"UNIVERSIDAD AUTÓNOMA GABRIEL RENE MORENO"

FACULTAD DE INGENIERÍA EN CIENCIAS DE LA COMPUTACIÓN Y TELECOMUNICACIONES

CARRERA EN INGENIERÍA EN SISTEMAS

MINERIA DE DATOS Y BIG DATA

Estudiante: Vanesa Vino Apaza

Registro: 220053243

Materia: Sistemas para el Soporte y la Toma de Decisiones

Docente: Ing. Miguel Peinado Pereira

Semestre: 2-2024

Sigla: INF-432

Grupo: SA

SANTA CRUZ – BOLIVIA

MINERIA DE DATOS Y BIG DATA

¿Qué es la minería de datos?
El Data Mining, también conocido como minería de datos, es un conjunto de técnicas que se realizan para explorar grandes cantidades de datos (Big Data). Encontrando patrones en los datos, el Data Mining nos puede ayudar a optimizar la toma de decisiones y la estrategia empresarial (Business Intelligence). Para alcanzar este objetivo, existen múltiples métodos matemáticos y estadísticos encapsulados en algoritmos, y que hoy en día también llamamos machine learning o inteligencia artificial.
El minado de datos es un conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos que expliquen el comportamiento de estos datos.

A pesar de que la idea del Data Mining puede parecer una innovación tecnológica muy reciente, en realidad este término apareció en los años sesenta conjuntamente con otros conceptos como, por ejemplo, el data fishing o data archeology. No obstante, no fue hasta los años ochenta cuando empezó su consolidación.

Su principal finalidad es explorar, mediante la utilización de distintas técnicas y tecnologías, bases de datos enormes de manera automática. El objetivo es encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos que se han ido recopilando con el tiempo. Estos patrones pueden encontrarse utilizando estadísticas o algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales.

Ventajas y desventajas del minado de datos
Los análisis de datos mediante el Data Mining pueden aportar numerosas ventajas a las empresas para la optimización de su gestión y tiempo, pero también para la captación y fidelización de clientes, que les permitirá aumentar sus ventas.

- Permite **descubrir información** que no esperábamos obtener. Esto se debe a su funcionamiento con algoritmos, ya que permite hacer muchas combinaciones distintas.
- Es capaz de **analizar bases de datos** con una enorme cantidad de datos.

- En 2024, con el uso creciente de algoritmos complejos de IA, las interpretaciones automáticas pueden ser útiles, pero los resultados de los modelos más avanzados pueden ser **difíciles de interpretar sin el uso de técnicas explicativas de IA.**
- Permite encontrar, atraer y **retener clientes**.
- La empresa puede mejorar la **atención al cliente** a partir de la información obtenida.
- Da a las empresas la posibilidad de **ofrecer a los clientes los productos** o servicios que necesitan.
- Antes de usar los modelos, estos son comprobados mediante estadísticas para verificar que las **predicciones** obtenidas son válidas.
- Ahorra costes a la empresa y abre nuevas oportunidades de negocio.

Sin embargo, también puede aparecer algún inconveniente a la hora de utilizar técnicas de Data Mining. Ahora, los costos de almacenamiento y procesamiento de datos en la nube han bajado significativamente debido a la competencia entre proveedores y la adopción masiva de servicios cloud escalables

Técnicas para el minado de datos

- **Asociación:** Se trata de una de las técnicas más utilizadas. En esta técnica, una transacción y la relación entre los elementos se utilizan para identificar un patrón. Esta es la razón por la que también se conoce como «técnica de relación». Se utiliza para realizar un análisis de la cesta de la compra, que se hace para conocer todos aquellos productos que los clientes compran juntos habitualmente, por ejemplo.
- **Agrupación o clustering:** Esta técnica crea agrupaciones de objetos significativos que comparten las mismas características. A menudo se confunde con la clasificación, pero si comprendes correctamente cómo funcionan estas dos técnicas no tendrás ningún problema. A diferencia de la clasificación, que coloca los objetos en clases predefinidas, la agrupación en clústeres coloca los objetos en clases definidas por nosotros.
- **Clasificación:** En 2024, las técnicas de clasificación y predicción han evolucionado con la incorporación de modelos de aprendizaje profundo (Deep Learning) y el procesamiento en tiempo real de datos. Las técnicas de clasificación ahora están

más automatizadas y precisas, mientras que las predicciones se basan en modelos más complejos que pueden analizar conjuntos de datos más variados y complejos, como datos no estructurados (imágenes, texto).

