

Sistemas de Data Mining

Sistemas de Data Warehousing

Arquitectura base.



Motivaciones

Problemática planteada:

- Acceso a Información para la toma de decisiones.



Sistemas de Data Warehousing

Algunos conceptos:

- Diccionario de Datos o Metadatos:
 - Asocia objetos del negocio a datos en BDs.
- Análisis multidimensional y herramientas OLAP:
 - Modelamiento del problema en dimensiones.
- Data Mining:
 - Búsqueda de correlaciones entre datos.
- Calidad de Datos
 - Se agregan criterios de Relevancia y Pertinencia de Datos.

La información y las organizaciones

Las organizaciones tienen necesidad de:

- Conocimiento:
 - Materia prima para toma de decisiones.
 - Es lo que se desea construir.
- Información:
 - Materia prima para conocer los fenómenos reales.
 - Un ítem de datos es información según el contexto de toma de decisiones.
- Datos:
 - Materia prima de la información.
 - Generados por procesos que no necesariamente los explotan.

Soportar múltiples tipos de usuarios

Diferentes niveles jerárquicos:

- Directivos.
- Gerentes de área.
- Mandos técnicos.

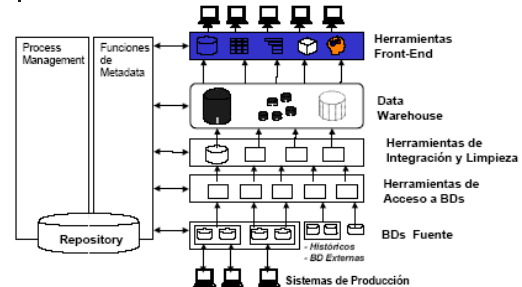
Diferentes funciones:

- Planificación.
- Control.
- Análisis.

Herramientas Consultas y Reportes

- **Funcionalidades base:**
 - Construir fácilmente consultas/reportes complejos.
 - Muy buenos para construir reportes no previstos.
 - Incorporan lenguajes para manejo de datos.
 - Incluyen funciones de todo tipo.
 - Ofrecen diferentes niveles de complejidad orientada a diferentes tipos de usuario:
 - Construcción de reporte complejo desde cero.
 - Construcción de reporte en base a templates.
 - Ejecución parametrizada de reportes.
 - Ejecución fija de reporte.

Data Mining en contexto DW



Herramientas OLAP

- **Funcionalidades base:**
 - Permiten consultar datos :
 - Interactivamente y en forma eficiente.
 - Usando mecanismos comprensibles para usuarios.
 - Una consulta corresponde a cruzar dimensiones y elegir la medida en el cruzamiento.
 - Funcionalidades adicionales:
 - Rankings.
 - Visualización gráfica.
 - Funcionalidades de herramientas:
 - Integración con BDs Relacionales.
 - Integración con herramientas de escritorio.
 - Interfaces tipo API.

Data Mining en contexto DW

- **Diferencias con OLAP.**
 - Data Mining usa mecanismos de:
 - Descubrimiento de información, Pattern-matching,
 - Deducción de reglas, ... y otros
- para determinar relaciones claves entre los datos.
- Los algoritmos de Data Mining pueden estudiar varias dimensiones de datos simultáneamente y descubrir los que tienen comportamiento especial.
 - La iniciativa es del algoritmo y no del usuario.

Data Mining

- **Objetivos:**
 - Explorar BDs buscando relaciones desconocidas entre los datos.
- **Por ejemplo:**
 - Relaciones entre enfermedades y decesos.
 - Algunas candidatas a nuevas causas de decesos.
 - Otras podrían ser datos erróneos.
- **Qué incluye ?**
 - Un conjunto muy amplio y heterogéneo de técnicas y herramientas.

Aplicación : criterios generales

- **Etapas en uso de DM:**
 - Identificación del problema.
 - Definición de la Estrategia de resolución.
 - Aplicación de DM para generar un Modelo.
 - Manipulación del Modelo obtenido.
 - Medición de resultados obtenidos.
- **DM provee feedback a otros procesos:**
 - Construcción del DW.
 - Estructuración de los datos.
 - Definición de indicadores.
 - Estructuración/Análisis de datos OLAP post-DM.
 - En base a resultados obtenidos.

