

SISTEMAS OPERATIVOS

TRABAJO MONOGRÁFICO REALIZADO POR:

**VERÓNICA S. BOGADO Y
MARIANA C. ARRUZAZABALA**

COMO ADSCRIPTAS A LA ASIGNATURA

“SISTEMAS OPERATIVOS”

SEPTIEMBRE - 2003

*Descubrimiento de
Conocimiento en Bases de
Datos
(KDD)*



*Minería de Datos
(MD)*

Introducción

De forma general, los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación de la información y ese modelo representen un valor agregado, entonces nos referimos al conocimiento. En la figura siguiente se ilustra la jerarquía que existe en una base de datos entre datos, información y conocimiento. Se observa igualmente el volumen que presenta en cada nivel y el valor que los responsables de las decisiones le dan en esa



jerarquía. El área interna dentro del triángulo representa los objetivos que se han propuesto. La separación del triángulo representa la estrecha unión entre dato e información, no así entre la información y el conocimiento.

La capacidad de generar y almacenar información creció considerablemente en los últimos tiempos, se ha estimado que la cantidad de datos en el mundo almacenados en bases de datos se duplica cada 20 meses. Es así que hoy las organizaciones tienen gran cantidad de datos almacenados y organizados, pero a los cuales no les pueden analizar eficientemente en su totalidad.

Con las sentencias SQL se puede realizar un primer análisis, aproximadamente el 80% de la información se obtiene con estas técnicas. El 20% restante, que la mayoría de las veces, contiene la información más importante, requiere la utilización de técnicas más avanzadas.

El Descubrimiento de Conocimiento en Bases de Datos (KDD) apunta a procesar automáticamente grandes cantidades de

datos para encontrar conocimiento útil en ellos, de esta manera permitirá al usuario el uso de esta información valiosa para su conveniencia.

Descubrimiento de Conocimiento en Bases de Datos (KDD)

El **KDD** es el

"Proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos ". (Fayyad et al., 1996)

El **objetivo fundamental** del KDD es encontrar conocimiento útil, válido, relevante y nuevo sobre un fenómeno o actividad mediante algoritmos eficientes, dadas las crecientes órdenes de magnitud en los datos. Al mismo tiempo hay un profundo interés por presentar los resultados de manera visual o al menos de manera que su interpretación sea muy clara. Otro aspecto es que la interacción humano-máquina deberá ser flexible, dinámica y colaboradora. El resultado de la exploración deberá ser interesante y su calidad no debe ser afectada por mayores volúmenes de datos o por ruido en los datos. En este sentido, los algoritmos de descubrimiento de información deben ser altamente robustos.

Metas

Las metas del KDD son:

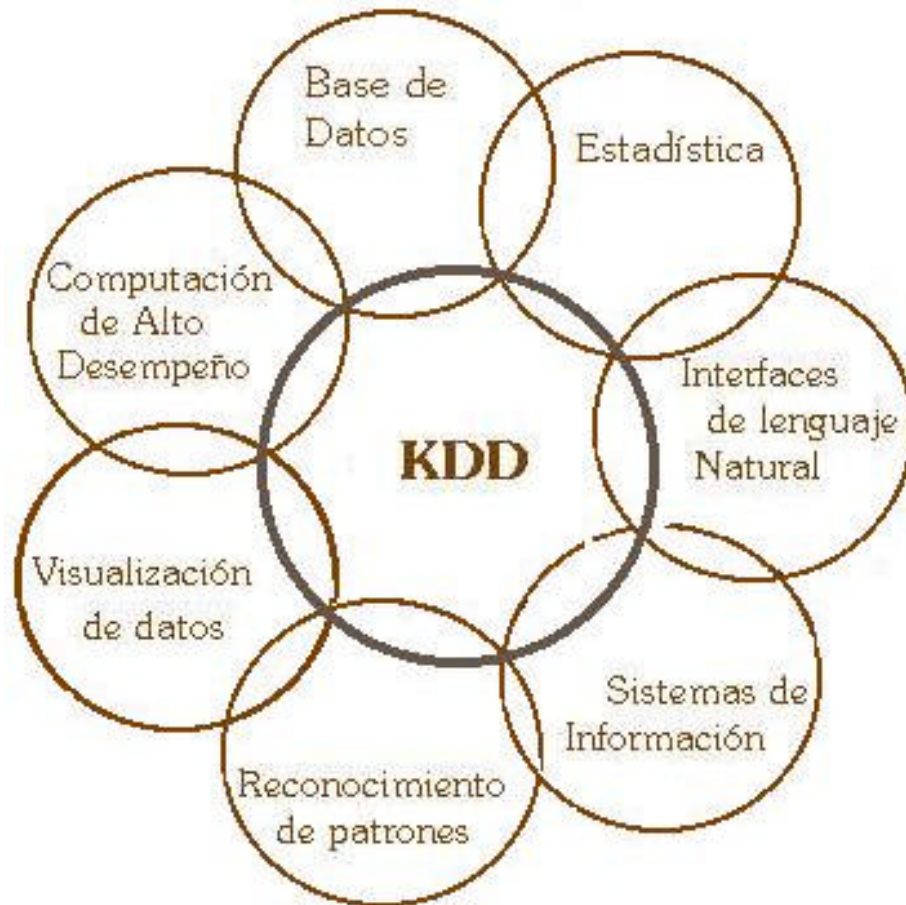
- Procesar automáticamente grandes cantidades de datos crudos.
- Identificar los patrones más significativos y relevantes.
- Presentarlos como conocimiento apropiado para satisfacer las metas del usuario.