- **Predicción:** Esta técnica predice la relación que existe entre las variables independientes y dependientes, así como las variables independientes por sí solas. Puede usarse para predecir ganancias futuras dependiendo de la venta. Supongamos que la ganancia y la venta son variables dependientes e independientes, respectivamente. Ahora, basándonos en lo que dicen los datos de ventas pasadas, podemos hacer una predicción de ganancias del futuro con una curva de regresión.
- **Patrones secuenciales:** Esta técnica tiene como objetivo utilizar datos de transacciones y luego identificar tendencias, patrones y eventos similares en ellos durante un período de tiempo. Los datos históricos de ventas se pueden utilizar para descubrir artículos que los clientes compraron juntos en diferentes épocas del año. Las empresas pueden entender esta información recomendando a los clientes que compren esos productos en momentos en que los datos históricos no sugieren que lo harían. Las empresas pueden utilizar ofertas y descuentos para impulsar esta recomendación.

Cómo llevar a cabo un minado de datos

Los mineros o exploradores de datos a la hora de llevar a cabo un análisis de minería de datos, deberán realizar los siguientes pasos:

1. **Investigación comercial**

Antes de empezar, deberás tener una idea completa de los objetivos de tu empresa, de los recursos disponibles y de los diversos escenarios actuales en consonancia con los requisitos. Esto sería muy útil de cara a crear un plan detallado que alcance los objetivos de la organización.

2. **Análisis de calidad**

A medida que vamos recopilando los datos de distintas fuentes, necesitaremos verificarlos y compararlos para garantizar que no hayan cuellos de botella en el proceso de integración de datos. La garantía de calidad ayuda a detectar cualquier anomalía en los datos, como la interpolación de datos faltantes, manteniendo los datos en plena forma antes de que se sometan a una extracción.

3. **Limpieza de datos**

Se trata de la selección, limpieza, enriquecimiento, reducción y transformación de la base de datos. Se calcula que el 90% del tiempo en este tipo de procesos se gasta en este paso.

4. **Transformación de datos**

Este paso consta de cinco sub-etapas. Los procesos involucrados hacen que los datos estén listos en conjuntos de datos finales.

- **Suavizado de datos:** Se elimina el ruido de los datos

Resumen de datos: La agregación de conjuntos de datos se aplica en este proceso

- **Generalización de datos:** Los datos se generalizan reemplazando cualquier dato de bajo nivel con conceptualizaciones de nivel superior
- **Normalización de datos**: Los datos se definen en rangos establecidos
- **Construcción de atributos de datos:** Los conjuntos de datos deben estar en el conjunto de atributos antes de la minería de datos

5. **Modelado de datos**

Por último, para una mejor identificación de los patrones de datos, se implementan varios modelos matemáticos en el conjunto de datos, basados en varias condiciones.

Tipos de datos que pueden ser minados

- **Datos almacenados en una base de datos**

Una base de datos también puede denominarse sistema de gestión de bases de datos o DBMS. Cada DBMS almacena datos que están relacionados entre sí de una forma u otra.

También tiene un conjunto de programas de software que se utilizan para administrar datos y proporcionar un fácil acceso a ellos. Estos programas de software sirven para muchas cosas, incluida la definición de la estructura de la base de datos o asegurarse de que la información almacenada permanezca segura y consistente.

- **Data warehouse**

Un almacén de datos o data warehouse es una única ubicación de almacenamiento de datos que recopila datos de varias fuentes y luego

los almacena en forma de plan unificado. Cuando los datos se almacenan en estos sistemas se someten a una limpieza, integración, carga y actualización.

