Minería de Datos

Conceptos:

Es un mecanismo de explotación, consistente en la búsqueda de información valiosa en grandes volúmenes de datos.

Otra definición: es el análisis de archivos y bitácoras de transacciones, trabaja a nivel del conocimiento con el fin de descubrir patrones, relaciones, reglas, asociaciones o incluso excepciones útiles para la toma de decisiones. La MD está muy ligada a los Data Warehouse

La Minería de Datos puede ser dividida en:

1. **Minería de datos predictiva (mdp):** usa primordialmente técnicas estadísticas.
2. **Minería de datos para el descubrimiento de conocimiento (mddc):** usa principalmente técnicas de inteligencia artificial.

Aplicaciones de la Minería de Datos.

Actualmente se aplica en áreas tales como:

1. **Aspectos climatológicos:** predicción de tormentas, etc.
2. **Medicina:** encontrar la probabilidad de una respuesta satisfactoria a un tratamiento médico.
3. **Mercadotecnia**: identificar clientes susceptibles de responder a ofertas de productos y servicios por correo, fidelidad de clientes, afinidad de productos, etc.
4. **Inversión en casas de bolsa y banca:** análisis de clientes, aprobación de préstamos, determinación de montos de crédito, etc.
5. **Detección de fraudes y comportamientos inusuales:** telefónicos, seguros, en tarjetas de crédito, de evasión fiscal, electricidad, etc.
6. **Análisis de canastas de mercado** para mejorar la organización de tiendas, segmentación de mercado (clustering).
7. **Determinación de niveles de audiencia de** programas televisivos.
8. **Industria y manufactura:** diagnóstico de fallas.

Técnicas de Minería de datos

1. **Análisis Preliminar de datos usando Query tools**: Es el 1º paso de un proyecto de MD, se aplica una consulta SQL al conjunto de datos, para rescatar algunos aspectos visibles antes de aplicar las técnicas.
2. **Técnicas de Visualización**: Son aptas para ubicar patrones en un conjunto de datos, puede usarse al comienzo de un proceso de MD para determinar la calidad de los datos.
3. **Redes neuronales artificiales**: Son modelos predecibles, no lineales que aprenden a través del entrenamiento.
4. **Reglas de Asociación:** Establecen asociaciones en base a los perfiles de los clientes sobre los cuales se realiza la MD.
5. **Algoritmos Genéticos:** Son técnicas de optimización que usan procesos tales como combinaciones genéticas y mutaciones, etc.
6. **Redes Bayesianas:** Buscan determinar relaciones causales que expliquen un fenómeno según los datos contenidos en una base de datos. Se han usado principalmente para realizar predicciones.
7. **Árbol de Decisión**: Son estructuras que representan conjuntos de decisiones, y estas decisiones generan reglas para la clasificación de un conjunto de datos.

Etapas de la Minería de Datos

La minería de datos se centra en llenar la necesidad de descubrir el por qué, para luego predecir y pronosticar las posibles acciones con cierto factor de confianza para cada predicción.

El análisis de archivos y bitácoras de transacciones, trabaja a nivel del conocimiento con el fin de descubrir **patrones**, relaciones, reglas, asociaciones o incluso excepciones útiles para la toma de decisiones

Estos patrones y tendencias se pueden recopilar y definir como un *modelo de minería de datos*. Los modelos de minería de datos se pueden aplicar en escenarios como los siguientes:

- **Pronóstico**: Cálculo de las ventas y predicción de las cargas del servidor o del tiempo de inactividad del servidor.
- **Riesgo y probabilidad:** Elección de los mejores clientes para la distribución de correo directo, determinación del punto de equilibrio probable para los escenarios de riesgo, y asignación de probabilidades a diagnósticos y otros resultados.

- **Recomendaciones:** Determinación de los productos que se pueden vender juntos y generación de recomendaciones.
- **Búsqueda de secuencias:** Análisis de los artículos que los clientes han introducido en el carrito de la compra y predicción de posibles eventos.
- **Agrupación:** Distribución de clientes o eventos en grupos de elementos relacionados, y análisis y predicción de afinidades.

Usuarios de la Minería de Datos

En la minería de datos se presentan los siguientes usuarios:

- Analistas Empresariales
- Peritos Estadísticos
- Profesionales en TI

Funciones de un Analista:

