Minería de Datos

Universidad

Politécnica de Puebla

UPP

JESÚS ANTONIO GONZÁLEZ BERNAL

Minería de Datos

Evolución de la Tecnología BD

- 1960's y antes
  - Creación de las BD en archivos primitivos
- 1970's hasta principios de los 1980's
  - BD Jerárquicas y de Red
  - BD Relacionales
  - Herramientas de modelado de datos (Entidad-Relación)
  - Indexado y técnicas de organización (B-trees, Hashing)
  - Lenguajes de queries SQL, etc.
  - Interfaces de usuario y reportes
  - Procesamiento y optimización de queries
  - Manejo transacciones (recuperación, control concurrencia)
  - OLTP (On Line Transaction Processing)

Minería de Datos

Evolución de la Tecnología BD

- 1980’s (Mediados al presente)
  - Sistemas de BD Avanzados
    - Modelos de datos avanzados: Extended-Relational, OO, Object-Relational, Deductivo
  - Orientados a aplicaciones
    - Espaciales, temporales, multimedia, activos, cientificos, bases de conocimiento

Minería de Datos

Evolución de la Tecnología BD

- 1980's (Finales al presente)
  - Data warehouse y OLAP (On Line Analytical Processing)
  - Minería de datos y descubrimiento de conocimiento

- 1990's (al presente)
  - Sistemas basados en XML
  - Web mining

- 2000 (a la fecha)
  - NUEVA GENERACIÓN DE SISTEMAS DE INFORMACIÓN INTEGRADOS

Minería de Datos

¿Qué es la minería de datos?

La tarea no trivial de extraer información implícita, previamente desconocida y potencialmente útil de bases de datos (Frawley et. al. 1992).

Minería de Datos

¿Qué es la minería de datos?

El proceso de descubrir conocimiento interesante de grandes cantidades de datos almacenadas en bases de datos, data warehouses u otro repositorio de información (Jiawei Han, Micheline Kamber 2001).

Minería de Datos

¿Qué es la minería de datos?

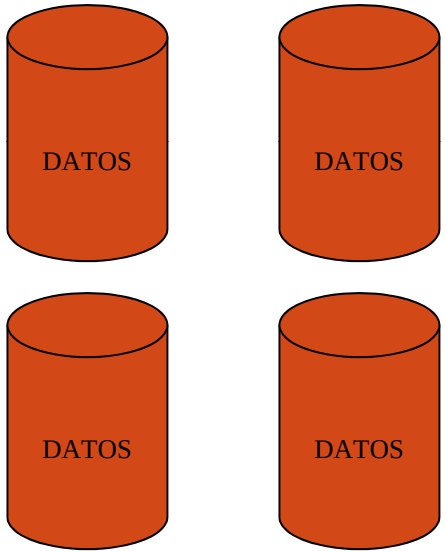
Sinónimos:

- Descubrimiento de Conocimiento en Bases de Datos
- Minería de conocimiento de bases de datos
- Extracción de conocimiento
- Análisis de datos y patrones
- Arqueología de datos

Minería de Datos

¿Cómo nació la minería de datos?

