1º CONTECSI Congreso Internacional de Gestión de la Tecnologia y Sistemas de Información
21-23 de Junio de 2004 USP/São Paulo/SP-Brasil

MINERÍA DE DATOS

MICHELL ANGELO FERRUCCIO[1]
ARTURO IVÁN GARCÍA ALONSO
SANDRA XIMENA GÓMEZ

ABSTRACT

What does data mining really mean? There is a lot of confusion among people looking to extract information from the databases. Usually they assume that data analysis is the same thing as data mining. In the traditional query approach, the analyst generates a series of questions based on his domain knowledge, perhaps guided by an hypothesis to be tested. The answers to these questions are used to deduce a pattern or verify the hypothesis about the data.

Data mining uses a variety of data analysis tools to discover patterns and relationships in data that can be used to make reasonably accurate predictions. It is a process, not a particular technique or algorithm. We would like to emphasize that the goal of data mining is prediction, generalizing a pattern to other data. Exploring and describing the database is merely the starting point.

RESUMEN

¿Que significa realmente Minería de datos? Existe una gran confusión entre aquellos que tratan de extraer información de las bases de datos; usualmente asumen que el análisis de datos es lo mismo que la minería de datos.
En el acercamiento tradicional por consultas, el analista genera una serie de preguntas basadas en su dominio de conocimiento, quizás guiado por una hipótesis que debe ser comprobada.

La minería de datos utiliza una variedad de herramientas de análisis de datos para descubrir patrones y relaciones en los datos, que pueden ser utilizados para efectuar predicciones razonablemente acertadas. Es un proceso, no una técnica particular o algoritmo, cuyo objetivo es la predicción a partir de la generalización de un patrón para ciertos datos. La exploración y descripción de la base de datos, es meramente el punto de partida.

INTRODUCCIÓN

La tecnología actual permite capturar y almacenar una gran cantidad de datos, sin embargo, estos datos por si solos no son suficientes para satisfacer las necesidades de información de sus propietarios. Tratar de buscar relaciones entre los datos y descifrar patrones o tendencias son retos de la vida moderna.

Una de las formas más sencillas para buscar relaciones es utilizar las habilidades humanas; cualquier persona sería capaz de identificarlas observando con detenimiento tablas con un cierto número de datos y posiblemente haciendo uso de un lápiz y papel.

No obstante las dimensiones de las bases de datos actuales hacen muy difícil para un humano, realizar algún tipo de análisis y extraer información importante.

En la actualidad existen herramientas que trabajan sobre una gran cantidad de datos y que proporcionan análisis estadístico de los mismos, que si bien es cierto, son de gran utilidad en algunos casos, por ejemplo, en la determinación de la época en donde se realizan más ventas en un almacén, no proporcionan toda la información posible.

[1] Estudiantes de Ingeniería de Sistemas. Universidad Nacional de Colombia. Sede Bogotá

El descubrimiento de conocimiento en base de datos (KDD) es una técnica que combina recursos desarrollados en el área de la inteligencia artificial para extraer información importante de las bases de datos, y dentro del cual se referencia a la minería de datos como un paso fundamental dentro del proceso. Asimismo, la minería de datos es fundamental en la investigación científica, y técnica, como herramienta de análisis y descubrimiento de conocimiento a partir de datos de observación o de resultados de experimentos.

LA MINERÍA DE DATOS

La minería de datos pretende encontrar información, que se pueda extraer de las bases de datos en un proceso de selección y aplicación de algoritmos de búsqueda de patrones, relaciones, reglas, asociaciones e incluso excepciones que sean útiles para la toma de decisiones.

ARQUITECTURA DE UN SISTEMA TIPICO DE MINERIA DE DATOS

Los componentes que constituyen un sistema típico de minería de datos, comprenden un conjunto de bases de datos, almacenes y depósitos de información y/o hojas de cálculo, sobre los que se ejecutan técnicas de limpieza e integración.

Otros componentes son:

Servidor de almacén de datos

Es responsable de buscar los datos relevantes, basados en las demandas del usuario de la minería de datos.

Base de conocimiento

Este es el dominio del conocimiento que se utiliza para guiar la búsqueda o evaluar la importancia de los patrones resultantes.

Motor de minería de datos

Es esencial e idealmente consiste en un grupo de módulos funcionales de tareas como la caracterización, asociación, clasificación, análisis de cluster, y análisis de evolución y desviación.

