso en el mercado, también esta crea un producto el cual tu estas ofreciendo, pero hay que ver que tan efectiva fue la implementación la empresa ¿va creciendo o decreciendo? , ha eso hace referencia este punto.

4. Contribuye a la toma de decisiones tácticas y estratégicas proporcionando un sentido automatizado para identificar información clave desde volúmenes de datos generados por procesos tradicionales y de e-Business.

5. Permite a los usuarios dar prioridad a decisiones y acciones, mostrando factores que tienen un mayor en un objetivo, también muestra qué segmentos de clientes son desechables y qué unidades de negocio son sobrepasados y el ¿por qué?
   Hace referencia a que gracias a datamining solo hay que preocuparse de las tomas de decisiones ya que gracias a esta tecnología va mostrando las diversas ventajas y desventajas como son algunas señaladas en este punto.

6. Proporciona poderes de decisión a los usuarios del negocio que mejor entienden el problema y el entorno y es capaz de medir las acciones y los resultados de la mejor forma.

   Gracias a datamining se pueden dividir los problemas en distintos sectores y esto provocara que en diversos sectores deba haber distintos grupos de trabajo especializados en el ámbito de ese problema para así optimizar el tiempo y recursos.

7. Genera Modelos descriptivos: en un contexto de objetivos definidos en los negocios permite a empresas, sin tener en cuenta la industria o el tamaño, explorar automáticamente, visualizar y comprender los datos e identificar patrones, relaciones y dependencias que impactan en los resultados finales de la cuenta de resultados (tales como el aumento de los ingresos, incremento de los beneficios, contención de costes y gestión de riesgos).

8. Genera Modelos predictivos: permite que relaciones no descubiertas e identificadas a través del proceso del datamining sean expresadas como reglas de negocio o modelos predictivos. Estos outputs pueden comunicarse en formatos tradicionales (presentaciones, informes, información electrónica compartida, embebidos en aplicaciones, etc.) para guiar la estrategia y planificación de la empresa.

Técnicas

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.

Entre las más utilizadas están:

1. *Redes Neuronales*

   Esta técnica de inteligencia artificial, en los últimos años se ha convertido en uno de los instrumentos de uso frecuente para detectar categorías comunes

en los datos, debido a que son capaces de detectar y aprender complejos patrones, y características de los datos.

Una de las principales características de las redes neuronales, es que son capaces de trabajar con datos incompletos e incluso paradójicos, que dependiendo del problema puede resultar una ventaja o un inconveniente. Además esta técnica posee dos formas de aprendizaje: supervisado y no supervisado.

2. *Arboles de Decisión*

   Está técnica se encuentra dentro de una metodología de aprendizaje supervisado. Su representación es en forma de árbol en donde cada nodo es una decisión, los cuales a su vez generan reglas para la clasificación de un conjunto de datos.

   Los árboles de decisión son fáciles de usar, admiten atributos discretos y continuos, tratan bien los atributos no significativos y los valores faltantes. Su principal ventaja es la facilidad de interpretación.

3. *Algoritmos Genéricos*

   Los algoritmos genéticos imitan la evolución de las especies mediante la mutación, reproducción y selección, como también proporcionan programas y optimizaciones que pueden ser usadas en la construcción y entrenamiento de otras estructuras como es el caso de las redes neuronales. Además los algoritmos genéticos son inspirados en el principio de la supervivencia de los más aptos.

4. *Clustering*

   Agrupan datos dentro de un número de clases preestablecidas o no, partiendo de criterios de distancia o similitud, de manera que las clases sean similares entre sí y distintas con las otras clases. Su utilización ha proporcionado significativos resultados en lo que respecta a los clasificadores o reconocedores de patrones, como en el modelado de

sistemas. Este método debido a su naturaleza flexible se puede combinar fácilmente con otro tipo de técnica de minería de datos, dando como resultado un sistema híbrido.

5. *Aprendizaje Automático*
   Esta técnica de inteligencia artificial es utilizada para inferir conocimiento del resultado de la aplicación de alguna de las otras técnicas antes mencionadas.

Modelos de Minería de Datos

Un modelo de minería de datos se crea mediante la aplicación de un algoritmo a los datos, pero es algo más que un algoritmo o un contenedor de metadatos: es un conjunto de datos, estadísticas y patrones que se pueden aplicar a los nuevos datos para generar predicciones y deducir relaciones.

Aplicaciones de Modelos de Minería de Datos

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

1. Pronóstico: cálculo de las ventas y predicción de las cargas del servidor o del tiempo de inactividad del servidor.
2. Riesgo y probabilidad: elegir los mejores clientes para correspondencia, determinar el punto de equilibrio probable para escenarios de riesgo, asignación de probabilidades a diagnósticos u otros resultados de destino
3. Recomendaciones: determinación de los productos que se pueden vender juntos y generación de recomendaciones.
4. Búsqueda de secuencias: análisis de los artículos que los clientes han introducido en el carrito de la compra y predicción de posibles eventos.

5. Agrupación: distribución de clientes o eventos en grupos de elementos relacionados, y análisis y predicción de afinidades.

Generación de Modelos de Minería de Datos

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:

1. Definir el Problema

El primer paso del proceso de minería de datos 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 se 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, se podría tener que volver a definir el proyecto.

2. Preparar los Datos

El segundo paso del proceso de minería de datos consiste en consolidar y limpiar los datos identificados en el paso anterior.

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 se esperan.

Por consiguiente, antes de empezar a generar los modelos de minería de datos, se deben identificar estos problemas y determinar cómo se corregirán. 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 se necesiten usar herramientas de generación de perfiles de datos, y de limpieza y filtrado automático de datos para explorar los datos y buscar incoherencias.