- Automatización de aplicaciones
- Uso de lectores de códigos de barras



Almacenaje masivo de información

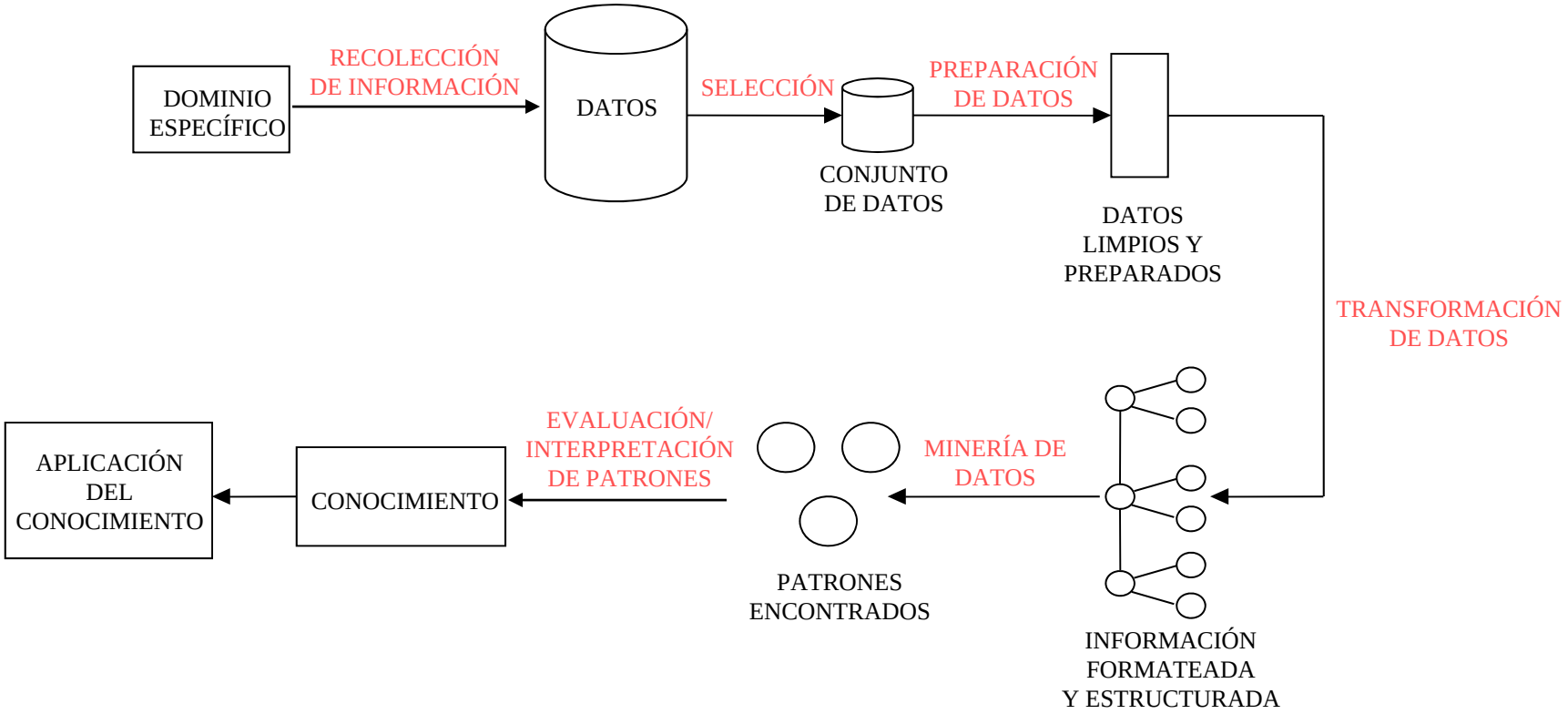
Necesidad de nuevas herramientas para analizar la información

Las herramientas estándares como la estadística y los queries a bases de datos no son suficientes

La minería de datos es parte del proceso de descubrimiento de conocimiento en bases de datos