Relación con otras disciplinas

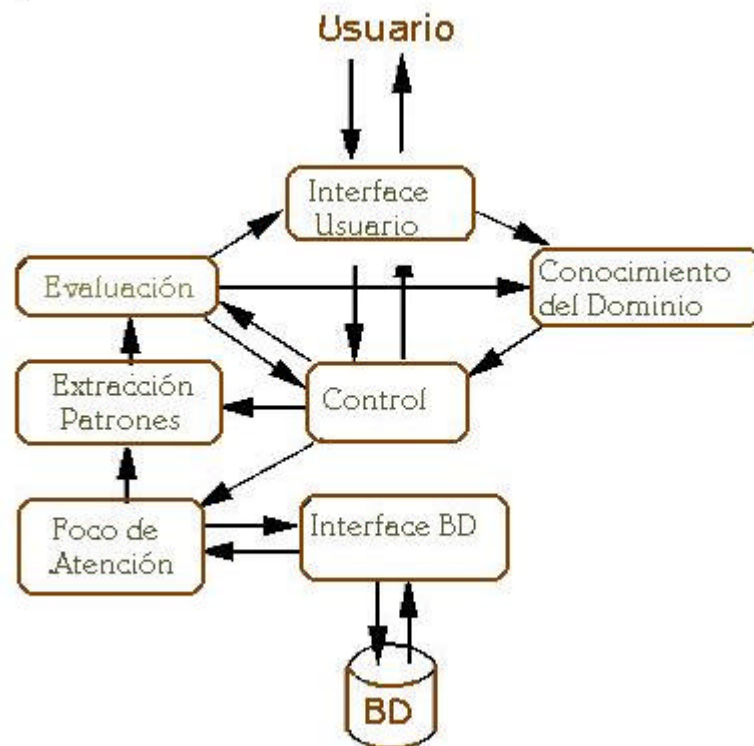
KDD nace como interfaz y se nutre de diferentes disciplinas:

- ***Sistemas de información / bases de datos:*** tecnologías de bases de datos y bodegas de datos, maneras eficientes de almacenar, acceder y manipular datos.
- ***Estadística, aprendizaje automático / IA*** (redes neuronales, lógica difusa, algoritmos genéticos, razonamiento probabilístico): desarrollo de técnicas para extraer conocimiento a partir de datos.
- ***Reconocimiento de patrones:*** desarrollo de herramientas de clasificación.
- ***Visualización de datos:*** interfaz entre humanos y datos, y entre humanos y patrones.
- ***Computación paralela / distribuida:*** cómputo de alto desempeño, mejora de desempeño de algoritmos debido a su complejidad y a la cantidad de datos.
- ***Interfaces de lenguaje natural a bases de datos.***

Gráficamente éstas relaciones pueden ser representadas de la siguiente manera:



Componentes



- **Conocimiento del dominio y preferencias del usuario:** Incluye el diccionario de datos, información adicional de las estructuras de los datos, restricciones entre campos, metas o preferencias del usuario, campos relevantes, listas de clases, jerarquías de generalización, modelos causales o funcionales, etc.
 - El objetivo del conocimiento del dominio es orientar y ayudar en la búsqueda de patrones interesantes (aunque a veces puede causar resultados contraproducentes).
 - Se tiene que hacer un balance entre eficiencia y completitud del conocimiento.

- **Control del descubrimiento:** Toma el conocimiento del dominio, lo interpreta y decide qué hacer (en la mayoría de los sistemas el control lo hace el usuario).

