Ingeniería del Software

DESCUBRA LA INFORMACION OCULTA DE LAS BASES DE DATOS BORROSAS

RESUMEN

El volumen de datos manejados por la sociedad actual crece a un ritmo acelerado, generando inconvenientes, debido a que el conocimiento más valioso suele aparecer oculto entre los datos recogidos, en forma de patrones o reglas que relacionan entre sí otras partes más superficiales de la información. Este conocimiento se ha venido obteniendo, tradicionalmente, mediante análisis manual, aplicando la inferencia inductiva sobre el conjunto de datos de partida; pero ante tal avalancha de información, es imposible pensar en dichos métodos para el análisis; ante esta problemática surge la *Minería de Datos* como una nueva técnica de análisis automático que busca extraer la información sin "ruido", pero con un alto grado de validez y certeza.

Palabras Claves:

- KDD
- Minería de Datos
- Data Mining
- Bases de Datos Borrosas
- Lógica Borrosa

Introducción:

El volumen de datos generados por la sociedad actual crece a un ritmo acelerado, este fenómeno se refleja, por ejemplo, en las observaciones satelitales de la Tierra, que fácilmente manipulan un volumen de datos del orden de un petabyte (10^{15} bytes), o en Internet, donde puede encontrarse cualquier cantidad de información que se pueda imaginar; algunos cálculos de estimación sugieren que esta cantidad se multiplica por 10 cada año. En situaciones mas cotidianas y con un volumen de datos menor, se encuentran entre otras, las operaciones bancarias con tarjetas de crédito o las transacciones realizadas en los supermercados.

Lo más preocupante, es que el conocimiento más valioso suele aparecer oculto entre los datos recogidos, en forma de patrones o reglas que relacionan entre sí otras partes más superficiales de la información. Este conocimiento se ha venido obteniendo, tradicionalmente, mediante análisis manual, aplicando la inferencia inductiva sobre el conjunto de datos de partida. Pero ante esta avalancha de información es imposible pensar en dichos métodos para el análisis, necesitando cada

vez más la ayuda de super-computadores que automaticen el proceso inductivo, para así analizar de forma inteligente las montañas de datos existentes y extraer de ellas ese conocimiento oculto y valioso.

La "Minería de datos" , como se conocen en la actualidad a las nuevas técnicas de análisis automático, busca extraer la información sin "ruido", pero con un alto grado de validez y certeza.

Artículo:

La Incertidumbre que se produce como consecuencia de la imprecisión y subjetividad propias de la actividad humana, está acompañada, de forma inherente por la información.

Muchos de los conocimientos que manejamos son vagos y borrosos, es decir, sus límites no están perfectamente determinados, y no por ello carecen de significación. La lógica borrosa y la teoría de conjuntos borrosos ofrecen un método natural para representar esa imprecisión y subjetividad humana, generada también por la forman como se ha adquirido el conocimiento a través del método hipótesis-deducción, donde la intuición del investigador pautaba la hipótesis. Este procedimiento es obsoleto en la medida que el volumen de datos de información o datos muestra va aumentando, llevando a que todas las investigaciones de "inducción" (intuición) de conocimiento, busquen automatizar este proceso.

La estadística ha aportado diversas técnicas para la extracción de información útil; sin embargo, estas técnicas no permiten la creación de reglas o leyes de conocimiento; para esto, son necesarias *técnicas de análisis inteligente*. En consecuencia, cada vez más investigaciones dentro de la Inteligencia Artificial, se dirigen a la inducción de conocimiento en bases de datos. Debido a esta creciente necesidad ha aparecido un nuevo campo: *La Minería de datos* (Data Mining), la cual aplica nuevos métodos matemáticos y técnicas software para el análisis inteligente de datos.

Así, surge un proceso que además de incluir el análisis inteligente de los datos con técnicas de data mining, incluye los pasos previos, como el filtrado y preprocesado de los datos; y los pasos posteriores, como la interpretación y validación del conocimiento extraído. Este proceso se denomina descubrimiento de conocimiento en bases de datos (KDD: knowledge discovery in databases).

Según [Frawley et al., 91] *El descubrimiento de conocimiento es la extracción no trivial de información implícita, previamente desconocida y potencialmente útil, a partir de un conjunto de datos. Dado un conjunto de hechos (datos) H, un lenguaje L, y alguna medida de la certidumbre C, definimos una regularidad (pattern) como una*

sentencia S en L que describe relaciones dentro de un subconjunto H s de H con una certidumbre c, de forma que S es más sencillo que la enumeración de todos los hechos de H s . Una regularidad que sea interesante y bastante cierta (según criterios definidos por el usuario) se denomina conocimiento. Un sistema de descubrimiento será un programa que toma como entrada el conjunto de hechos y extrae las regularidades existentes.

KDD

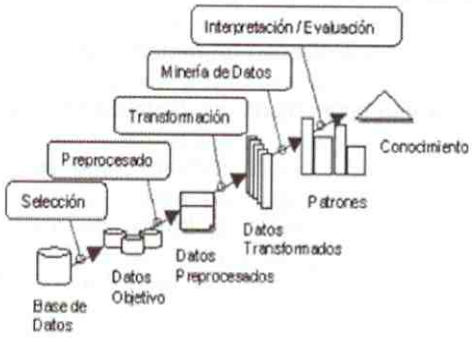
Cuando el conocimiento se extrae partiendo de los datos de una base de datos, se tiene KDD.