Modulo de evaluación de patrones

Emplea las medidas de interés e interactúa con los módulos como foco de la búsqueda hacia los patrones relevantes.

GUI

Módulo que comunica los usuarios con el sistema, permitiendo especificar la consulta o tarea a ejecutar, y proporcionado información que ayude a enfocar la búsqueda.

TIPOS DE ARQUITECTURA

La pregunta clave que se realiza es, si se debe integrar o combinar los sistemas de minería de datos con los sistemas de bases de datos o almacenes de datos.
Por lo tanto surgen los esquemas: *Sin enganche, Enganche débil, Enganche semi-fuerte, Enganche fuerte.*

Sin enganche

Significa que el sistema de minería de datos (DM) no utilizará ninguna función de un sistema de bases de datos o de almacén de datos. Buscará por algún origen particular, como por ejemplo un archivo del sistema, procesará los datos utilizando algunos algoritmos de minería de datos, y luego almacenará los resultados en otro archivo.

Enganche débil

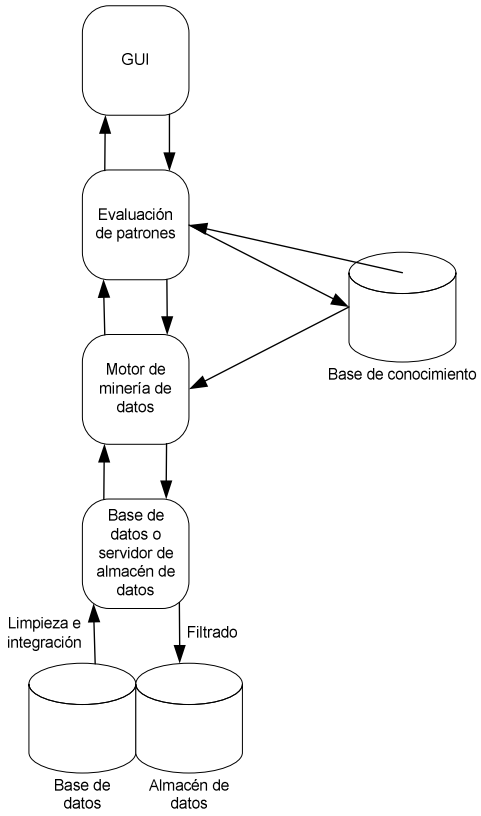
Trae datos desde un depósito administrado por un sistema de bases de datos (DB) o por un almacén de datos (DW), y luego almacena los resultados en un lugar establecido, que puede ser un archivo o en un sistema DB/DW.

Enganche semi-fuerte

Toma algunas primitivas de la minería de datos de los sistemas DB/DW, como ordenamiento, indización, agregación, y cómputos preestablecidos.

Enganche fuerte

El sistema de minería de datos esta integrado en el sistema DB/DW. Es decir, que el sistema de DM es un componente funcional de un sistema de información.



Arquitectura de un sistema típico de minería de datos.

ESTRUCTURA DE LA MINERÍA DE DATOS

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

Las tareas de la minería de datos se pueden clasificar en dos categorías: minería de datos descriptiva y minería de datos predictiva. Junto a éstas existen otras tareas complementarias como la segmentación de datos, el análisis de dependencias y la identificación de anomalías; las cuales se pueden utilizar tanto en descripción como en predicción.

La descripción es normalmente usada para realizar un análisis preliminar de los datos. Busca derivar descripciones concisas de características de los datos: medias, desviaciones estándares, etc.

En *la predicción* los datos son objetos caracterizados por atributos que pertenecen a diferentes clases. La meta es inducir un modelo para poder predecir una clase dados los valores de los atributos (conocimiento inductivo).

Para ello, se usan por ejemplo, árboles de decisión, reglas, redes neuronales etc.

La segmentación consiste en separar los datos en subgrupos o clases que puedan ser particionados en una forma uniforme, y que constituyan intervalos que parezcan intuitivos o naturales.

En *el análisis de dependencias* el valor de un elemento puede usarse para predecir el valor de otro. 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.

La detección de desviaciones, casos extremos o anomalías busca 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.

Los componentes básicos de los métodos o técnicas de minería son:

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

2. *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*).