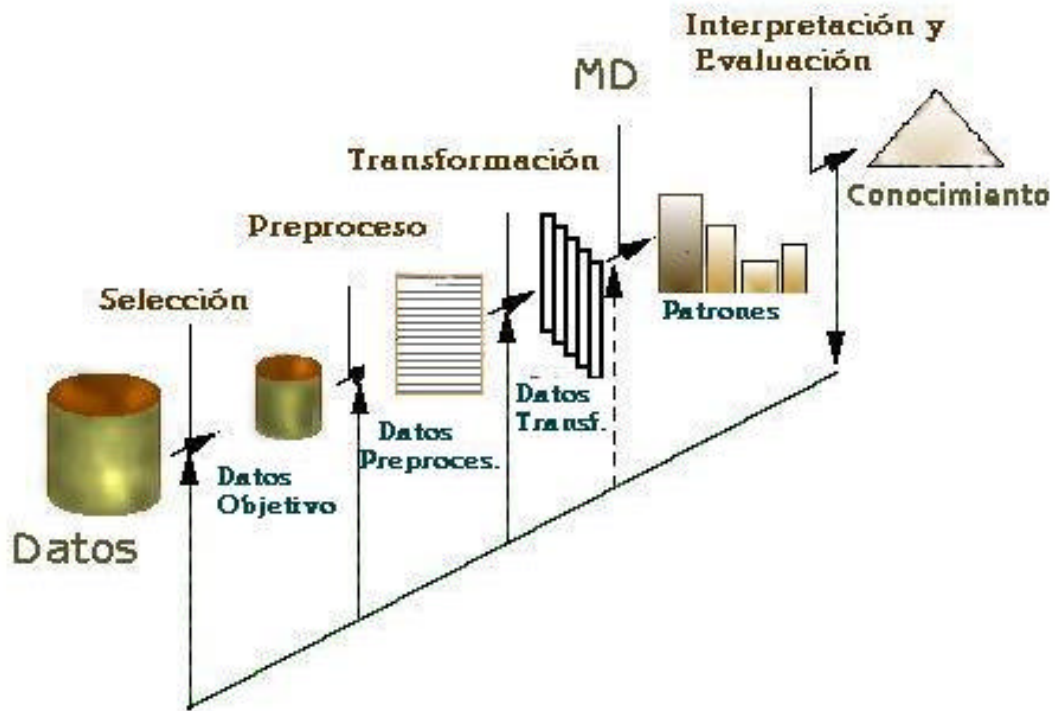
- **Interfaces:** Con la base de datos y con el usuario.

- **Foco de atención:** Especifica qué tablas, campos y registros accesar. Tiene que tener mecanismos de selección aleatoria de registros tomando muestras estadísticamente significativas, puede usar predicados para seleccionar un subconjunto de los registros que comparten cierta característica, etc.
 - Algunas técnicas para enfocar la atención incluyen:
 - Agregación: junta valores (por ejemplo, los más bajos y los más altos)
 - Partición de datos: sobre la base de valores de atributos (por ejemplo, sólo aquellos datos que tengan ciertos valores)
 - Proyección: ignorar algún(os) atributo(s)
 - Partición y proyección implican menos dimensiones. Agregación y proyección implican menos dispersión.

- **Extracción de patrones:** Donde patrón se refiere a cualquier relación entre los elementos de la base de datos. Pueden incluir medidas de incertidumbre. Aquí se aplican una gran cantidad de algoritmos de aprendizaje y estadísticos.

- **Evaluación:** Un patrón es interesante en la medida que sea confiable, novedoso y útil respecto al conocimiento y los objetivos del usuario. La evaluación normalmente se le deja a los algoritmos de extracción de patrones que generalmente están basados en significado estadístico (sin embargo, no es ni debe ser el único criterio).

El proceso de KDD



El proceso de KDD consiste en usar métodos de minería de datos (algoritmos) para extraer (identificar) lo que se considera como conocimiento de acuerdo a la especificación de ciertos parámetros usando una base de datos junto con pre-procesamientos y post-procesamientos.

Se estima que la extracción de patrones (minería) de los datos ocupa solo el 15% - 20% del esfuerzo total del proceso de KDD.

El proceso de descubrimiento de conocimiento en bases de datos involucra varios pasos:

- **Determinar las fuentes de información:** que pueden ser útiles y dónde conseguirlas.
- **Diseñar el esquema de un almacén de datos (Data Warehouse):** que consiga unificar de manera operativa toda la información recogida.

- **Implantación del almacén de datos:** que permita la "navegación" y visualización previa de sus datos, para discernir qué aspectos puede interesar que sean estudiados. Esta es la etapa que puede llegar a consumir el mayor tiempo.

- **Selección, limpieza y transformación de los datos que se van a analizar:** la selección incluye tanto una criba o fusión horizontal (filas) como vertical (atributos). La limpieza y preprocesamiento de datos se logra diseñando una estrategia adecuada para manejar ruido, valores incompletos, secuencias de tiempo, casos extremos (si es necesario), etc.

