"Uso y diseño de Mineros de Datos"

RESUMEN. Los mineros son productos recientes que trabajan, principalmente, sobre bases de datos relacionales que contienen información extensa. Trabajan por lo general de manera asíncrona (como demonios o procesos autónomos), y buscan de manera exhaustiva datos anómalos, interesantes, desviaciones, tendencias o anomalías. En este artículo se describen sus características y elementos básicos, algunos de sus principales usos y de los problemas a resolver.

1. INTRODUCCIÓN

Un sistema de Minería de Datos está formado por varios programas de cómputo, que realizan la búsqueda en una base de datos, de manera automática, de tendencias, desviaciones, anomalías, patrones y situaciones "interesantes." Estas desviaciones o tendencias son reportadas inmediatamente, o más a menudo en un archivo, para su posterior visualización y decisión final. Existen varios algoritmos generales para "minería de datos". El Sistema a menudo posee un configurador que permite particularizarlos a casos y situaciones específicas; una agenda de trabajo guía a cada minero en sus búsquedas. La prioridad de búsqueda (orden en la agenda), los criterios de "interés" y "tolerancia" y la definición de "situación análoga" son modificables por el usuario, de suerte que un minero originalmente posee un criterio general pero quizá fuera de foco de lo que es "interesante" para su usuario, y termina -merced a numerosas intervenciones o enseñanzas de su usuario- como una colección de límites de valores y criterios de éxito muy específicos, que busca en determinada área de la base de datos y bajo condiciones o predicados igualmente ajustados a la medida.

El sistema es un auxiliar valioso para las áreas de planeación, evaluación, apoyo a la toma de decisiones, y otras donde la necesidad de análisis semi-exhaustivo sea extensa, debido al gran número de datos. Uno de estos sistemas (construido por el autor) está programado en una mezcla de Informix y C, bajo el Sistema Operativo Unix SCO. Este artículo explica un poco su uso y otro poco su arquitectura o construcción.

Existen varios sistemas de minería de datos, generalmente hechos a la medida, para análisis de transacciones de tarjetas de crédito ("¿será fraudulenta esta compra? Ya van tres televisores que compra en media hora"), de ventas en cadenas de tiendas ("He notado que los pantalones vaqueros talla 32 largo 31 color morado ya no se venden tanto"; "ha habido una demanda superior a lo esperado de chiles en nogada"), tendencias en enfermedades, en accidentados, etc. Los mineros trabajan muy a gusto con grandes cantidades de datos; sus competidores son sistemas de análisis estadístico de datos, reconocedores de patrones (clasificadores supervisados y no supervisados), y visualizadores de información (donde el usuario es el que hace el análisis y se percata de la anomalía o situación interesante).

2. LA BÚSQUEDA DE INFORMACIÓN RELEVANTE EN UN MAR DE DATOS RUTINARIOS.

La concentración de información en una sola base de datos (llamada Base de Datos Corporativos o Corporativa), donde se conjunta información sumariada de toda la empresa, es posible a costos razonables en la actualidad debido a tres tecnologías disparadoras: la proliferación de computadoras personales y de bajo costo; el abaratamiento del disco Winchester (menos de un dólar de EEUU el megabyte), y la instalación de redes de inter-comunicación de datos.

El objeto principal de la base de datos corporativa es la de compartir (en la cúpula de la organización) el mismo juego de datos en toda la empresa; hacer seguimiento, control, evaluación, escenarios alternos; comparación de lo presupuestado o planeado con la vida real (resultados de la operación de la empresa en determinado lapso transcurrido), y hacer análisis de desviaciones y reportes no planeados.

2.1 .Análisis de situaciones interesantes

El análisis de situaciones interesantes se lleva a cabo actualmente por medios sobre todo manuales, a saber:

* Mediante preguntas pre-formuladas, escritas en un lenguaje de tercera generación (Cobol, digamos) o de cuarta, los que dan reportes diseñados manualmente, pero cuyo resultado nos es útil para medir desempeño, productividad, costos unitarios, etc. Son los "indicadores de gestión." Las salidas o respuestas son generalmente numéricas.

* Mediante preguntas espontáneas, no planeadas, utilizando el lenguaje SQL (lenguaje estándar de consulta) o uno más gráfico, como el QBE (pregunta mediante ejemplos). Las salidas o respuestas son textuales: números y letras, en columnas.