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

REPRESENTACION DEL CONOCIMIENTO E INFERENCIAS PRODUCIDAS

Con base en las tareas definidas en la estructura de la minería de datos se han identificado una serie de técnicas:

- Herramientas de consulta
- Visualización
- Arboles de decisión
- Reglas de asociación
- Redes neuronales
- Algoritmos genéticos
- Redes Bayesianas

En éstas técnicas la representación del conocimiento se puede encontrar en forma de tablas, árboles o reglas.

Como se verá más adelante, estás técnicas pueden ser muy buenas en algunas tareas mientras que en otras son deficientes; por esta razón, generalmente se realiza una combinación de algunas de ellas.

Herramientas de consulta

Un lenguaje de consulta necesita una sintaxis común, que permita a los usuarios especificar la forma en que se visualizan los patrones descubiertos en una o varias formas, incluyendo reglas, tablas, tablas cruzadas, diagramas de barras, árboles de decisión, cubos, etc. Por tanto se definió el lenguaje DMQL para dicho propósito.

Existe otro lenguaje, llamado MSQL, que utiliza la sintaxis de SQL y ciertas primitivas como ordenamiento y agrupamiento. Como se pueden generar grandes cantidades de reglas, este lenguaje proporciona primitivas –como *GetRule* y *SelectRule*- para generación y selección de reglas.

Otros esfuerzos en diseño de lenguajes para minería de datos, incluyen *MINE RULE operator* que sigue la sintaxis de SQL y sirve como consulta de generación de reglas para asociación.

Recientemente, Microsoft™ propuso un lenguaje llamado OLE DB for Data Mining (DM). Es un esfuerzo notable hacia la estandarización de las primitivas del lenguaje de minería de datos. Teniendo un lenguaje estándar se ayudará a fortalecer la industria de minería de datos facilitando el desarrollo de plataformas y sistemas de DM, y el tomar parte de los resultados de DM.

Técnicas de Visualización

Las técnicas de visualización consisten en mostrar en espacios de dos o más dimensiones información sobre dos o más atributos. Son útiles en las tareas de segmentación, es decir, permiten realizar de manera rápida la identificación de subconjuntos que pueden ser de interés, por lo que son usadas generalmente al principio del proceso de minería de datos.

Árboles de decisión

Los árboles de decisión son usados en tareas predictivas de clasificación.

Para el caso de la minería de datos, en un árbol de decisión se toma un atributo que pueda ser de importancia y se divide en dos por un cierto valor umbral. Esto se hace una y otra vez hasta que se obtiene una respuesta adecuada.

Lo anterior se puede ilustrar mejor con un ejemplo. Supóngase que se tiene un producto X y se desea saber dentro de un grupo de personas con un rango de ingresos y de edades, cuáles son las que tienden a comprar más dicho producto. El árbol de decisión se muestra en la figura 1.
En el ejemplo se puede apreciar que no solo se trata con un atributo, la edad, sino que a partir del segundo nivel también es importante el ingreso en uno de los subgrupos.

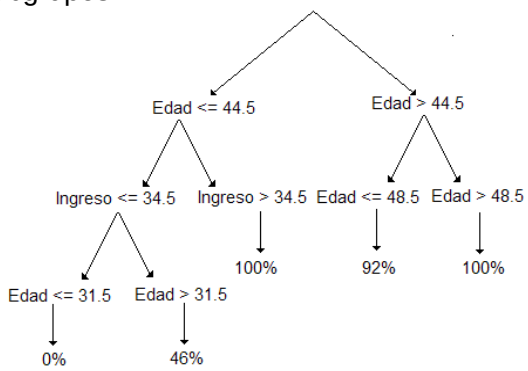


Fig. 1 Arbol de decisión para el producto X

La información que brindan los árboles de decisión es de gran utilidad debido a que permite identificar claramente los subgrupos. Además su estructura de árbol proporciona mayor entendimiento y permite al usuario ver el proceso de decisión en su totalidad desde que se realiza la primera división.

