www.monografias.com

Data Mining

Descubriendo Información Oculta

Data Mining, ***la extracción de información oculta y predecible de grandes bases de datos***, es una poderosa tecnología nueva con gran potencial para ayudar a las compañías a concentrarse en la información más importante de sus Bases de Información (Data Warehouse). Las herramientas de Data Mining predicen futuras tendencias y comportamientos, permitiendo en los negocios tomar decisiones proactivas y conducidas por un conocimiento acabado de la información (knowledge-driven). Los ***análisis prospectivos*** automatizados ofrecidos por un producto así van más allá de los eventos pasados provistos por herramientas retrospectivas típicas de sistemas de soporte de decisión. Las herramientas de Data Mining pueden responder a preguntas de negocios que tradicionalmente consumen demasiado tiempo para poder ser resueltas y a los cuales los usuarios de esta información casi no están dispuestos a aceptar. Estas herramientas exploran las bases de datos en busca de patrones ocultos, encontrando información predecible que un experto no puede llegar a encontrar porque se encuentra fuera de sus expectativas.

Muchas compañías ya colectan y refinan cantidades masivas de datos. Las técnicas de Data Mining pueden ser implementadas rápidamente en plataformas ya existentes de software y hardware para acrecentar el valor de las fuentes de información existentes y pueden ser integradas con nuevos productos y sistemas pues son traídas en línea (on-line). Una vez que las herramientas de Data Mining fueron implementadas en computadoras cliente servidor de alta performance o de procesamiento paralelo, pueden analizar bases de datos masivas para brindar respuesta a preguntas tales como, "¿Cuáles clientes tienen más probabilidad de responder al próximo mailing promocional, y por qué? y presentar los resultados en formas de tablas, con gráficos, reportes, texto, hipertexto, etc.

Los Fundamentos del Data Mining

Las técnicas de Data Mining son el resultado de un largo proceso de investigación y desarrollo de productos. Esta evolución comenzó cuando los datos de negocios fueron almacenados por primera vez en computadoras, y continuó con mejoras en el acceso a los datos, y más recientemente con tecnologías generadas para permitir a los usuarios navegar a través de los datos en tiempo real. Data Mining toma este proceso de evolución más allá del acceso y navegación retrospectiva de los datos, hacia la entrega de información prospectiva y proactiva. Data Mining está listo para su aplicación en la comunidad de negocios porque está soportado por tres tecnologías que ya están suficientemente maduras:

- Recolección masiva de datos
- Potentes computadoras con multiprocesadores
- Algoritmos de Data Mining

Las bases de datos comerciales están creciendo a un ritmo sin precedentes. Un reciente estudio del META GROUP sobre los proyectos de Data Warehouse encontró que el 19% de los que contestaron están por encima del nivel de los 50 Gigabytes, mientras que el 59% espera alcanzarlo en el segundo trimestre de 1997. En algunas industrias, tales como ventas al por menor (retail), estos números pueden ser aún mayores. MCI Telecommunications Corp. cuenta con una base de datos de 3 terabytes + 1 terabyte de índices y overhead corriendo en MVS sobre IBM SP2. La necesidad paralela de motores computacionales mejorados puede ahora alcanzarse de forma más costo - efectiva con tecnología de computadoras con multiprocesamiento paralelo. Los algoritmos de Data Mining utilizan técnicas que han existido por lo menos desde hace 10 años, pero que sólo han sido implementadas recientemente como herramientas maduras, confiables, entendibles que consistentemente son más performantes que métodos estadísticos clásicos.

En la evolución desde los datos de negocios a información de negocios, cada nuevo paso se basa en el previo. Por ejemplo, el acceso a datos dinámicos es crítico para las aplicaciones de navegación de datos (drill through applications), y la habilidad para almacenar grandes bases de datos es crítica para Data Mining.

Los componentes esenciales de la tecnología de Data Mining han estado bajo desarrollo por décadas, en áreas de investigación como estadísticas, inteligencia artificial y aprendizaje de máquinas. Hoy, la madurez de estas técnicas, junto con los motores de bases de datos relacionales de alta performance, hicieron que estas tecnologías fueran prácticas para los entornos de data warehouse actuales.