Desplegando los resultados de las preguntas anteriores en forma tabular; en gráficas de tipo pastel, barra, etc.; en histogramas o diagramas de dispersión; en forma de nubes o cúmulos; sobre mapas para formar lo que se ha denominado "Cartas Temáticas" que despliegan su contenido sobre cartas geográficas.

2.1.1. Desventajas del análisis manual

Este tipo de análisis manual o visual, llevado a cabo por una persona, tiene varias desventajas:

* El tiempo disponible es limitado. Muchas situaciones interesantes se soslayan (y no se detectan) simplemente porque toma tiempo pensar o imaginarse la pregunta, elaborarla; expresarla (en SQL, por ejemplo) de manera correcta sintácticamente; interpretarla o compilarla; hacer la búsqueda en la base de datos, y analizar (ver, contemplar) la respuesta para determinar si es "interesante" o no. El usuario se cansa, o se le acaba el tiempo.

* El usuario no percibe una situación interesante. "No la ve".

* No se le ocurre buscar en determinado lugar, y ahí estaba algo muy interesante.

Por consiguiente, hace falta un tipo de análisis automatizado, que complemente al análisis manual. Es decir, un sistema donde la computadora proponga lugares [regiones de la base de datos] donde buscar cosas interesantes; donde los criterios de interés se propongan automáticamente o estén predefinidos por la computadora, y/o modificados por el usuario; donde las búsquedas se vayan efectuando poco a poco; y donde se vayan recopilando aquéllos datos que resultaron (usando criterios predefinidos en la máquina ó, mejor aún, modificados por el usuario) interesantes, a efecto de que el usuario pueda verlos y determinar por sí mismo qué tan interesantes o rutinarios en realidad son.

2.2. Objetivo de un sistema de Minería de Datos

Ayudar al planificador o gerente en la toma de decisiones, mediante la detección automática de anomalías, desviaciones, tendencias, patrones, y situaciones "interesantes." El criterio de "interés" originalmente viene dado por el sistema, pero el usuario puede modificarlo (mediante retroalimentación) poco a poco. Es decir, el sistema de minería de datos propuesto tiene la virtud de adaptarse o aprender de su usuario, por lo que conforme pase el tiempo, reflejará los gustos, intereses y preocupaciones del usuario mismo.

Un programa que ayude a buscar situaciones interesantes con los criterios correctos, complementa enormemente una labor que hasta ahora se ha considerado "intelectual" y de alto nivel, privativa de los gerentes, planificadores y administradores. Además, la búsqueda se realiza fuera de horas pico, usando tiempos de máquina excedentes.

3. .ESTRUCTURA DE UN SISTEMA DE MINERÍA DE DATOS

El sistema que construí tiene los siguientes componentes funcionales: los algoritmos o programas que buscan (los mineros); en dónde se busca (la base de datos); qué se busca (tipificación de lo interesante o anómalo); cómo se busca (orden, secuencia de búsqueda, qué se busca primero); y qué se hace con lo encontrado (almacenamiento de hallazgos).

Quién busca
: Los mineros

Dónde
: En la base de datos

Qué
: Criterio de interesante

Cómo
: Agenda

Almacenamiento de lo encontrado
: Cofre de tesoros

3.1. Los mineros

La espina dorsal del sistema de minería de datos la forman varios algoritmos generales de búsqueda y detección de "situaciones interesantes". Estos algoritmos se guían por un árbol de conceptos que describe la relación que guardan los diferentes temas y rubros en la base de datos.

Un minero consta de dos partes: un extractor que saca o extrae cierto conjunto de datos que podrían contener algo de interés, y un módulo "revisor" o "verificador" que, mediante análisis matemáticos o estadísticos, dictamina si hubo algo interesante en el subconjunto de datos extraídos. Esta división tiene un sentido práctico, pues refleja la situación de la base de datos Informix (estoy describiendo el sistema que se construyó), que yace en un servidor Unix ; en tanto que los módulos revisores o verificadores son clientes que están situados en PC's bajo DOS. Estas dos partes trabajan bajo el esquema cliente-servidor, o en un programa monolítico si el extractor (junto con la base de datos) y el revisor yacen en la misma computadora.