- **Seleccionar y aplicar el método de minería de datos apropiado:** esto incluye la selección de la tarea de descubrimiento a realizar, por ejemplo, clasificación, agrupamiento o clustering, regresión, etc. La selección de él o de los algoritmos a utilizar. La transformación de los datos al formato requerido por el algoritmo específico de minería de datos. Y llevar a cabo el proceso de minería de datos, se buscan patrones que puedan expresarse como un modelo o simplemente que expresen dependencias de los datos, el modelo encontrado depende de su función (clasificación) y de su forma de representarlo (árboles de decisión, reglas, etc.), se tiene que especificar un criterio de preferencia para seleccionar un modelo dentro de un conjunto posible de modelos, se tiene que especificar la estrategia de búsqueda a utilizar (normalmente está predeterminada en el algoritmo de minería).

- **Evaluación, interpretación, transformación y representación de los patrones extraídos:** Interpretar los resultados y posiblemente regresar a los pasos anteriores. Esto puede involucrar repetir el proceso, quizás con otros datos, otros algoritmos, otras metas y otras estrategias. Este es un paso crucial en donde se requiere tener conocimiento del dominio. La interpretación puede beneficiarse de procesos de visualización, y sirve también para borrar patrones redundantes o irrelevantes.

- **Difusión y uso del nuevo conocimiento.**

- Incorporar el conocimiento descubierto al sistema (normalmente para mejorarlo) lo cual puede incluir resolver conflictos potenciales con el conocimiento existente.
- El conocimiento se obtiene para realizar acciones, ya sea incorporándolo dentro de un sistema de desempeño o simplemente para almacenarlo y reportarlo a las personas interesadas.

En este sentido, KDD implica un proceso interactivo e iterativo involucrando la aplicación de varios algoritmos de minería de datos.

Minería de datos

La Minería de Datos es la *etapa de descubrimiento* en el proceso de KDD:

«paso consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos preprocesados» (Fayyad et al., 1996)

Aunque se suelen usar indistintamente los términos KDD y Minería de Datos.

Principales características y objetivos de la Minería de Datos

- Explorar los datos se encuentran en las profundidades de las bases de datos, como los almacenes de datos, que algunas veces contienen información almacenada durante varios años.
- En algunos casos, los datos se consolidan en un almacén de datos y en mercados de datos; en otros, se mantienen en servidores de Internet e Intranet.
- El entorno de la minería de datos suele tener una arquitectura cliente-servidor.
- Las herramientas de la minería de datos ayudan a extraer el "mineral" de la información enterrado en archivos corporativos o en registros públicos, archivados
- El "minero" es, muchas veces un usuario final con poca o ninguna habilidad de programación, facultado por "barrenadoras de datos" y otras poderosas herramientas indagatorias para efectuar preguntas ad hoc y obtener rápidamente respuestas.
- "Hurgar y sacudir" a menudo implica el descubrimiento de resultados valiosos e inesperados.
- Las herramientas de la minería de datos se combinan fácilmente y pueden analizarse y procesarse rápidamente.
- Debido a la gran cantidad de datos, algunas veces resulta necesario usar procesamiento en paralelo para la minería de datos.
- La minería de datos produce cinco tipos de información:

- asociaciones
 - secuencias
 - clasificaciones
 - agrupamientos
 - pronósticos
- Los mineros de datos usan varias herramientas y técnicas.

La minería de datos es un proceso que invierte la dinámica del método científico en el siguiente sentido:

En el método científico, primero se formula la hipótesis y luego se diseña el experimento para coleccionar los datos que confirmen o refuten la hipótesis. Si esto se hace con la formalidad adecuada (cuidando cuáles son las variables controladas y cuáles experimentales), se obtiene un nuevo conocimiento.

En la minería de datos, se coleccionan los datos y se espera que de ellos emerjan hipótesis. Se busca que los datos describan o indiquen por qué son como son. Luego entonces, se valida esa hipótesis inspirada por los datos en los datos mismos, será numéricamente significativa, pero experimentalmente inválida. De ahí que la minería de datos debe presentar un enfoque exploratorio, y no confirmador. Usar la minería de datos para confirmar las hipótesis formuladas puede ser peligroso, pues se está haciendo una inferencia poco válida.

La minería de datos es una tecnología compuesta por etapas que integra varias áreas y que no se debe confundir con un gran software. Durante el desarrollo de un proyecto de este tipo se usan diferentes aplicaciones software en cada etapa que pueden ser estadísticas, de visualización de datos o de inteligencia artificial, principalmente. Actualmente existen aplicaciones o herramientas comerciales de minería de datos muy poderosas que contienen un sinfín de utilerías que facilitan el desarrollo de un proyecto. Sin embargo, casi siempre acaban complementándose con otra herramienta.