El Alcance de Data Mining

El nombre de Data Mining deriva de las similitudes entre buscar valiosa información de negocios en grandes bases de datos - por ej.: encontrar información de la venta de un producto entre grandes montos de Gigabytes almacenados - y minar una montaña para encontrar una veta de metales valiosos. Ambos procesos requieren examinar una inmensa cantidad de material, o investigar inteligentemente hasta encontrar exactamente donde residen los valores. Dadas bases de datos de suficiente tamaño y calidad, la tecnología de Data Mining puede generar nuevas oportunidades de negocios al proveer estas capacidades:

- **Predicción automatizada de tendencias y comportamientos**. Data Mining automatiza el proceso de encontrar información predecible en grandes bases de datos. Preguntas que tradicionalmente requerían un intenso análisis manual, ahora pueden ser contestadas directa y rápidamente desde los datos. Un típico ejemplo de problema predecible es el marketing apuntado a objetivos (targeted marketing). Data Mining usa datos en mailing promocionales anteriores para identificar posibles objetivos para maximizar los resultados de la inversión en futuros mailing. Otros problemas predecibles incluyen pronósticos de problemas financieros futuros y otras formas de incumplimiento, e identificar segmentos de población que probablemente respondan similarmente a eventos dados.
- **Descubrimiento automatizado de modelos previamente desconocidos.** Las herramientas de Data Mining barren las bases de datos e identifican modelos previamente escondidos en un sólo paso. Otros problemas de descubrimiento de modelos incluye detectar transacciones fraudulentas de tarjetas de créditos e identificar *datos anormales* que pueden representar errores de tipeado en la carga de datos.

Las técnicas de Data Mining pueden redituar los beneficios de automatización en las plataformas de hardware y software existentes y puede ser implementadas en sistemas nuevos a medida que las plataformas existentes se actualicen y nuevos productos sean desarrollados. Cuando las herramientas de Data Mining son implementadas en sistemas de procesamiento paralelo de alta performance, pueden analizar bases de datos masivas en minutos. Procesamiento más rápido significa que los usuarios pueden automáticamente experimentar con más *modelos* para entender datos complejos. Alta velocidad hace que sea práctico para los usuarios analizar inmensas cantidades de datos. Grandes bases de datos, a su vez, producen mejores predicciones.

Las bases de datos pueden ser grandes tanto en profundidad como en ancho:

- **Más columnas**. Los analistas muchas veces deben limitar el número de variables a examinar cuando realizan análisis manuales debido a limitaciones de tiempo. Sin

embargo, variables que son descartadas porque parecen sin importancia pueden proveer información acerca de modelos desconocidos. Un Data Mining de alto rendimiento permite a los usuarios explorar toda la base de datos, sin preseleccionar un subconjunto de variables.

- **Más filas.** Muestras mayores producen menos errores de estimación y desvíos, y permite a los usuarios hacer inferencias acerca de pequeños pero importantes segmentos de población.

Las técnicas más comúnmente usadas en Data Mining son:

- **Redes neuronales artificiales:** modelos predecible no-lineales que aprenden a través del entrenamiento y semejan la estructura de una red neuronal biológica.
- **Arboles de decisión:** estructuras de forma de árbol que representan conjuntos de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos. Métodos específicos de árboles de decisión incluyen Arboles 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.
- **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* 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.

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.

¿Cómo Trabaja el Data Mining?

¿Cuán exactamente es capaz Data Mining de decirle cosas importantes que usted desconoce o que van a pasar? La técnica usada para realizar estas hazañas en Data Mining se llama *Modelado*. Modelado es simplemente el acto de construir un modelo en una situación donde usted conoce la respuesta y luego la aplica en otra situación de la cual desconoce la respuesta. Por ejemplo, si busca un galeón español hundido en los mares lo primero que podría hacer es investigar otros tesoros españoles que ya fueron encontrados en el pasado. Notaría que esos barcos frecuentemente fueron encontrados fuera de las costas de Bermuda y que hay ciertas características respecto de las corrientes oceánicas y ciertas rutas que probablemente tomara el capitán del barco en esa época. Usted nota esas similitudes y arma un modelo que incluye las características comunes a todos los sitios de estos tesoros hundidos. Con estos modelos en mano sale a buscar el tesoro donde el modelo indica que en el pasado hubo más probabilidad de darse una situación similar. Con un poco de esperanza, si tiene un buen modelo, probablemente encontrará el tesoro.

Este acto de construcción de un modelo es algo que la gente ha estado haciendo desde hace mucho tiempo, seguramente desde antes del auge de las computadoras y de la tecnología de Data Mining. Lo que ocurre en las computadoras, no es muy diferente de la manera en que la gente construye modelos. Las computadoras son cargadas con mucha información acerca de una variedad de situaciones donde una respuesta es conocida y luego el software de Data Mining en la computadora debe correr a través de los datos y distinguir las características de los datos que llevarán al modelo. Una vez que el modelo se construyó, puede ser usado en situaciones similares donde usted no conoce la respuesta.

