MINERÍA DE DATOS

Teleprocesos y Sistemas Distribuidos

Licenciatura en Sistemas

de Información

FACENA - UNNE

Octubre - 2003

CONTENIDO

- Qué es Data Warehousing
- Data Warehouse
- Objetivos del Data Warehouse
- Cómo trabaja el Data Warehouse
- En qué se lo puede usar
- Sistemas de Data Warehouse y Oltp
- El descubrimiento del conocimiento (KDD)
- Metas de KDD

- Técnicas de KDD
- Data Marts
- Minería de datos (MD)
- Aplicaciones de MD
- Técnicas de MD
- Algoritmos de MD
- Etapas principales del proceso de MD
- Extensiones de MD
- ¿Por qué usar Data Mining?
- Conclusiones

¿QUÉ ES DATA WAREHOUSING?

Es una **técnica** para **consolidar** y **administrar datos** desde variadas fuentes con el propósito de responder preguntas de negocios y tomar decisiones.

El proceso de *Data Warehousing* debe proveer:

- la información correcta,
- a la persona indicada,
- en el formato adecuado,
- y en el tiempo preciso.

Consolidar datos desde una variedad de fuentes → **Transformación de Datos**.

Manejar grandes volúmenes de datos → **Procesamiento y Administración de Datos**.

Acceder a los datos de una forma más directa, y analizarlos para obtener relaciones complejas entre los mismos → **Acceso a los Datos y Descubrimiento o Data Mining**.

Estos desarrollos tecnológicos, constituyen un ***Data Warehouse o Bodega de Datos***.

DATA WAREHOUSE

Es un sistema para el almacenamiento y distribución de cantidades masivas de datos.

Según ***Inmon*** (1992): "Un DW es una colección de datos integrados orientados a temas, integrados, no-volátiles y variables en el tiempo, organizados para soportar necesidades empresariales".

Susan Osterfeldt (1993): "Considero al DW como algo que provee dos beneficios empresariales reales: *Integración y Acceso a los datos*..."

Según, Bill Inmon, hay cuatro características que describen un almacén de datos:

- **orientado al sujeto**
- **variación-temporal**
- **integrados**
- **no son inestables**

OBJETIVOS DEL DATA WAREHOUSE

- Proveer una visión única de los clientes en toda la empresa.
- Poner tanta información comercial en manos de tantos usuarios diferentes como sea posible.
- Mejorar el tiempo de espera que insumen los informes habituales.
- Predecir compras de productos y aumentar la productividad.
- Mejorar la capacidad de respuesta a problemas comerciales.

¿CÓMO TRABAJA EL DATA WAREHOUSE?

- Extrae la información operacional.
- La transforma a formatos consistentes.
- Entrega la información al usuario.

¿EN QUÉ SE LO PUEDE USAR?

- Mejorar el proceso de toma de decisiones.
- Manejar las relaciones de marketing.
- Análisis de rentabilidad.
- Reducir costos o incrementar ingresos.

DATA MARTS

- Es un pequeño Data Warehouse, para un determinado número de usuarios, y para un área funcional específica de la compañía.

- Es un subconjunto de una bodega de datos para un propósito específico. Su función es apoyar a otros sistemas para la toma de decisiones.

SISTEMAS DE DATA WAREHOUSE Y OLTP

- Organización
  - Número de usuarios
    - Trabajo y Tiempo de procesamiento
      - Tamaño
        - Normalización
          - Diseño
            - Actualización
              - Estabilidad

El Descubrimiento de Conocimiento (KDD)

- Se define como "la extracción no trivial de información implícita, desconocida, y potencialmente útil de los datos".

- El proceso de KDD toma los resultados tal como vienen de los datos,
  - los transforma en información útil y entendible.

- KDD puede usarse como un medio de recuperación de información, de la misma manera que los agentes inteligentes realizan la recuperación de información en la Web.

METAS DE KDD

- procesar automáticamente grandes cantidades de datos crudos,
- identificar los patrones más significativos y relevantes,
- y presentarlos como conocimiento apropiado para satisfacer las metas del usuario.

TÉCNICAS DE KDD

Método de Clasificación.

- Es el más usado de todos los métodos de KDD.
- Agrupa los datos de acuerdo a similitudes o clases.
- Existen numerosas herramientas disponibles que son automatizadas.

Método Probabilístico.

- Utiliza modelos de representación gráfica.
- Se basa en las probabilidades e independencias de los datos.
- Puede usarse en los sistemas de diagnóstico, planeación y sistemas de control.

Método Estadístico.

- Usa la regla del descubrimiento y se basa en las relaciones de los datos.
- Es usado para generalizar los modelos en los datos y construir las reglas de los modelos nombrados.
- Por ejemplo: el proceso analítico en línea (OLAP).

Método Bayesian de KDD.

- Es un modelo gráfico que usa directamente los arcos para formar una gráfica acíclica.
- Se usa muy frecuentemente las redes de Bayesian cuando la incertidumbre se asocia con un resultado que puede expresarse en términos de una probabilidad.
- Este método es usado para los sistemas de diagnóstico.

MINERÍA DE DATOS

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 MD puede ser dividida en:

- minería de datos predictiva (mdp): usa primordialmente técnicas estadísticas.
- minería de datos para el descubrimiento de conocimiento (mddc): usa principalmente técnicas de inteligencia artificial.