Reglas de Asociación

Las reglas de Asociación relacionan un conjunto de pares atributo-valor con otros

pares atributo-valor. Una regla de asociación es de la forma:

Sean X e Y atributos e I un conjunto de registros

$$X \Rightarrow Y, \text{con } X \subset I, Y \subset I, X \cap Y = 0$$

- Se cumple en *la BD* con confianza *c* si *c*% de un subconjunto de registros que pertenecen a la BD que contienen *X* también contienen *Y*
- Soporte de la regla = soporte de $X \cup Y$ (tiene soporte *s* en la BD si *s*% de las transacciones en la BD contienen $X \cup Y$)

Las reglas de asociación tienen carácter predictivo y son de gran utilidad en muchas disciplinas, entre las que se encuentran la medicina, la mercadotecnia y las finanzas. Esta utilidad dependerá de la escogencia de un umbral de confianza y un umbral de soporte adecuados.
Algunos ejemplos de reglas de asociación son:

- Si cliente compra A y B entonces compra C el 80% de las veces.

- El 90% de los que ven las páginas A y B ven *después* la página C (Reglas de asociación secuencial).

Redes Neuronales

Las redes neuronales están compuestas por elementos simples operando en paralelo. Estos elementos están inspirados por el sistema nervioso biológico. La función de la red es determinada por las conexiones entre elementos.

Una red neuronal se puede entrenar para mejorar una función particular ajustando los valores de las conexiones entre los elementos (entrenamiento de la red). La red es ajustada comparando la salida obtenida con el resultado esperado (Véase Figura 2).

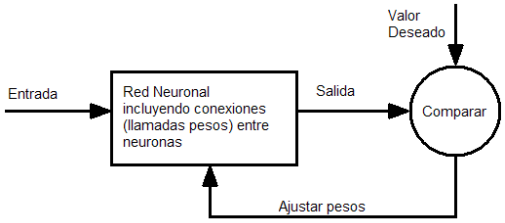


Fig 2. Entrenamiento de una red neuronal

La arquitectura de una red neuronal incluye unos nodos de entrada y unos nodos de salida, los cuales se conectan por medio de unos nodos ocultos o capas internas que se encuentran entre las capas exteriores y que actúan como una caja negra (Véase Figura 3).

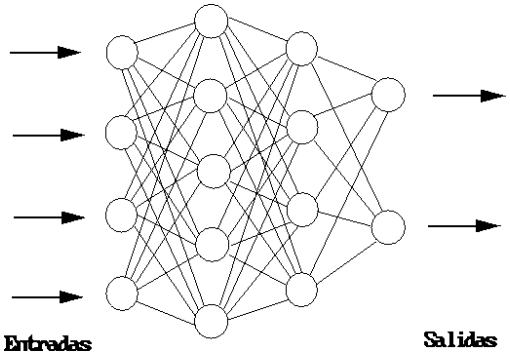


Fig 3. Arquitectura de una red neuronal

Las redes neuronales han sido entrenadas para realizar funciones complejas en distintos campos de aplicación incluyendo reconocimiento de patrones y sistemas de control.

Algoritmos Genéticos

Introducidos por John Holland en 1970, los algoritmos genéticos establecen una analogía entre el conjunto de soluciones de un problema y el conjunto de individuos de una población natural, codificando la información de cada solución en un vector a modo de cromosoma.

La idea de éstos algoritmos es la de encontrar soluciones aproximadas a

problemas de gran complejidad computacional mediante un proceso de evolución simulada. Para esto se introduce una función de evaluación de los cromosomas (calidad o “fitness”), que está basada en la función objetivo del problema. Igualmente se introduce un mecanismo de selección de manera que los cromosomas con mejor evaluación sean escogidos para "reproducirse" más a menudo que aquellos que tienen la peor.

Para llevar a la práctica el esquema anterior y concretarlo en un algoritmo, hay que especificar los siguientes elementos:

- Una representación cromosómica
- Una población inicial
- Una medida de evaluación
- Un criterio de selección / eliminación de cromosomas
- Una o varias operaciones de recombinación
- Una o varias operaciones de mutación

La representación cromosómica puede realizarse con cadenas binarias (unos y ceros), o con un número limitado de términos de un alfabeto.

La población inicial suele ser generada aleatoriamente. De no ser aleatoria, se sugiere que esta población inicial sea lo suficientemente variada para que represente una gran parte de la población y se evite la convergencia prematura.

Respecto a la evaluación de los cromosomas, se suele utilizar la calidad como medida de la bondad según el valor de la función objetivo en el que se puede añadir un factor de penalización para controlar la infactibilidad.

Los Operadores de Cruzamiento mas utilizados son:

- De un punto: Se elige aleatoriamente un punto de ruptura en los padres y se intercambian sus bits.
- De dos puntos: Se eligen dos puntos de ruptura al azar para intercambiar.
- Uniforme: En cada bit se elige al azar un padre para que contribuya con su bit al del hijo, mientras que el segundo hijo recibe el bit del otro padre.