Si alguien le dice que tiene un modelo que puede predecir el uso de los clientes, ¿Cómo puede saber si es realmente un buen modelo? La primera cosa que puede probar es pedirle que aplique el modelo a su base de clientes - donde usted ya conoce la respuesta. Con Data Mining, la mejor manera para realizar esto es dejando de lado ciertos datos para aislarlos del proceso de Data Mining. Una vez que el proceso está completo, los resultados pueden ser testeados contra los datos excluidos para confirmar la validez del modelo. Si el modelo funciona, las observaciones deben mantenerse para los datos excluidos.

Una arquitectura para Data Mining

Para aplicar mejor estas técnicas avanzadas, éstas deben estar totalmente integradas con el data warehouse así como con herramientas flexibles e interactivas para el análisis de negocios. Varias herramientas de Data Mining actualmente operan fuera del warehouse, requiriendo pasos extra para extraer, importar y analizar los datos. Además, cuando nuevos conceptos requieren implementación operacional, la integración con el warehouse simplifica la aplicación de los resultados desde Data Mining. El Data warehouse analítico resultante puede ser aplicado para mejorar procesos de negocios en toda la organización, en áreas tales como

manejo de campañas promocionales, detección de fraudes, lanzamiento de nuevos productos, etc.

El punto de inicio ideal es un data warehouse que contenga una combinación de datos de seguimiento interno de todos los clientes junto con datos externos de mercado acerca de la actividad de los competidores. Información histórica sobre potenciales clientes también provee una excelente base para prospecting. Este warehouse puede ser implementado en una variedad de sistemas de bases relacionales y debe ser optimizado para un acceso a los datos flexible y rápido.

Un server multidimensional OLAP permite que un modelo de negocios más sofisticado pueda ser aplicado cuando se navega por el data warehouse. Las estructuras multidimensionales permiten que el usuario analice los datos de acuerdo a como quiera mirar el negocio - resumido por línea de producto, u otras perspectivas claves para su negocio. El server de Data Mining debe estar integrado con el data warehouse y el server OLAP para insertar el análisis de negocios directamente en esta infraestructura. Un avanzado, metadata centrado en procesos define los objetivos del Data Mining para resultados específicos tales como manejos de campaña, prospecting, y optimización de promociones. La integración con el data warehouse permite que decisiones operacionales sean implementadas directamente y monitoreadas. A medida que el data warehouse crece con nuevas decisiones y resultados, la organización puede "minar" las mejores prácticas y aplicarlas en futuras decisiones.

Este diseño representa una transferencia fundamental desde los sistemas de soporte de decisión convencionales. Más que simplemente proveer datos a los usuarios finales a través de software de consultas y reportes, el server de Análisis Avanzado aplica los modelos de negocios del usuario directamente al warehouse y devuelve un análisis proactivo de la información más relevante. Estos resultados mejoran los metadatos en el server OLAP proveyendo una estrato de metadatos que representa una vista fraccionada de los datos. Generadores de reportes, visualizadores y otras herramientas de análisis pueden ser aplicadas para planificar futuras acciones y confirmar el impacto de esos planes.