Estrategias para Data Mining

- **Introducción.**
 - Las estrategias para Data Mining corresponden al tipo de estudio que se desea realizar.
 - Las estrategias no son algoritmos en si mismas, sino formas de encarar el problema planteado.
 - Cada estrategia generará un Modelo, a través de la ejecución de un algoritmo.
- **Algunas estrategias.**
 - Clasificación, Clustering, Asociación,
 - Optimización, Predicción.

Estrategias: Visualización

- **Objetivo.**
 - Representar situaciones de problema en forma visual, de forma de facilitar su análisis.
- **Ejemplo:**
 - Mostrar las distribuciones de ventas de productos en ciudades, teniendo en cuenta las características demográficas.
- **Observaciones.**
 - Se basa en técnicas de Interfase Hombre-Máquina y de comunicación de información en forma gráfica.

Estrategias : Clasificación

- **Objetivo:**
 - Clasificar registros según una variable objetivo, teniendo en cuenta valores de otros atributos.
- **Ejemplo:**
 - Se tiene una BD histórica con datos variados de clientes y un atributo de calificación de calidad (variable objetivo).
 - Dado un nuevo registro, del cual se desconoce su valor de variable objetivo, se quiere clasificar según los valores de los atributos.
- **Observaciones:**
 - Es de tipo aprendizaje dirigido, ya que se define la variable objetivo

Estrategias: Asociación

- **Objetivo.**
 - Generar reglas de tipo IF A1,...An THEN B, donde A1 ...,An son fenómenos en el problema.
- **Ejemplo:**
 - Se tiene una BD con tickets de supermercado. Y se quiere generar reglas que relacionen los productos comprados, hora de compra, día, mes, y perfil de cliente.
 - IF TipoCliente=1 AND CompraProd=p1 THEN CompraProd=P2;
- **Observaciones.**
 - También se lo llama Market Basket Análisis.

Estrategias: Clustering

- **Objetivo.**
 - Generar grupos con registros según su "similitud" en valores de atributos variados.
- **Ejemplo:**
 - Dada la BD del caso de Clasificación, generar grupos de clientes que tienen comportamiento similar sobre el conjunto de atributos.
- **Observaciones.**
 - Se trata de aprendizaje no-dirigido.
 - Se modela como un espacio n-dimensional de puntos, con una dimensión or atributo y un punto por registro.

Estrategias: Optimización.

- **Objetivo.**
 - Seleccionar una combinación de productos (o resultados) que mejor alcanza los objetivos de negocios.
- **Ejemplo:**
 - Lograr una combinación de cantidades producidas en diferentes productos que tienen sus costos y precios de venta.
- **Observaciones.**
 - Son casos de optimización lineal y no-lineal.

Estrategias: Estimación

- **Objetivo.**
 - Realizar clasificaciones pero con una variable objetivo continua y no discreta.
- **Ejemplo:**
 - Para el caso de los clientes, tomar como variable la ganancia esperada que generan.

Proceso: Definición de estudio

- **Definición.**
 - Consiste en definir los resultados a obtener, el tipo de estrategia y el alcance del estudio.
- **Aspectos a resolver:**
 - Definir los límites.
 - De qué se parte y qué se quiere obtener.
 - Elegir el tipo de estudio, incluyendo la estrategia.
 - Especificar los elementos a analizar.
 - Datos relevantes, valores resultados.
 - Definición de la muestra.
 - ¿ Como tomar una muestra representativa ?

El Proceso de Data Mining.

- **Introducción.**
 - Aplicar Data Mining corresponde más a un proceso que a una operación individual.
- **Pasos:**
 - Preparación de datos.
 - Definición de estudio.
 - Construcción de Modelo.
 - Entender y aplicar el Modelo.

Proceso: Construcción de Modelo

- **Definición.**
 - Consiste en construir un modelo abstracto que representa el problema y que manipulándolo se tratan de resolver los requerimientos.
- **Aspectos a resolver:**
 - Precisión (accuracy).
 - Comprensibilidad (understandability).
 - Qué entradas afectan la salida.
 - Por qué tiene éxito o falla.
 - Performance.
 - Qué tan rápido genera el modelo.
 - Qué tan rápido se obtienen las conclusiones deseadas.