Los principales pasos o procesos dentro de KDD son los siguientes:

1. Desarrollo y entendimiento del dominio de la aplicación, el conocimiento relevante y los objetivos del usuario final.
2. Creación del conjunto de datos objetivo, seleccionando el subconjunto de variables o ejemplos sobre los que se realizará el descubrimiento.

Data Mining

3. Preprocesado de los datos: eliminación de ruido, estrategias para manejar valores ausentes, normalización de los datos, etc.
4. Transformación y reducción de los datos: Incluye la búsqueda de características útiles de los datos según sea el objetivo final, la reducción del número de variables y la proyección de los datos sobre espacios de búsqueda en los que sea más fácil encontrar una solución.
5. Elección del tipo de sistema para minería de datos. Esto depende de si el objetivo del proceso de KDD es la clasificación, regresión, agrupamiento de conceptos (clustering), detección de desviaciones, etc.
6. Elección del algoritmo de minería de datos.
7. Minería de datos. En este paso se realiza la búsqueda de conocimiento con una determinada representación del mismo. El éxito de la minería de datos depende en gran parte de la correcta realización de los pasos previos, por parte del usuario.
8. Interpretación del conocimiento extraído, con posibilidad de iterar de nuevo desde el primer paso.
9. Consolidación del conocimiento descubierto, incorporándolo al sistema, o simplemente documentándolo y enviándolo a la parte interesada.



Procesos dentro de KDD

Este proceso presenta algunas limitaciones al tratar con *Datos Dinámicos*, datos cambiados o actualizados constantemente; *Datos Incompletos,* datos con valores nulos y el *Ruido e Incertidumbre*

dado por el tipo de valores de los atributos, por su medida y sobre todo el continuo crecimiento del tamaño de las bases de datos que no permite el análisis de toda la información.

La representación del conocimiento cuenta también con diferentes metodologías, entre las que se destacan:

- *Representaciones basadas en la Lógica de Proposiciones Extendidas*, denominadas O+ o representación objeto-atributo-valor, en las cuales pueden incluirse los *Árboles de Decisión,* que se basan en la partición del conjunto muestra; las *Reglas de Producción,* consideradas del tipo Si-Entonces y las *Listas de Decisión*.
- *Representaciones Estructuradas*, en las cuales se encuentran las redes semánticas y los marcos. Las primeras, constan de un conjunto de nodos (conceptos) unidos entre sí por diferentes clases de enlaces asociativos (relaciones); y los marcos, son un tipo de estructura, formada por un nombre y campos definidos para el modelo especifico.
- *Representaciones basadas en ejemplos*, donde el conocimiento se presenta mediante ejemplos representativos, basándose en similitudes entre los datos.
- *Representaciones basadas en redes neuronales*, que son sistemas formados por un conjunto de elementos de computación llamados neuronas artificiales; estas neuronas están interconectadas a través de enlaces con unos pesos asociados, que representan el conocimiento en la red.

Conclusiones y resultados:

Algunas de las áreas de aplicación de técnicas de análisis inteligente de información son:

- Predicción automática de tendencias y comportamientos.
- Marketing dirigido: Analizar datos sobre envíos por correo publicitarios para identificar el segmento más apropiado para realizar un nuevo mailing.
- Análisis de las ventas de una compañía farmacéutica para reforzar las acciones de marketing en los hospitales y médicos de mayor impacto.
- Identificación de mejores clientes para el lanzamiento de una nueva tarjeta de crédito.
- Descubrimiento automático de patrones ocultos .
- Análisis de datos de ventas de productos para identificar aquellos que sin estar relacionados entre sí, se compran juntos a menudo.
- Detección de transacciones fraudulentas realizadas con tarjeta de crédito
- Detección de errores de grabación de datos
- Estudios de respuesta a un posible cambio de precios .
- Segmentación y Clustering .
- Dividir la base de datos de clientes en segmentos relativamente homogéneos basados en conductas

estudiadas.

- Una organización bancaria puede estudiar qué grupo de usuarios tiene una alta probabilidad de cancelar su cuenta en función de determinados parámetros y a continuación realizar acciones específicas para evitar que ocurra.
- Análisis de los datos obtenidos a partir de instrumental científico. Esto permite el análisis de los datos para investigación, la formación de hipótesis y teorías.

Finalmente, se puede destacar la importancia de la Minería de datos, como una nueva técnica que ofrece una forma de manipular la incertidumbre asociada a la realidad y conduce a extraer un conocimiento mas certero, de más calidad, un conocimiento muchas veces ignorado, invisible y encriptado en el universo de información contenido en las bases de datos borrosas.

Referencias bibliográficas:

•Michael J. A. Berry, Gordon Linoff . **Data Mining Techniques, for marketing, sales and customer support**, Segunda Edición, EUA, *editorial wiley.* 2000.

http://dns1.mor.itesm.mx/~emorales/Cursos/KDD01/node2.html

http://www.datamining.com

Autores:

Daveiva Barrera Celis

e-mail: vita_5@hotmail.com

Carlos Francisco Páez Durán

e-mail: carlosfpd@hotmail.com

Erika Liliana Pabón Pereira

e-mail: erikapabon@hotmail.com

Elcy Patricia Peñaloza Leal

e-mail: elcypat@hotmail.com

Andrea Paola Sánchez Pérez

e-mail: phahola@hotmail.com