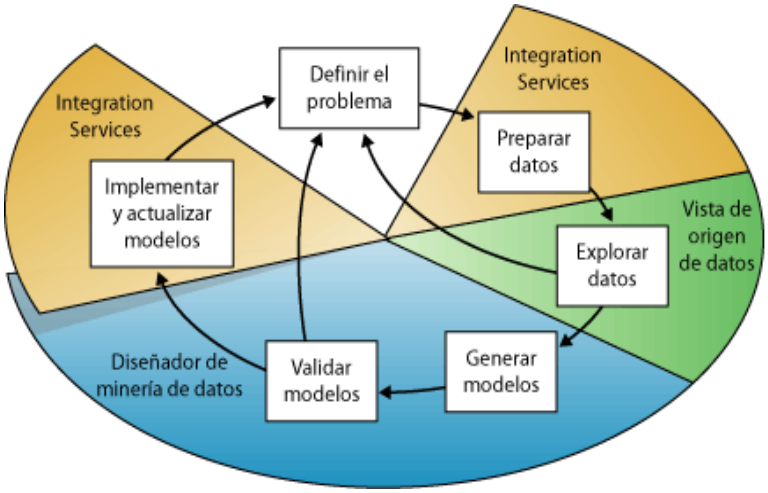
La generación de un modelo de minería de datos forma parte de un proceso mayor que incluye desde la formulación de preguntas acerca de los datos y la creación de un modelo para responderlas, hasta la implementación del modelo en un entorno de trabajo.

Este proceso se puede definir mediante los seis pasos básicos siguientes:

1. Definir el problema
2. Preparar los datos
3. Explorar los datos
4. Generar modelos
5. Explorar y validar los modelos

6. Implementar y actualizar los modelos

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



- 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.

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 (KPI) 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 del proceso de minería de datos, como se resalta en el siguiente diagrama, consiste en explorar los datos preparados.

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 dificultar 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.

Generar Modelos

Deberá 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. Para obtener más información acerca de cómo se relacionan las estructuras de minería de datos con los modelos de minería de datos, vea Arquitectura lógica (Analysis Services - Minería de datos).

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. Para obtener más información, vea Requisitos y consideraciones de procesamiento (minería de datos).

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.

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. SQL Server pone a su disposición el lenguaje DMX, que podrá usar para crear consultas de predicción, y el Generador de consultas de predicción, que le ayudará a generar las consultas. Para obtener más información, vea Referencia de Extensiones de minería de datos (DMX).
- Crear consultas de contenido para recuperar estadísticas, reglas o fórmulas del modelo. Para obtener más información, vea Consultas de minería de datos.
- Incrustar la funcionalidad de minería de datos directamente en una aplicación. Puede incluir Objetos de administración de análisis (AMO), que contiene un conjunto de objetos que la aplicación pueda utilizar para crear, cambiar, procesar y eliminar estructuras y modelos de minería de datos. También puede enviar mensajes XML for Analysis (XMLA) directamente a una instancia de Analysis Services. Para obtener más información, vea Development (Analysis Services - Data Mining).
- Utilizar Integration Services para crear un paquete en el que se utilice un modelo de minería de datos para dividir de forma inteligente los datos entrantes en varias tablas. Por ejemplo, si una base de datos se actualiza continuamente con clientes potenciales, puede utilizar un modelo de minería de datos junto con Integration Services para dividir los datos entrantes en clientes que probablemente compren un producto y clientes que probablemente no compren un producto. Para obtener más información, vea Typical Uses of Integration Services.
- Crear un informe que permita a los usuarios realizar consultas directamente en un modelo de minería de datos existente. Para obtener más información, vea Reporting Services en herramientas de datos de SQL Server (SSRS).
- Actualizar los modelos después de la revisión y análisis. Cualquier actualización requiere que vuelve a procesar los modelos. Para obtener más información, vea Procesar objetos de minería de datos.
- 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. Para obtener más información, vea Administración de las soluciones y los objetos de minería de datos.

Extensiones de la minería de datos

- **Web Mining:** Consiste en aplicar las técnicas de MD a documentos y servicios de la Web. Las herramientas de Web Mining analizan y procesan los logs para producir información significativa.
- **Text Mining:** Se refiere a examinar una colección de documentos y descubrir información no contenida en ningún documento individual de la colección. Dado que el 80 % de la información de una compañía se almacena en forma de documentos, existen técnicas que apoyan al TM.

Algunos software que usan la minería de datos

Clementine de SPSS.

Las organizaciones utilizan el conocimiento extraído con Clementine para:

- retener a los clientes rentables,
- identificar oportunidades de venta cruzada,
- detectar fraudes,
- reducir riesgos y mejorar la prestación de servicios a la administración,
- alcanzar un mayor nivel de conocimiento de sus clientes on line, y por lo tanto, mejorar el diseño de sus sitios web.

PolyAnalyst 4.5 de Megaputer.

http://www.megaputer.com

Megaputer: es líder en negocios y software inteligentes para Web. Ofrece las mejores herramientas para Data Mining, Text Mining y Web Mining.

Plataformas:

- Microsoft Windows XP/NT/2000
- Para UNIX y Linux 2001
- Además requiere la instalación de Microsoft Excel.

¿Por qué usar Minería de Datos?

- Ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios.
- Contribuye a la toma de decisiones tácticas y estratégicas.

- Proporciona poder de decisión a los usuarios del negocio, y es capaz de medir las acciones y Resultados de la mejor forma.
- Genera modelos descriptivos: permite a empresas, explorar y comprender los datos e Identificar patrones, relaciones y dependencias que impactan en los resultados finales.
- Genera modelos predictivos: permite que relaciones no descubiertas través del proceso del Dm sean expresadas como reglas de negocio.