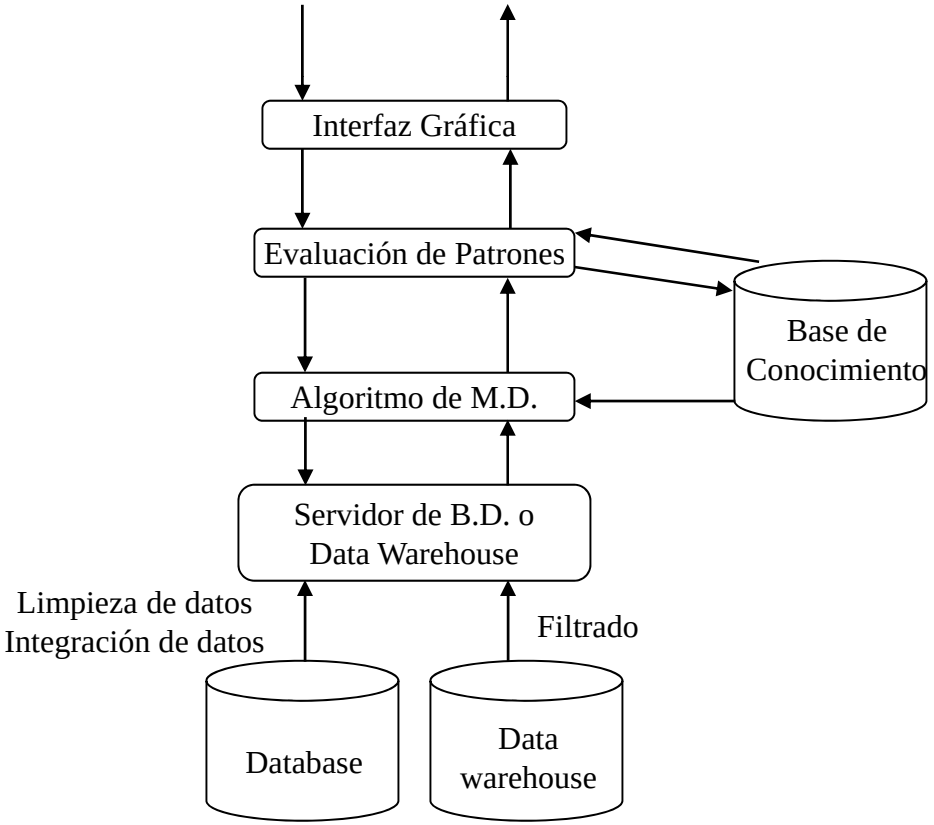
Minería de Datos

Proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD Process)



Minería de Datos

Arquitectura de un Sistema Típico de Minería de Datos



(Han and Kamber 2001)

Minería de Datos

Arquitectura de un Sistema Típico de Minería de Datos

- Base de datos
  - Puede ser: base de datos, datawarehouse, hoja de cálculo u otra clase de repositorio
  - A estos datos se le aplican técnicas de limpieza e integración

- Servidor de bases de datos
  - Utilizado para obtener la información relevante según el proceso de minería de datos

Minería de Datos

Arquitectura de un Sistema Típico de Minería de Datos

- Base de conocimiento
  - Conocimiento del dominio para guiar la búsqueda, evaluar que tan interesantes son los patrones
  - Creencias de los datos (del usuario: lo que se espera de los datos para descubrir comportamientos inesperados)
  - Umbrales de evaluación
  - Conocimiento previo
  - Meta-datos

Minería de Datos

Arquitectura de un Sistema Típico de Minería de Datos

- Algoritmo de minería de datos
  - Modular para realizar distintos tipos de análisis
    - Caracterización
    - Asociación
    - Clasificación
    - Análisis de grupos
    - Evolución (en espacio o tiempo)
    - Análisis de desviaciones

Minería de Datos

Arquitectura de un Sistema Típico de Minería de Datos

- Módulo de Evaluación de Patrones
  - Medidas de que tan interesante es un patrón
  - Interactúa con el algoritmo de M.D. para guiar la búsqueda hacia patrones interesantes

Minería de Datos

Arquitectura de un Sistema Típico de Minería de Datos

- Interfaz gráfica
  - Interacción con el usuario
  - Elección de la tarea de minería de datos
  - Proveer información para enfocar la búsqueda
  - Ayudar a evaluar los patrones
  - Explorar los patrones encontrados y la base de datos original
  - Visualizar los patrones en distintas formas

Minería de Datos

- Minería de datos
  - Utiliza métodos basados en:
    - Tecnología de Bases de Datos
    - Estadística
    - Aprendizaje automático
    - Cómputo de alto rendimiento
    - Reconocimiento de patrones
    - Redes Neuronales
    - Visualización de Datos
    - Recuperación de Información
    - Procesamiento de imágenes y señales
    - Análisis de Datos Espaciales

Minería de Datos

- Perspectiva de Bases de Datos
  - Énfasis en Eficiencia y Escalabilidad para grandes bases de datos
  - Algoritmo escalable
    - Tiempo de ejecución crece linealmente en proporción al tamaño de la base de datos dados los recursos disponibles de memoria principal y espacio en disco

Minería de Datos

- Repositorios de Datos
  - Base de Datos Relacional
    - DBMS
    - Lenguajes: (i.e., DDL, DML, DQL, etc.)
    - Tablas, atributos, tuplas
    - Modelo E-R
    - Normalización
    - Acceso a datos “Queries”
  - Análisis de datos
    - Tendencias
    - Patrones
    - Desviaciones

Minería de Datos

- Repositorios de Datos
  - Data Warehouse
    - Repositorio de información recopilada de varias fuentes bajo un esquema unificado y usualmente reside en un solo sitio
    - Construcción
      - Limpieza de datos
      - Transformación de datos
      - Integración de datos
      - Carga de los datos
      - Actualización periódica de los datos
    - Datos organizados sobre temas de alto nivel (cliente, proveedor, actividad, parte)