Historia

La idea de minería de datos no es nueva. Ya desde los años sesenta los estadísticos manejaban términos como data fishing, data mining o data archaeology con la idea de encontrar

correlaciones sin una hipótesis previa en bases de datos con ruido. A principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de minería de datos y KDD. A finales de los años ochenta sólo existían un par de empresas dedicadas a esta tecnología; en 2002 existen más de 100 empresas en el mundo que ofrecen alrededor de 300 soluciones. Las listas de discusión sobre este tema las forman investigadores de más de ochenta países. Esta tecnología ha sido un buen punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios.

La evolución de sus herramientas en el transcurso del tiempo puede dividirse en cuatro etapas principales:

- Colección de datos (1960)
- Acceso de datos (1980)
- Almacén de datos y apoyo a las decisiones (principios de la década de 1990)
- Minería de datos inteligente (finales de la década de 1990)

Aplicaciones actuales de la Minería de Datos

- Análisis de la cesta de la compra mediante reglas de asociación.
- Modelos para análisis de riesgos (seguros, créditos...).
- Evaluación de campañas publicitarias.
- Análisis de la fidelidad de clientes (*churning*).
- Análisis de valores de bolsa.
- Detección y prevención de fraude en comercio electrónico.
- Modelos de tráfico a partir de datos GPS.
- Perfiles de usuarios de redes.
- Detección de intrusos en redes.

Herramientas y técnicas del KDD

Existen muchas metodologías del descubrimiento del conocimiento en uso y bajo desarrollo. Algunas de estas técnicas son genéricas, mientras otros son de dominio específico.

Características básicas que comparten todas las técnicas KDD :

- **Grandes cantidades de datos:** para poder derivar un conocimiento adicional.

- **Eficiencia debido al volumen de datos.**

- **Exactitud:** es un elemento esencial para asegurar que el descubrimiento del conocimiento es válido

- **Uso de un lenguaje de alto nivel:** los resultados deberán ser presentados de una manera entendible para el ser humano

- **Uso de alguna forma de aprendizaje automatizado:** técnicas de aprendizaje inteligente que van examinando los datos a través de procesos automatizados.

- **Producción de resultados interesantes:** debe tener un valor potencial para el usuario.

KDD proporciona la capacidad para descubrir información nueva y significativa usando los datos existentes. KDD rápidamente excede la capacidad humana para analizar grandes cantidades de datos. La cantidad de datos que requieren procesamiento y análisis en grandes bases de datos exceden las capacidades humanas y la dificultad de transformar los datos con precisión es un conocimiento que va más allá de los límites de las bases de datos tradicionales. Por consiguiente, la utilización plena

de los datos almacenados depende del uso de técnicas del descubrimiento del conocimiento.

Técnicas de KDD

Las técnicas de aprendizaje podrán ser supervisadas o no supervisadas. En general, las técnicas de aprendizaje dirigidas disfrutan de un rango de éxito definido por la utilidad del descubrimiento del conocimiento.

KDD típicamente combina métodos automatizados con la interacción humana para asegurar resultados exactos, útiles, y entendibles.

Método Probabilístico

Utiliza modelos de representación gráfica para comparar las diferentes representaciones del conocimiento. Estos modelos están basados en las probabilidades e independencias de los datos. Estos son útiles para aplicaciones que involucran incertidumbre y aplicaciones estructuradas tal que una probabilidad puede asignarse a cada uno de los ``resultados'' o pequeña cantidad del descubrimiento del conocimiento. Las técnicas probabilísticas pueden 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. El algoritmo de aprendizaje inductivo puede seleccionar automáticamente trayectorias útiles y atributos para construir las reglas de una base de datos con muchas relaciones'. Este tipo de inducción es usado para generalizar los modelos en los datos y construir las reglas de los modelos nombrados. El proceso analítico en línea (OLAP) es un ejemplo de un método orientado a la estadística.

Método de clasificación

Es el método más viejo y más usado de todos los métodos de KDD. Este método agrupa los datos de acuerdo a similitudes o clases. Hay muchos tipos de clasificación de técnicas y numerosas herramientas disponible que son automatizadas.

- **Método Bayesian:** es un modelo gráfico que usa directamente los arcos exclusivamente para formar un [sic] gráfica acíclica. Usa los medios probabilísticos y gráficos de representación, pero también es considerado un tipo de clasificación.

Redes de Bayesian: se usan cuando la incertidumbre se asocia con un resultado y puede expresarse en términos de una probabilidad. Este método cuenta con un dominio del conocimiento codificado y ha sido usado para los sistemas de diagnóstico.