Proceso: Preparación de datos.

- **Definición.**
 - Consiste en la generación de una base de datos sobre la cual se pueda aplicar el estudio deseado.
- **Aspectos a resolver:**
 - Limpieza de datos.
 - Valores nulos.
 - Derivación de datos.
 - Integración (merge) de datos.

Proceso: Entender y aplicar el Modelo

- **Definición.**
 - Consiste en asociar el modelo resultante al problema real de forma de comprenderlo.
- **Implica:**
 - Validar los resultados del modelo.
 - Extraer elementos relevantes y descartar las distorsiones.
 - Concluir qué fenómeno ocurre u ocurrir

Modelos y sus características

- **Modelos de Data Mining:**
 - Un *Modelo* es una representación de un problema que, instanciado con valores, genera resultados.
 - Por ejemplo: se tienen modelos predictivos, de clasificación, series de tiempo, clustering, etc.
 - Los modelos poseen ciertos atributos:
 - Underfitting y Overfitting.
 - Dirigido o no dirigido.
 - Explicabilidad de resultado.
 - Facilidad de aplicación.

Modelos y sus características

- **Explicabilidad.**
 - Resulta clarificante de interés conocer las razones que determinan los resultados.
 - Diferentes técnicas aportan distintos niveles de explicabilidad sobre sus resultados.
- **Facilidad de aplicación.**
 - Está asociado a la facilidad de uso, de comprensión de los resultados, de claridad de los resultados, de practicidad y conexión a bases de datos.

Modelos y sus características

- **Underfitting y Overfitting:**
 - **Overfitting:** más info que la deseable.
 - Todos los elementos se comportan como el set de entrenamiento (memorización del training set).
 - Se tiene información redundante dentro de los campos considerados, obteniendo un modelo trivial.
 - **Underfitting:** menos info que la deseable.
 - No se llegan a obtener patrones de interés sobre los datos (e.g. con bajo impacto predictivo).
 - Puede ser consecuencia de la des-actualización de modelos en el tiempo.

Algoritmos de Data Mining.


- **Introducción.**
 - El Modelo resultante del proceso de Data Mining es generado por algoritmos a través de productos de software.
- **Tipos de algoritmos.**
 - Árboles de Decisión.
 - Algoritmos Genéticos.
 - Redes Neuronales.
 - Estadísticos.
 - Algoritmos avanzados de asociación.
 - Algoritmos para Optimización.

Modelos y sus características

- **Dirigidos vs. No dirigidos.**
 - **Dirigidos:** la forma de la salida del modelo se especifica previo a su construcción.
 - El modelo se entrena sobre casos donde la salida está determinada (e.g. red neuronal con salida a estimar conocida).
 - **No dirigidos:** el propio modelo determina cuál será su salida.
 - Por ejemplo: estrategia de clustering donde el modelo son los clusters identificados.


Técnicas para Data Mining

- **La elección de una combinación particular de técnicas dependerá**
 - Problema a resolver / análisis DM. naturaleza de los datos disponibles.
 - Características conocidas sobre los tipos de
 - Modelos generados por las técnicas:
 - Underfitting & Overfitting
 - Dirigidos vs No dirigidos
 - Explicabilidad
 - Facilidad de aplicación




Data Mining

- **Síntesis.**
 - Area con fuertes componentes matemáticas.
 - Nuevos productos:
 - Accesibles en precio.
 - Explotables por usuarios no expertos.
 - Se prevee un gran impacto:
 - en el diseño de Sistemas DW.
 - en la explotación de Sistemas DW.
 - Todavía trabajo por hacer en la integración a los Sistemas DW.



Resumen

- **Herramientas Front-End:**
 - Tipos muy diferentes:
 - Desde planillas ... OLAP ... Data Mining.
 - OLAP es la actualmente dominante.
 - Data Mining es la emergente.
 - Enfoque :
 - Usabilidad por parte de usuarios finales.
 - Conexión a la Arquitectura del SDW.
- **Proceso de Carga y Actualización.**
 - Corresponde a la actividad más costosa.



*Caminante no hay Camino
se Hace Camino al Andar*

Gracias