- **Data transaccional**

La base de datos transaccional almacena registros que se capturan como transacciones. Por ejemplo, reservas de vueltos, compras, clics en un sitio web, etc. Cada registro de transacciones tiene una identidad única. También engloba todos los elementos que la han convertido en una transacción.

¿Qué es el Big Data?

El Big Data, también denominado macrodatos o datos a gran escala, es un conjunto de tecnologías y herramientas que permiten trabajar con datos masivos. El concepto de Big Data está caracterizado por las 5 Vs. Un volumen extremo de datos, con una alta velocidad y variedad son las características más conocidas. No obstante, hablando de Big Data también tenemos que tomar en cuenta la calidad de los datos descrita por la veracidad y el valor que representan para la organización.

Características del big data

Para que la información pueda considerarse parte de los macrodatos debe cumplir algunos criterios que, además de la cantidad, tienen que ver con la calidad. Te dejamos cinco características que te ayudarán a identificar qué son los macrodatos dentro de todo el flujo de información:

1. **Volumen**

Aquí nos referimos a medidas que no pueden reducirse a simples gigabytes, sino a cantidades que exigen un almacenamiento mucho más complejo y con una capacidad de almacenamiento enorme a nivel terabyte.

2. **Veracidad**

Por supuesto, la información que abarca el big data debe ser verdadera y comprobable. No se trata de agregar números a capricho o inventados: su origen tiene que proceder de una fuente fidedigna y que se pueda consultar en cualquier momento.

3. **Velocidad**

Ahora tenemos la ventaja de capturar datos en tiempo real, así que una de las características del big data también es la velocidad con la que

pueden recolectarse. Al fin y al cabo muchos de ellos provienen de acciones en aplicaciones, sitios web o redes sociales.

4. **Variedad**

También es información que se obtiene de distintas fuentes, lo que permite que no haya sesgo pero tampoco límites de lo que se aprende del análisis. Más adelante, te presentamos cuáles son las principales fuentes del big data. Si están bien gestionados, los macrodatos nutren el conocimiento que un negocio tiene de su mercado, clientes, público, industria, etc.

5. **Valor**

Al final de su análisis, el big data ayudará a tomar acciones y a reconocer oportunidades para mejorar la empresa en varios de sus niveles. Si no se encuentra una utilidad a la información, ya sea para aumentar ventas, tener procesos eficientes, optimizar sus departamentos, acercarse más a sus clientes, atraer inversionistas o cualquier acción que signifique una mejora, entonces no tendrías por qué considerarlos parte de tus macrodatos.

Fuentes de big data

Como ya te adelantamos, los macrodatos provienen de diferentes fuentes. Algunas de ellas son:

1. **Las personas**

Entregan datos cuando envían un formulario, interactúan en sus redes sociales, mandan un correo electrónico o contestan un mensaje de texto.

2. **Las máquinas**

Nos referimos al GPS o una red de wifi, pero también incluye a un contador de energía eléctrica de un edificio, las cámaras de vigilancia en el centro de la ciudad o los que registran la entrada y salida de autos en un estacionamiento público. Aquí también entra el internet de las cosas (IoT), ya que estos dispositivos conectados a internet también generan datos valiosos y precisos de los usuarios.

3. **Estadísticas biométricas**

Provienen de la seguridad en distintos niveles y abarcan el registro de huellas dactilares, de voz o el reconocimiento de rostro.

4. **Estadísticas de marketing digital**

Suelen ser resultado de las acciones que los usuarios realizan cuando responden a un anuncio o cuando ingresan a un sitio web; las palabras clave que atraen más lectores en un blog pueden ser un ejemplo.

5. **Estadísticas de transacciones de datos**

Corresponden a las compras en línea, el envío de dinero por medios electrónicos, reservas de hotel o aviones, etcétera.

Categorías del big data

Además de las fuentes, el big data se divide en estas categorías:

1. **Datos estructurados**

Son los que tienen tamaño y formato definido (generalmente numérico) y pueden gestionarse con hojas de cálculo y bases de datos.