Minería de Datos

- Repositorios de Datos
  - Data Warehouse (continuación...)
    - Datos desde una perspectiva histórica (resúmenes de varios años)
    - Modelado sobre una estructura multidimensional
      - Cubos de datos
    - Análisis de Datos
      - OLAP
        - Utiliza información previa sobre el dominio para presentar los datos a diferentes niveles de abstracción (drill-down, roll-up para ver diferentes niveles de agrupación de información
      - Se requiere más análisis de datos

Minería de Datos

- Repositorios de Datos
  - Base de Datos Transaccional
    - Cada registro es una transacción
      - Número de transacción y lista de elementos de la transacción
    - Análisis de datos
      - Qué elementos se venden bien juntos?
      - “Market basket data analysis”

Minería de Datos

- Repositorios de Datos Avanzados
  - Bases de datos Orientadas a Objetos
  - Bases de datos Objeto-Relacionales
  - Bases de datos Espaciales
  - Bases de datos Temporales y de Series de Tiempo
  - Bases de datos de Texto
  - Bases de datos Multimedia
  - Bases de datos Heterogéneas
  - El World Wide Web

Minería de Datos

- ¿Para qué usamos la minería de datos?
  - Predecir
    - Utilizar algunas variables o campos en una base de datos para predecir valores desconocidos o futuros.
  - Describir
    - Encontrar patrones que describan la información (interpretables por el hombre)

Minería de Datos

Tareas de la minería de datos

- Clasificación
- Regresión
- Agrupamiento o clustering (encontrar clases)
- Sumarización (describir clases o conceptos)
- Modelos de dependencias
- Detección de cambios y desviaciones
- Asociación
- Análisis de Evolución (cambios en el tiempo)

Minería de Datos

Clasificación de Sistemas de M.D.

- Tipos de bases de datos sobre los que se hace M.D. (Datawarehouse, transaccional, relacional, OO, etc.)
- Tipo de conocimiento minado (caracterización, discriminación, asociación, clasificación, agrupamiento, etc.)
- Tipos de técnicas utilizadas (aprendizaje automático, estadística, visualización, ...)
- Adaptaciones para Aplicaciones (DNA, e-mail, etc...)

Minería de Datos

Componentes de un Algoritmo de Minería de Datos

- Modelo de representación
- Modelo de evaluación
- Método de búsqueda

Minería de Datos

Componentes de un Algoritmo de Minería de Datos

Modelo de Representación

- Lenguaje para describir los patrones
  - Árbol de decisiones
  - Lógica de primer grado
  - Gráfico

Minería de Datos

Componentes de un Algoritmo de Minería de Datos
Modelo de Evaluación

- Características del patrón encontrado
  - ¿Útil?
  - ¿Novedoso?
  - ¿Entendible?
  - ¿Efectivo para predecir?
  - ¿Medidas objetivas?
    - Soporte
    - Confianza

Minería de Datos

Componentes de un Algoritmo de Minería de Datos
Método de Búsqueda

- Búsqueda de parámetros
  - Para optimizar el modelo de evaluación
    - Parámetros de redes neuronales
    - Parámetro de espacio en “beam search”
- Búsqueda del modelo
  - Itera sobre la búsqueda de parámetros y elige el mejor resultado

Minería de Datos

Métodos de Minería de Datos

- Árboles de decisión y reglas
  - ID3, C4.5
- Regresión no lineal y métodos de clasificación
  - Redes Neuronales (Backprogagation)
- Métodos basados en ejemplos
  - Método del vecino más cercano
- Modelos gráficos de dependencias probabilísticas
  - Redes Bayesianas
- Modelos de aprendizaje relacional (ILP)
  - FOIL, Progol
- Asociaciones
  - Agrawal

Minería de Datos

¿Quiénes son los usuarios?