La operación de mutación más sencilla, y una de las más utilizadas consiste en reemplazar con cierta probabilidad el valor de un bit. Otra forma es recombinar elementos tomados al azar sin considerar su fitness.

Redes Bayesianas

Una red bayesiana es un grafo acíclico dirigido en el que cada nodo representa una variable y cada arco una dependencia probabilística, en la cual se especifica la probabilidad condicional de cada variable dados sus padres. La variable a la que apunta el arco es dependiente (causa-efecto) de la que está en su origen.

La topología o estructura de la red brinda información sobre las dependencias probabilísticas entre las variables y sus dependencias condicionales dada otra(s) variable(s).

Entre las características que poseen las redes bayesianas, se puede destacar que permiten aprender sobre relaciones de dependencia y causalidad, permiten combinar conocimiento con datos, evitan el sobre-ajuste de los datos y pueden manejar bases de datos incompletas.

Rendimiento de las técnicas

Las técnicas para obtención y representación del conocimiento mencionadas anteriormente no tienen el mismo rendimiento en todas las tareas. Algunas trabajan mejor en unas más que en otras. Los siguientes son los usos de

algunas de éstas técnicas en cada una de las tareas:

- Las reglas de asociación son buenas en tareas de descripción y predicción, y análisis de dependencias
- Las redes neuronales trabajan bien en tareas predictivas, al igual que los algoritmos genéticos.
- Las herramientas de consulta son usadas para las tareas de segmentación y descripción.
- Las redes bayesianas funcionan en el análisis de dependencias.
- Los árboles de decisión se comportan de manera satisfactoria tanto en tareas de descripción como en tareas de predicción.

Es importante resaltar que éstas técnicas no son las únicas que existen, pero si hacen parte de las más utilizadas.

TIPOS DE RAZONAMIENTO

La representación del conocimiento en la minería de datos puede realizarse de diversas formas: redes bayesianas, redes neuronales, árboles de decisión, etc, razón por la cual también utiliza diferentes tipos de razonamiento, dentro del proceso de identificación de patrones.

El razonamiento lógico suele aplicarse cuando se emplean, por ejemplo, árboles de decisión para representar el conocimiento, y puede realizarse hacia adelante, a partir de los estados iniciales o hacia atrás, a partir de los estados objetivo; la diferencia entre ellos radica en la forma de construir el árbol.

En algunas ocasiones, sin embargo, puede resultar adecuado describir las creencias sobre las que no se tiene certeza, en las que existen algunas evidencias que las apoyan. En este caso el razonamiento estadístico es de gran utilidad; mientras en los sistemas lógicos solo es necesario encontrar una demostración para poder asegurar el valor cierto de una proposición; en los sistemas que son inciertos, como por ejemplo los representados en redes bayesianas, el cálculo de una creencia en una proposición necesita que se combinen todos los tipos de razonamiento posibles.

Por otro lado, también se suele utilizar el razonamiento en la mayoría de las representaciones, tanto para construir el modelo como para realizar predicciones sobre una gran cantidad de datos a partir de operaciones matemáticas, por ejemplo al comparar totales de diferentes atributos.

OPERADORES EMPLEADOS

Con base en las diferentes formas de representar el conocimiento y los tipos de razonamiento que se han expuesto en el artículo, se puede afirmar que básicamente la minería de datos utiliza operadores lógicos y matemáticos en su proceso de inferencia y construcción de patrones.

En la mayoría de los modelos se suelen emplear operaciones matemáticas, sobre todo cuando se realizan predicciones a través de cálculos estadísticos. De la misma forma se realizan operaciones de tipo lógico cuando se realizan comparaciones y se toman decisiones.

RELACION PERTINENTE DEL ENTORNO

El uso de la minería de datos está determinado por una serie de criterios específicos prácticos y técnicos:

Criterios prácticos:

- Existe potencialmente un impacto significativo.
- No hay métodos alternativos.
- Existe soporte del cliente para su desarrollo.
- No existen problemas de legalidad o violación a información privilegiada.