2. **Datos no estructurados**

Están desorganizados y no se encuentran dentro de un formato que jerarquice la información. Hablamos de archivos de texto o PDF, o información no numérica (como correos) que recopilamos de redes sociales, imágenes o videos.

3. **Datos semiestructurados**

Combinan los estructurados con los no estructurados porque, aunque la mayoría se encuentra desordenada o sin formato definido, es posible clasificar datos valiosos gracias al uso de etiquetas.

Características del big data

Para que la información pueda considerarse parte de los macrodatos debe cumplir algunos criterios que, además de la cantidad, tienen que ver con la calidad. Te dejamos cinco características que te ayudarán a identificar qué son los macrodatos dentro de todo el flujo de información:

1. **Volumen**

Aquí nos referimos a medidas que no pueden reducirse a simples gigabytes, sino a cantidades que exigen un almacenamiento mucho más complejo y con una capacidad de almacenamiento enorme a nivel terabyte.

2. **Veracidad**

Por supuesto, la información que abarca el big data debe ser verdadera y comprobable. No se trata de agregar números a capricho o

inventados: su origen tiene que proceder de una fuente fidedigna y que se pueda consultar en cualquier momento.

3. **Velocidad**

Ahora tenemos la ventaja de capturar datos en tiempo real, así que una de las características del big data también es la velocidad con la que pueden recolectarse. Al fin y al cabo muchos de ellos provienen de acciones en aplicaciones, sitios web o redes sociales.

4. **Variedad**

También es información que se obtiene de distintas fuentes, lo que permite que no haya sesgo pero tampoco límites de lo que se aprende del análisis. Más adelante, te presentamos cuáles son las principales fuentes del big data. Si están bien gestionados, los macrodatos nutren el conocimiento que un negocio tiene de su mercado, clientes, público, industria, etc.

5. **Valor**

Al final de su análisis, el big data ayudará a tomar acciones y a reconocer oportunidades para mejorar la empresa en varios de sus niveles. Si no se encuentra una utilidad a la información, ya sea para aumentar ventas, tener procesos eficientes, optimizar sus departamentos, acercarse más a sus clientes, atraer inversionistas o cualquier acción que signifique una mejora, entonces no tendrías por qué considerarlos parte de tus macrodatos.

Fuentes de big data

Como ya te adelantamos, los macrodatos provienen de diferentes fuentes. Algunas de ellas son:

1. **Las personas**

Entregan datos cuando envían un formulario, interactúan en sus redes sociales, mandan un correo electrónico o contestan un mensaje de texto.

2. **Las máquinas**

Nos referimos al GPS o una red de wifi, pero también incluye a un contador de energía eléctrica de un edificio, las cámaras de vigilancia en el centro de la ciudad o los que registran la entrada y salida de autos en un estacionamiento público. Aquí también entra el internet de las cosas (IoT), ya que estos dispositivos conectados a internet también generan datos valiosos y precisos de los usuarios.

4. **Estadísticas biométricas**

Provienen de la seguridad en distintos niveles y abarcan el registro de huellas dactilares, de voz o el reconocimiento de rostro.

5. **Estadísticas de marketing digital**

Suelen ser resultado de las acciones que los usuarios realizan cuando responden a un anuncio o cuando ingresan a un sitio web; las palabras clave que atraen más lectores en un blog pueden ser un ejemplo.

6. **Estadísticas de transacciones de datos**

Corresponden a las compras en línea, el envío de dinero por medios electrónicos, reservas de hotel o aviones, etcétera.

Categorías del big data

Además de las fuentes, el big data se divide en estas categorías:

1. **Datos estructurados**

Son los que tienen tamaño y formato definido (generalmente numérico) y pueden gestionarse con hojas de cálculo y bases de datos.

2. **Datos no estructurados**

Están desorganizados y no se encuentran dentro de un formato que jerarquice la información. Hablamos de archivos de texto o PDF, o información no numérica (como correos) que recopilamos de redes sociales, imágenes o videos.

3. **Datos semiestructurados**

Combinan los estructurados con los no estructurados porque, aunque la mayoría se encuentra desordenada o sin formato definido, es posible clasificar datos valiosos gracias al uso de etiquetas.