3. Explorar los Datos

El tercer paso del proceso de minería de datos consiste en explorar los datos preparados. Se deben 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 podrían 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 dificultar el ajustar un modelo a los datos.

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

4. Generar Modelos

El cuarto paso del proceso de minería de datos consiste en generar el modelo o modelos de minería de datos.

Se debe definir qué columnas de datos se desea que se usen; para ello, se 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 se generan agregados y otra información estadística que se puede usar para el análisis.

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 se pueden utilizar los parámetros para ajustar cada algoritmo y se pueden 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.

5. Explorar y Validar los Modelos

El quinto paso del proceso de minería de datos 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.

6. Implementar y Actualizar los Modelos

El último paso del proceso de minería de datos 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, se pueden llevar acabo diferentes tareas, dependiendo de las necesidades. Las siguientes son algunas de las tareas que puede realizar:

a) Usar los modelos para crear predicciones que luego se podrán usar para tomar decisiones comerciales.
b) Crear consultas de contenido para recuperar estadísticas, reglas o fórmulas del modelo.
c) Crear un informe que permita a los usuarios realizar consultas directamente en un modelo de minería de datos existente.
d) Actualizar los modelos después de la revisión y análisis.
e) 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.

Minería de Textos – Text Mining

Es una de las ramas de la lingüística computacional que trata de obtener información y conocimiento a partir de conjuntos de datos que en principio no tienen un orden o no están dispuestos en origen para transmitir esa información
Es una técnica clave en un mundo como el actual en el que continuamente se recogen datos desde distintas perspectivas y de muchos aspectos diferentes de todas las actividades propias de los seres humanos.

El Text Mining no se debe confundir con la recuperación de la información, que es la recuperación automática de documentos relevantes mediante indexaciones de textos, clasificación, categorización, etc. La información que realmente le interesaría a la minería de textos es aquella contenida en esos documentos pero de manera general, es decir, no está contenida en un texto en concreto sino que es la información global que tienen todos los registros, textos, documentos... de la colección en común. Es un análisis de los datos compartidos por todos los textos de la colección que se ofrece de manera indirecta, es decir, son informaciones que la colección dará a los especialistas pero que no fue específicamente incluida en esa colección en el momento de su creación para su posterior difusión a los usuarios.

La Minería de Textos comprende tres 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.
- ✓ Por último se realizaría lo que antes se definía como minería de datos para encontrar asociaciones entre esos datos claves previamente extraídos de entre los textos

Aplicaciones

Es muy útil para todas las compañías, administraciones y organizaciones en general que por las características propias de su funcionamiento, composición y actividades generan gran cantidad de documentos y que están interesadas en obtener información a partir de todo ese volumen de datos. Les puede servir para conocer mejor a sus clientes, cuáles son sus hábitos, preferencias, etc.

Etapas

Es 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:

1. *Determinación de Objetivos*

   Aclarar que es lo que se está buscando con esta investigación, acotando hasta qué punto se quiere profundizar en la misma y definiendo claramente los limites.

2. *Preprocesamiento de los Datos*

   Es 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.

3. *Determinación del Modelo*

   Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse unas técnicas u otras.

4. *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.

Minería de Opiniones o Análisis de Sentimientos

La Minería de Opiniones se refiriere a una serie de aplicaciones de técnicas del procesamiento del lenguaje natural, lingüística computacional y minería de textos, que tienen como objetivo la extracción de información subjetiva a partir de contenidos generados por los usuarios, como puedan ser comentarios en blogs, o reviews de productos. Con este tipo de tecnologías, se puede extraer un valor tangible y directo, como pueda ser “positivo”/”negativo”, a partir de un comentario textual.

En líneas generales, existen dos tipos de tareas relacionadas con la Minería de Opiniones:

1. *Detección de la polaridad:* O lo que es lo mismo, ser capaces de determinar si una opinión es positiva o negativa. Más allá de una polaridad básica, también se puede querer obtener un valor numérico dentro de un rango determinado, que de una determinada forma trate de obtener un "rating" objetivo asociado a una determinada opinión.

2. *Análisis del sentimiento basado en características:* O lo que es lo mismo, ser capaces de determinar las distintas características del producto tratadas en la opinión o review escrita por el usuario, y para cada una de esas características mencionadas en la opinión, ser capaces de extraer una polaridad. Este tipo de acercamientos son mucho más complejos y de un grano mucho más fino que la detección de la polaridad.

Conclusión

Las Minerías de Datos, Textos y Opiniones son herramientas muy importantes para analizar la información de una empresa u organización y sirven para pronosticar en base a las tendencias que han estado presentes a lo largo de un periodo.

La tecnología aplicada en la administración, trata de proporcionar medios que faciliten el control de una organización, procurando prevenir errores que se pudieran dar.

Estas son las herramientas del presente y del futuro, por lo cual cada vez son utilizadas por más empresas y esto hace que cada día sean necesarias más personas especializadas en tema.

Bibliografía

- Microsoft (2014). Obtenido de:

  https://msdn.microsoft.com/es-es/library/ms174949.aspx

- Sinnexus. Obtenido de: http://www.sinnexus.com/business_intelligence/datamining.aspx
- Minería de Datos. Obtenido de: http://mineria-datos-actualidad.blogspot.mx/2012/06/por-que-usar-data-mining.html
- Minería de Textos. Obtenido de: http://textmining.galeon.com/

Este Artículo fue elaborado por Ing. Josué Pacheco Ortiz, bajo auspicio del Maestro Fernando Aguirre y Hernández, de la materia Fundamentos de Ingeniería Administrativa, de la Maestría en Ingeniería Administrativa, del Instituto Tecnológico Nacional de México, Campus Orizaba. Y apoyado bajo beca Conacyt.