- Negocios --> Para construir modelos a partir de grandes bases de datos
  - Información transaccional
  - Datawarehouses
- Consumidores --> Para filtrar información de grandes bases de datos
  - Por ejemplo del Web
- Investigadores --> Para analizar grandes bases de datos

Minería de Datos

Aplicaciones de Minería de Datos

- Astronomía
  - Clasificación de estrellas y galaxias
- Análisis de Mercado y Administración
  - Perfil de clientes
    - ¿Qué tipos de clientes compran que productos?
      - Clasificación o Agrupamiento (clustering)
    - ¿Qué productor se compran normalmente juntos?
      - Reglas de asociación
  - Descubrir las relaciones entre características personales y el tipo de productos que se compran
  - Descubrir correlaciones entre compras

Minería de Datos

Más Aplicaciones de Minería de Datos

- Finanzas
  - Compañías de inversión hacen transacciones en la bolsa de valores basándose en resultados de Minería de Datos
  - Predicción de flujo de efectivo
- Detección de fraude
  - Utilizan bases de datos históricas para crear modelos de comportamiento fraudoliento y utilizar Minería de Datos para identificar nuevos fraudes.
    - Seguros de autos
    - Seguros médicos
    - Lavado de dinero
    - Telefónicos
    - Tratamiento médico inapropiado

Minería de Datos

Aun Más Aplicaciones de Minería de Datos

- Deportes
  - Para interpretar las estadísticas
- Web
  - Analizar logs en general
  - Analizar el comportamiento de los usuarios de un sitio
- E-mail
  - Clasificar e-mail y repartirlo al departamento adecuado
- Personalización
  - Hacer recomendaciones de acuerdo a características conocidas del usuario
- Recursos humanos
  - Ayudar a seleccionar empleados

Minería de Datos

Todavía Más Aplicaciones de Minería de Datos

- Bancos
  - Analizar clientes para otorgar crédito
- Medicina
  - Aplicaciones que buscan nuevos medicamentos
  - Análisis de secuencias de genes
  - Predecir si un compuesto causa cáncer
  - Análisis de secuencias de proteínas

Minería de Datos

Ejemplo

- ¿Será un buen día para jugar tenis?

Vista	Temperatura	Humedad	Viento	¿Jugar?
Soleado	Alta	Alta	Falso	No
Soleado	Alta	Alta	Verdadero	No
Nublado	Alta	Alta	Falso	Si
Lluvioso	Media	Alta	Falso	Si
Lluvioso	Baja	Normal	Falso	Si
Lluvioso	Baja	Normal	Verdadero	No
Nublado	Baja	Normal	Verdadero	Si
Soleado	Media	Alta	Falso	No
Soleado	Baja	Normal	Falso	Si
Lluvioso	Media	Normal	Falso	Si
Soleado	Media	Normal	Verdadero	Si
Nublado	Media	Alta	Verdadero	Si
Nublado	Alta	Normal	Falso	Si
Lluvioso	Media	Alta	Verdadero	No

Minería de Datos

Ejemplo

- 4 atributos
  - Vista: soleado, nublado o lluvioso
  - Temperatura: alta, media o baja
  - Humedad: alta o normal
  - Viento: falso o verdadero
- Espacio de búsqueda
  - 36 posibles combinaciones (3x3x2x2=36)

Minería de Datos

Ejemplo

- Reglas (Lista de decisiones)
  - Si Vista=Soleado y Humedad=Alta Entonces Jugar=No
  - Si Vista=Lluviosa y Viento=Verdadero Entonces Jugar=no
  - Si Vista=Nublado Entonces Jugar=Si
  - Si Humedad=Normal Entonces Jugar=Si
  - Si Ninguna de las otras reglas aplica Entonces Jugar=Si

Minería de Datos

¿Aprendizaje Automático o Minería de Datos?

- Dos comunidades
  - Bases de datos
  - Aprendizaje automático
- Manejo de grandes cantidades de datos
- ¿Cuántos datos se necesitan para hacer minería de datos?
  - PAC Learning
- Algoritmos eficientes y escalables

Minería de Datos

Retos en Minería de Datos

- Metodología de MD e interacción con el usuario
  - MD para diferentes tipos de conocimiento en bases de datos
  - MD interactiva de conocimiento a múltiples niveles de abstracción
  - Incorporación de conocimiento previo
  - Lenguajes de consultas de MD y MD ad hoc
  - Presentación y visualización de los resultados de MD
  - Manejo de datos ruidosos o incompletos
  - Evaluación de patrones

Minería de Datos

Retos en Minería de Datos

- Desempeño
  - Eficiencia y escalabilidad de los algoritmos de MD
  - Algoritmos de MD paralelos, distribuidos e incrementales
- Diversidad de los tipos de datos
  - Manejo de tipos de datos relacionales y complejos
  - MD de información de bases de datos heterogéneas y sistemas de información global