- **Descubrimiento de patrones y de datos:** es otro tipo de clasificación que sistemáticamente reduce una base de datos grande a unos cuantos archivos informativos. Si el dato es redundante y poco interesante se elimina, la tarea de descubrir los patrones en los datos se simplifica. Este método trabaja en la premisa de un dicho viejo, ``menos es más''. El descubrimiento de patrones y las técnicas de limpia de datos son útiles para reducir volúmenes enormes de datos en las aplicaciones, tal como aquellos encontrados al analizar las grabaciones de un sensor automatizado. Una vez que las lecturas del sensor se reducen a un tamaño manejable usando la técnica de limpia de datos, pueden reconocerse con más facilidad los patrones de datos.
- **El método del árbol de decisión** usa las reglas de producción, construidas como figuras gráficas basado en datos premisos, y clasificación de los datos según sus atributos. Este método requiere clases de datos que son discretos y predefinidos. El uso primario de este método es para predecir modelos que pueden ser apropiados para cualquier clasificación o técnicas de regresión.

Método de desviación y tendencia del análisis

La base de este método es el *método de detección por filtrado*. Normalmente las técnicas de análisis y desviación son aplicadas temporalmente en las bases de datos. Una buena aplicación para este tipo de KDD es el análisis de tráfico en las grandes redes de telecomunicaciones.

AT&T usa tales sistemas para localizar e identificar circuitos que exhiben la desviación (conducta defectuosa). El volumen

total de datos que requieren análisis generan una técnica imperativa automatizada. Este tipo de tendencia de análisis también podría demostrar utilidad en los datos astronómicos y oceanográficos, ya que sus datos están basados en el tiempo y volumen.

Otros Métodos

- **Redes neuronales** son particularmente útiles para el reconocimiento de patrones y algunas veces se pueden agrupar con los métodos de clasificación.
- **Algoritmos genéticos** son usados para la clasificación, son similares a las redes neuronales aunque estas son consideradas más poderosos.

Método híbrido

También es llamado método multi-paradigmático. Combina la potencia de más de un método, aunque la implementación puede ser más difícil. Algunos de los métodos comúnmente usados combinan técnicas de visualización, inducción, redes neuronales y los sistemas basados en reglas para llevar a cabo el descubrimiento de conocimiento deseado. También se han usado bases de datos deductivas y algoritmos genéticos en los métodos híbridos.

Herramientas adicionales

La tendencia es proveer al usuario herramientas y facilidades para poder realizar KDD.

Desde este punto de vista, el proceso de KDD involucra interacciones complejas a través del tiempo entre un humano y una base de datos usando un conjunto de herramientas heterogéneas.

- Ayudas para analizar datos; entendimiento de la estructura, cobertura y calidad de los datos
- Herramientas para seleccionar herramientas, ajustarlas y refinar el modelo
- Visualización de datos y de patrones

- Integración de módulos (la salida de uno sirva de entrada en otro)
- Segmentación (selección) y discretización de datos
- Incorporación de conocimiento del dominio
- Interpretación de salidas
- Descubrimiento de la tarea a realizar
- Limpieza de datos (sin eliminar datos interesantes)
- Acoplamiento fuerte con bases de datos
- Desarrollo de algoritmos más eficientes, escalables y su paralelización

Minería de Datos: algoritmos y modelos

El proceso de minería involucra ajustar modelos o determinar patrones a partir de datos. Este ajuste normalmente es de tipo estadístico, en el sentido que se permite un cierto ruido o error dentro del modelo.

Los algoritmos de minería de datos realizan en general tareas de descripción (de datos y patrones), de predicción (de datos desconocidos) y de segmentación (de datos). Otras, como análisis de dependencias e identificación de anomalías se pueden utilizar tanto para descripción como para predicción.

- **Descripción:** se utiliza para el análisis preliminar de los datos (resumen, características de los datos, casos extremos, etc.). Con esto, el usuario se familiariza con los datos y sus estructuras. Busca derivar descripciones concisas de características de los datos (medias, desviaciones estándares, etc.).
- **Predicción :**
 - **Clasificación:** Los datos son objetos caracterizados por atributos que pertenecen a diferentes clases (etiquetas discretas).