Criterios técnicos:

- Existen suficientes datos.
- Atributos relevantes.
- Poco ruido (datos incompletos o erróneos) en los datos.
- Conocimiento del dominio

Son muchas las aplicaciones que tiene la Minería de datos, entre las que se incluyen principalmente:

- Análisis biomédico y DNA:

Integración semántica de DB distribuidas de genomas.
Búsqueda de similitudes entre secuencias de DNA.
Identificación de ocurrencias de secuencias.
Conexión de genes a diferentes etapas del desarrollo de enfermedades.
Análisis genético.

- Análisis financiero:

Diseño y construcción de almacenes de datos para análisis de datos multidimensionales.
Predicción de pagos de préstamos y análisis de políticas de créditos para clientes.
Clasificación de clientes para mercadeo.
Detección de lavado de dinero y crímenes financieros.

- Industria minorista:

Diseño y construcción de almacenes de datos, basadas en los beneficios de la Minería de Datos.
Análisis multidimensional de ventas, clientes, productos, tiempo y región.
Análisis de efectividad de campañas de ventas.
Retención de clientes (análisis de lealtad de clientes).
Recomendaciones de compra.

- Industria de Telecomunicaciones:

Análisis multidimensional de datos de telecomunicaciones.
Análisis de patrones fraudulentos e identificaciones de patrones inusuales.
Asociación multidimensional y análisis de patrones secuenciales.
Utilización de herramientas de visualización en análisis de datos de telecomunicaciones.

Algunos casos específicos son:

- En el campo de las empresas de telecomunicaciones, es famoso el caso de la detección de fraudes en cuentas telefónicas llevado a cabo por la British Telecom que, gracias a la Minería de Datos pudo establecer que una clase de fraudes muy importante y repetitiva se estaba produciendo exclusivamente en una región muy limitada de Inglaterra. El estudio de la enorme cantidad de cuentas (y llamados telefónicos), con otras herramientas nunca hubiese sido posible.

- En el ámbito policial: Interpol y las organizaciones policiales asociadas, en conjunto con los grandes bancos internacionales, ha podido detectar y desmantelar importantes redes de lavado de dinero proveniente del narcotráfico, analizando las transferencias de montos algo por debajo de los límites a partir de los cuales los diferentes países exigen documentación justificada.

- En el mundo de la medicina: Un campo bastante desconocido es el de la interacción de los principios activos de diversos medicamentos consumidos conjuntamente. La Minería de Datos aplicada a grandes cantidades de fichas clínicas de pacientes sometidos a medicación múltiple, ha permitido descubrir efectos secundarios

indeseables, información de vital importancia para los médicos tratantes.

En cuanto al entorno específico del campus universitario se podrían mencionar algunas posibles aplicaciones:

- Identificación del rendimiento académico debido a factores económicos.

- La influencia del desarrollo social y familiar del individuo en el desempeño académico.

- La tendencia de seleccionar electivas de tipo administrativas en determinado estrato social.

CONCLUSIONES

El análisis de datos almacenados en una base de datos tiene un carácter altamente exploratorio. El usuario está en busca de nueva información, de nuevos patrones que le sugieran relaciones entre diferentes aspectos de su actividad cotidiana. Si el usuario tuviera el conocimiento de todas esas asociaciones, no necesitaría el análisis de los datos.

La generación de información cada vez mayor y su almacenamiento, sin herramientas adecuadas de análisis, está propiciando pérdidas de información valiosa para la toma de decisiones, por lo que los sistemas para descubrir conocimiento en bases de datos son bastante útiles en diferentes dominios en donde no es fácil formalizar el conocimiento.

Las herramientas actuales aún requieren de una alta participación de un usuario humano, pues son interactivas y requieren la guía del experto. Sin embargo, se espera que en el futuro la identificación de patrones sea mucho más automatizada, simplemente porque los volúmenes de información por analizar crecen mucho más que los recursos humanos para analizarlos.

No obstante, se deben tener ciertas consideraciones para el uso de la minería de datos ya que descubrir conocimiento en bases de datos no es, ni será la solución total para resolver problemas de una organización altamente compleja, ni las herramientas actuales son totalmente adecuadas, debido a que los patrones encontrados no siempre cumplen con las expectativas del usuario.

BIBLIOGRAFÍA

ADRIAANS Pieter, Zantige Dolf. Data Mining. Addison Wesley. 1996.

HAN, Jiawei y KAMBER, Micheline. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers. 2001.

RICH Elaine, Knight Kevin. Inteligencia Artificial. McGraw Hill.1994.