Glosario de Términos de Data Mining

- **Algoritmos genéticos:** Técnicas de optimización que usan procesos tales como combinación genética, mutación y selección natural en un diseño basado en los conceptos de evolución natural.
- **Análisis de series de tiempo (time-series):** Análisis de una secuencia de medidas hechas a intervalos específicos. El tiempo es usualmente la dimensión dominanate de los datos.
- **Análisis prospectivo de datos:** Análisis de datos que predice futuras tendencias, comportamientos o eventos basado en datos históticos.
- **Análisis exploratorio de datos:** Uso de técnicas estadísticas tanto gráficas como descriptivas para aprender acerca de la estructura de un conjunto de datos.
- **Análisis retrospectivo de datos:** Análisis de datos que provee una visión de las tendencias , comportamientos o eventos basado en datos históricos.
- **Árbol de decisión:** Estructura en forma de árbol que representa un conjunto de decisiones. Estas decisiones generan reglas para la *clasificación* de un conjunto de datos. Ver *CART* y *CHAID.*
- **Base de datos multidimensional:** Base de datos diseñada para procesamiento analítico on-line (*OLAP*). Estructurada como un hipercubo con un eje por dimensión.
- **CART Árboles de clasificación y regresión**: Una técnica de *árbol de decisión* usada para la *clasificación* de un conjunto da datos. Provee un conjunto de reglas que se pueden aplicar a un nuevo (sin clasificar) conjunto de datos para predecir cuáles registros darán un cierto resultado. Segmenta un conjunto de datos creando 2 divisiones. Requiere menos preparación de datos que *CHAID* .
- **CHAID Detección de interacción automática de Chi cuadrado**: Una técnica de *árbol de decisión* usada para la *clasificación* de un conjunto da datos. Provee un conjunto de reglas que se pueden aplicar a un nuevo (sin clasificar) conjunto de datos para predecir cuáles registros darán un cierto resultado. Segmenta un conjunto de datos utilizando tests de chi cuadrado para crear múltiples divisiones. Antecede, y requiere más preparación de datos, que *CART.*
- **Clasificación**: Proceso de dividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo "más cercano" posible a otro, y grupos diferentes estén lo "más lejos" posible uno del otro, donde la distancia está medida con respecto a variable(s) específica(s) las cuales se están tratando de predecir. Por ejemplo, un problema típico de clasificación es el de dividir una base de datos de compañías en grupos que son lo más homogéneos posibles con respecto a variables como "posibilidades de crédito" con valores tales como "Bueno" y "Malo".
- **Clustering (agrupamiento)**: Proceso de dividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo "más cercano" posible a otro, y grupos diferentes estén lo "más lejos" posible uno del otro, donde la distancia está medida con respecto a todas las variables disponibles.
- **Computadoras con multiprocesadores**: Una computadora que incluye múltiples procesadores conectados por una red. Ver *procesamiento paralelo.*
- **Data cleansing:** Proceso de asegurar que todos los valores en un conjunto de datos sean consistentes y correctamente registrados.
- **Data Mining**: La extracción de información predecible escondida en grandes bases de datos.
- **Data Warehouse**: Sistema para el almacenamiento y distribución de cantdades masivas de datos
- **Datos anormales**: Datos que resultan de errores (por ej.: errores en el tipeado durante la carga) o que representan eventos inusuales.
- **Dimensión**: En una base de datos relacional o plana, cada campo en un registro representa una dimensión. En una *base de datos multidimensional*, una dimensión es un conjunto de entidades similares; por ej.: una base de datos multidimensional de ventas podría incluir las dimensiones Producto, Tiempo y Ciudad.

- **Modelo analítico**: Una estructura y proceso para analizar un conjunto de datos. Por ejemplo, un *árbol de decisión* es un modelo para la *clasificación* de un conjunto de datos
- **Modelo lineal:** Un *modelo analítico* que asume relaciones lineales entre una variable seleccionada (dependiente) y sus predictores (variables independientes).
- **Modelo no lineal**: Un *modelo analítico* que no asume una relación lineal en los coeficientes de las variables que son estudiadas.
- **Modelo predictivo:** Estructura y proceso para predecir valores de variables especificadas en un conjunto de datos.
- **Navegación de datos:** Proceso de visualizar diferentes dimensiones, "fetas" y niveles de una *base de datos multidimensional*. Ver *OLAP*.
- **OLAP Procesamiento analítico on-line (On Line Analitic prossesing):** Se refiere a aplicaciones de bases de datos orientadas a array que permite a los usuarios ver, navegar, manipular y analizar *bases de datos multidimensionales*.
- **Outlier:** Un item de datos cuyo valor cae fuera de los límites que encierran a la mayoría del resto de los valores correspondientes de la muestra. Puede indicar *datos anormales*. Deberían ser examinados detenidamente; pueden dar importante información.
- **Procesamiento paralelo**: Uso coordinado de múltiples procesadores para realizar tareas computacionales. El procesamiento paralelo puede ocurrir en una *computadora con múltiples procesadores* o en una red de estaciones de trabajo o PCs.
- **RAID**: Formación redundante de discos baratos (Redundant Array of inexpensive disks). Tecnología para el almacenamiento paralelo eficiente de datos en sistemas de computadoras de alto rendimiento.
- **Regresión lineal:** Técnica estadística utilizada para encontrar la mejor relación lineal que encaja entre una variable seleccionada (dependiente) y sus predicados (variables independientes).
- **Regresión logística:** Una regresión lineal que predice las proporciones de una variable seleccionada categórica, tal como Tipo de Consumidor, en una población.
- **Vecino más cercano:** 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* 1). Algunas veces se llama la técnica del vecino *k*-más cercano.

- **SMP Multiprocesador simétrico (Symmetric multiprocessor):** Tipo de *computadora con multiprocesadores* en la cual la memoria es compartida entre los procesadores

Terabyte: Un trillón de bytes.

Trabajo realizado por

Cynthia Presser Carne

CynthiaP@CicBue.com