- La meta es inducir un modelo para poder predecir una clase dados los valores de los atributos.
 - Se usan por ejemplo, árboles de decisión, reglas, análisis de discriminantes, etc.
- **Estimación** o **Regresión:** las clases son continuas.
 - La meta es inducir un modelo para poder predecir el valor de la clase dados los valores de los atributos.
 - Se usan por ejemplo, árboles de regresión, regresión lineal, redes neuronales, kNN, etc.
- **Segmentación:** separación de los datos en subgrupos o clases interesantes. Las clases pueden ser exhaustivas y mutuamente exclusivas o jerárquicas y con traslapes. Se puede utilizar con otras técnicas de minería de datos: considerar cada subgrupo de datos por separado, etiquetarlos y utilizar un algoritmo de clasificación. Se usan algoritmos de clustering, SOM (*self-organization maps*), EM (*expectation maximization*), k-means, etc. Normalmente el usuario tiene una buena capacidad de formar las clases y se han desarrollado herramientas visuales interactivas para ayudar al usuario.
- **Análisis de dependencias:** El valor de un elemento puede usarse para predecir el valor de otro. La dependencia puede ser probabilística, puede definir una red de dependencias o puede ser funcional (leyes físicas). También se ha enfocado a encontrar si existe una alta proporción de valores de algunos atributos que ocurren con cierta medida de confianza junto con valores de otros atributos. Se pueden utilizar redes Bayesianas, redes causales, y reglas de asociación.

- **Detección de desviaciones, casos extremos o anomalías:** Detectar los cambios más significativos en los datos con respecto a valores pasados o normales. Sirve para filtrar grandes volúmenes de datos que son menos probables de ser interesantes. El problema está en determinar cuándo una desviación es significativa para ser de interés.

La aplicación de los algoritmos de minería de datos requiere la realización de una serie de actividades previas encaminadas a preparar los datos de entrada debido a que, en muchas ocasiones dichos datos proceden de fuentes heterogéneas, no tienen el formato adecuado o contienen ruido. Por otra parte, es necesario interpretar y evaluar los resultados obtenidos.

El proceso completo consta de las siguientes etapas [Cabena et al., 1998]:

1. Determinación de los objetivos

2. Preparación de los datos

3. Selección: Identificación de las fuentes de información externas e internas y selección del subconjunto de datos necesario.

4. Preprocesamiento: estudio de la calidad de los datos y determinación de las operaciones de minería que se pueden realizar.

5. Transformación de datos: conversión de datos en un modelo analítico.

6. Minería de datos: tratamiento automatizado de los datos seleccionados con una combinación apropiada de algoritmos.

7. Análisis de resultados: interpretación de los resultados obtenidos en la etapa anterior, generalmente con la ayuda de una técnica de visualización.

8. Asimilación del conocimiento: aplicación del conocimiento descubierto.

Aunque los pasos anteriores se realizan en el orden en que aparecen, el proceso es altamente iterativo, estableciéndose

retroalimentación entre los mismos. Además, no todos los pasos requieren el mismo esfuerzo, generalmente la etapa de preprocesamiento es la más costosa ya que representa aproximadamente el 60 % del esfuerzo total, mientras que la etapa de minería sólo representa el 10%.

Componentes básicas de los modelos de minería de datos

Lenguaje de representación del modelo: es muy importante que se sepan las suposiciones y restricciones en la representación empleada para construir modelos.

Evaluación del modelo: En cuanto a predictividad se basa en técnicas de validación cruzada (*cross validation*) en cuanto a calidad descriptiva del modelo se basan en principios como el de máxima verosimilitud (*maximum likelihood*) o en el principio de longitud de descripción mínima o MDL (*minimum description length*).

Método de búsqueda: se puede dividir en búsqueda de parámetros y búsqueda del modelo, y determinan los criterios que se siguen para encontrar los modelos (hipótesis).

Modelos de Minería de Datos

- **Predictivos o Basados en la Memoria**

Técnicas: Clasificación, Predicción de valores.

Ejemplos: ¿Cuál es el riesgo de este cliente?, ¿Se quedará el cliente?

Los modelos predictivos requieren de un set de pruebas y de interacciones de entrenamiento:

1. Selección de pruebas.
2. Minado inicial.
3. Resultado.
4. Aplicación de una segunda muestra representativa.
5. Análisis de los resultados .

6. Interacciones hasta lograr un modelo consistente.

7. Aplicar al negocio.

- **Descriptivos**

Técnicas : Asociación, Segmentación o 'Clustering'

Ejemplos: Un cliente que compra productos dietéticos es tres veces más probable que compre caramelos.

Recomendaciones para implementar la tecnología de Minería de Datos

- **Primer Paso**

- Construir la Infraestructura (Data Warehouse).

- Implementar el proceso y crear la cultura

- **Segundo Paso**

- Incorporar Minería de datos al proceso.

- Seleccionar la herramienta.

- Identificar áreas de aplicación o problemas de negocio en dónde la minería de datos puede ayudarnos.

Técnicas más comúnmente usadas en Minería de Datos

- **Redes neuronales artificiales:** modelos predecible no-lineales que aprenden a través del entrenamiento y semejan la estructura de una red neuronal biológica.