APLICACIONES DE MD

Actualmente se aplica en áreas tales como:

- **aspectos climatológicos**: predicción de tormentas, etc.
- **medicina**: encontrar la probabilidad de una respuesta satisfactoria a un tratamiento médico.
- **mercadotécnia**: identificar clientes susceptibles de responder a ofertas de productos y servicios por correo, fidelidad de clientes, afinidad de productos, etc.
- **inversión en casas de bolsa y banca**: análisis de clientes, aprobación de prestamos, determinación de montos de crédito, etc.

- **detección de fraudes y comportamientos inusuales**: telefónicos, seguros, en tarjetas de crédito, de evasión fiscal, electricidad, etc.
- **análisis de canastas de mercado** para mejorar la organización de tiendas, segmentación de mercado (clustering).
- **determinación de niveles de audiencia** de programas televisivos.
- **industria y manufactura**: diagnóstico de fallas.

TÉCNICAS DE MD

- **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.
- **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.
- **Redes neuronales artificiales**: son modelos predecibles, no lineales que aprenden a través del entrenamiento.

- **Reglas de Asociación:** establecen asociaciones en base a los perfiles de los clientes sobre los cuales se realiza la MD.
- **Algoritmos Genéticos**: son técnicas de optimización que usan procesos tales como combinaciones genéticas y mutaciones, etc.
- **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.
- **Árbol de Decisión**: son estructuras que representan conjuntos de decisiones, y estas decisiones generan reglas para la clasificación de un conjunto de datos.

ALGORITMOS DE MINERÍA DE DATOS

- **supervisados o predictivos**: predicen el valor de un atributo de un conjunto de datos, conocidos otros atributos.

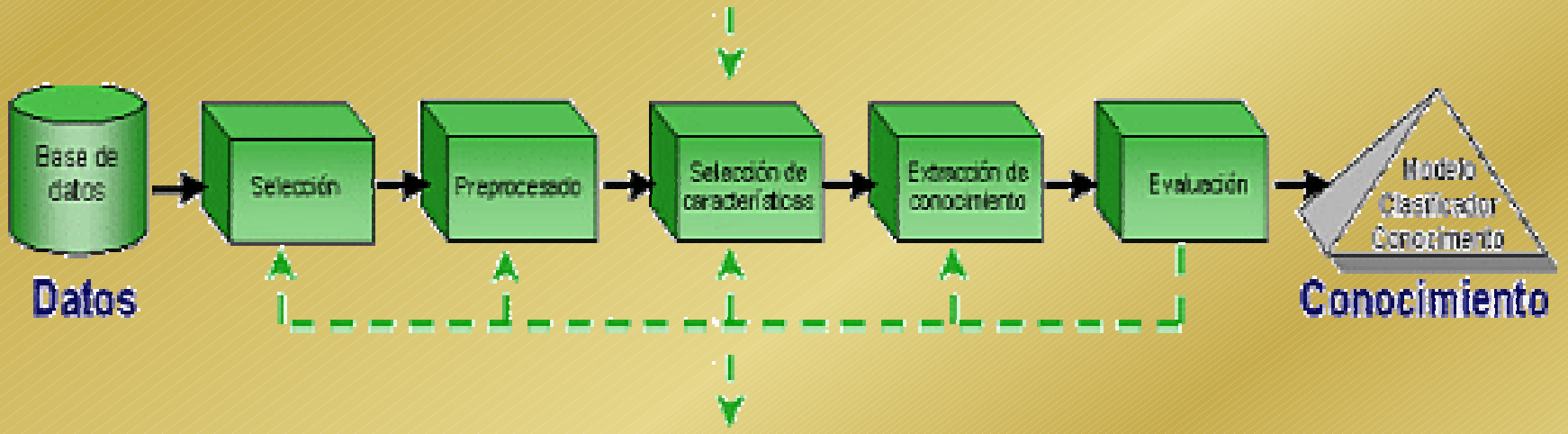
  A partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción de datos cuya etiqueta es desconocida.

- **no supervisados o del descubrimiento del conocimiento**: con estos algoritmos se descubren patrones y tendencias en los datos actuales. El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio de ellas.

ETAPAS PRINCIPALES DEL PROCESO DE MD

1. **Determinación de los objetivos**: delimitar los objetivos que el cliente desea bajo la orientación del especialista en Data Mining.
2. **Preprocesamiento de los datos**: se refiere a la selección, limpieza, enriquecimiento, reducción y la transformación de las bases de datos.
3. **Determinación del modelo**: se comienza con un análisis estadístico de los datos, y luego se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación.
4. **Análisis de los resultados**: verifica si los resultados obtenidos son coherentes y los compara con los obtenidos por el análisis estadístico y de visualización gráfica.

FASES DEL PROCESO



EXTENSIONES DEL DATA MINING

- **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

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 DATA MINING?

- 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.

CONCLUSIONES

El desarrollo de la tecnología de MD está en un momento crítico. Existen elementos que la hacen operable, pero por otra parte, hay factores que pueden crear un descrédito a esta tecnología, como ser:

- los productos a comercializar son, actualmente costosos, y los consumidores pueden hallar una relación costo/beneficio improductiva,

- se requiere de mucha experiencia para utilizar herramientas de la tecnología, o que sea muy fácil hallar patrones equívocos, triviales o no interesantes,

- la posibilidad de resolver los aspectos técnicos de hallar patrones en tiempo o espacio,

- además, hoy en día, las corporaciones comercializan con millones de perfiles personales, sin que aquellos a los que se refieren los datos intercambiados, estén en posibilidad de intervenir, entonces, se llega a pensar que presenta un peligro o riesgo para la privacidad de los clientes.

FIN

Apuntes en: www.exa.unne.edu.ar

Gracias por su atención.-