- **Árboles de decisión:** estructuras de forma de árbol que representan conjuntos de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos.
Se realizan cortes sobre una variable, lo cual limita su expresividad, pero facilita su comprensión. Generalmente se usan técnicas heurísticas en su construcción. Los métodos específicos de árboles de decisión incluyen Árboles de Clasificación y Regresión (CART: Classification And Regression Tree) y Detección de Interacción Automática de Chi Cuadrado (CHAI: Chi Square Automatic Interaction Detection)
- **Algoritmos genéticos:** técnicas de optimización que usan procesos tales como combinaciones genéticas, mutaciones y selección natural en un diseño basado en los conceptos de evolución. Inspirados en el principio de la supervivencia de los más aptos. La recombinación de soluciones buenas en promedio produce mejores soluciones. Es una analogía con la evolución natural.
 - **Programación Genética:** se basan en la evolución de programas de cómputos que permitan explicar o predecir con mínimo error un determinado fenómeno.
- **Método del vecino más cercano:** una técnica que clasifica cada registro en un conjunto de datos basado en una combinación de las clases del/de los k registro (s) más similar/es a él en un conjunto de datos históricos (donde $k \geq 1$). Algunas veces se llama la técnica del vecino k-más cercano.
- **Regla de inducción:** la extracción de reglas if-then de datos basados en significado estadístico.
- **Modelos gráficos de dependencias probabilísticas:** básicamente redes bayesianas, en donde la evaluación se basa en probabilidad y el encontrar el modelo en heurísticas.
- **Clustering:** agrupan datos cuya *distancia* multidimensional intraclass es *pequeña* e interclass es *grande*. (incluye clasificadores difusos). Agrupa los datos basándose en las similitudes de los mismos.
Ej: descripción de cada uno de los consumidores. En este caso se agruparía consumidores con características similares y al mismo tiempo se maximizarían las diferencias entre los distintos

grupos de consumidores. Existen diferentes técnicas de clustering y cada una de las mismas tiene sus propias aproximaciones para descubrir las aproximaciones que existen entre sus datos.

- **Análisis de Enlace (Link analysis):** describe una familia de técnicas que determinan asociaciones entre los registros de datos. El tipo de análisis de enlace más conocido es el Análisis de la canasta de mercado; en este caso los registros son los items comprados por un cliente durante la misma transacción y debido a que la técnica fue derivada del análisis de los datos de un supermercado, se considera que estos se encuentran en la misma canasta al momento de la compra o transacción. El análisis de la Canasta de Mercados descubre la combinación de items que fueron comprados por diferentes consumidores, y por asociación o enlace se puede determinar que tipos de productos son comprados juntos. El análisis de enlace no se restringe solo al análisis de la canasta de mercado, teniendo en cuenta que la canasta es un grupo de registros de datos la técnica puede ser usada en cualquier situación donde haya un número grande de grupos de registros de datos.
- **Análisis de Frecuencia (Frequency análisis):** comprende aquellas técnicas de minería de datos que son aplicadas al análisis de registros ordenados en el tiempo o cualquier conjunto de datos que puedan ser ordenado en el tiempo. Estas técnicas de minería de datos intenta detectar secuencias o subsecuencias similares en los datos ordenados.

Muchas de estas tecnologías han estado en uso por más de una década en herramientas de análisis especializadas que trabajan con volúmenes de datos relativamente pequeños. Estas capacidades están ahora evolucionando para integrarse directamente con herramientas OLAP y de Data Warehousing.

Bibliografía

<http://dns1.mor.itesm.mx/~emorales/Cursos/KDD01>

http://www.lania.mx/spanish/actividades/newsletters/1999-otono-invierno/retos_mineria.html

<http://dns1.mor.itesm.mx/~emorales/Cursos/KDD01/principal.html>

➤ Documentos en PDF:

Aplicación de técnicas de minería de datos en la Construcción y validación de modelos predictivos y Asociativos a partir de especificaciones de requisitos De software. *María N. Moreno García, Luis A. Miguel Quintales, Francisco J. García Peñalvo y M.José Polo Martín-Universidad de Salamanca-*

Bodegas de Datos como Apoyo a la Toma de Decisiones. *Dr. José Torres Jiménez*

Minería de Datos e Minería de Datos e Inteligencia de Negocios Inteligencia de Negocios. *Francisco J. Cantú-Centro de Sistemas Inteligentes-*

Minería y Almacenes de Datos.
[http://usuarios.lycos.es/sachavir.](http://usuarios.lycos.es/sachavir)

Minería de datos- Conceptos y Objetivos. www.daedalus.es

Mining Your Own Business in Retail Using DB2 Intelligent Miner for Data . ibm.com/redbooks

Minería de Datos-*José Hernández Orallo.*