El extractor se guía por el árbol de conceptos para accesar la base de datos, y por los "criterios de interés" para saber qué conviene extraer para su análisis posterior por el verificador o revisor.

El verificador hace su trabajo usando parámetros que le permiten clasificar o desechar los datos a él presentados como "interesantes" o "rutinarios."

3.1. Mineros típicos

a) Buscador dirigido diferido. Se le da el criterio de búsqueda (mediante un predicado o filtro), el que generalmente se programas en C ó en un lenguaje especial parecido al SQL pero que toma en cuenta el árbol de conceptos. Ejemplo: minero de umbrales. Se le dan las fórmulas a evaluar para comparar si exceden de un límite o umbral (ventas - 10 * rechazados). Es decir, se "arma" manualmente al minero. Ejemplo: minero de pendientes: repórtame si durante tres meses seguidos han bajado las ventas en 10% o más. En términos técnicos, este buscador realiza un ajuste de curvas.

b) Buscador de índices de productividad. Cuando se conoce la función a evaluar, con una fórmula se le definen los índices de productividad o desempeño. El minero evalúa el índice y reporta por ejemplo las n mejores regiones y las m peores.

c) Funciones booleanas. Partiendo las variables reales en un número pequeño de clases de equivalencia, si es necesario, este minero trata de ver las cosas "en blanco y negro". Para él lo que está pasando es que hay una función booleana f(v1, v2, .., vk) que vale 1 si hay anomalía o interés en ese punto, y 0 si no la hay.

d) Correlaciones entre señales del tiempo. Se buscan las tendencias de varias señales o variables cuando se les considera que fluyen a través del tiempo; es decir, la correlación entre f1(t) y f2(t). Un caso interesante es cuando f1 no es variable endógena (no es un dato de la compañía, sino es un dato exógeno, por ejemplo, número de teléfonos instalados por cada 100 mil habitantes). Este tipo de mineros que buscan simultáneamente en más de una variable son difíciles de programar.

3.2. La base de datos

Los datos están guardados en una base de datos relacional (o, menos comúnmente, en otro tipo de base). La búsqueda se hace sobre datos numéricos (ventas, número de reprobados, cantidad de medidores anómalos, ...); es decir, las variable(s) independiente(s) son numéricas: aquéllas que determinan la anormalidad de una situación. Empero, lo que se busca puede ser no numérico. Por ejemplo, con base en las ventas mensuales de energía eléctrica (variable independiente, numérica) se puede concluir que la SUCURSAL PUEBLA (variable textual) es "la anómala" o "la interesante."

Podrá haber una gran variedad de tablas, diseñadas para otros usos de la base de datos. Muchas de estas tablas carecen de tesoros y conviene señalarlas como estériles a los mineros. Por ejemplo, una tabla que liga el número con el nombre del Municipio; otra tabla que describe la dirección, teléfono y código postal del Gerente de cada centro de trabajo. Esto se hace de manera directa: se le dice al minero en qué tabla buscar (y ya).

3.3. El árbol de conceptos

Los datos con los que los mineros trabajan se refieren a "coordenadas" o ejes. Como ejes típicos tenemos: el tiempo; eje geográfico; eje de productos. Ejemplo: Se vendieron $812 el 18 de abril de 1996 en Salina Cruz, Oax., en zapatos Bostonianos negros talla 7. Una tabla puede tener varios ejes. Ejemplo de otros ejes: tipo de enfermedad; grado militar en el ejército; tipo de empleado; tipos de carreras y planes de estudio. Cada uno de estos eje generalmente forma un árbol porque contiene distintos niveles de agregación.. Los días se agregan en semanas, en meses, en años. Los martillos, desarmadores y formones se agrupan en herramientas; éstas, junto con clavos, tornillos y bisagras, en ferretería, etc.

Cada punto o medición (una venta, en el ejemplo) es un punto en un cubo o hipercubo de varias dimensiones. A lo largo de cada dimensión existe un árbol (por ejemplo, el árbol de Oficinas Nacionales - División - Zona - Agencia) que se encuentra "aplanado" para poder representarse a lo largo de esa dimensión. Para aclarar representamos un cuadrado (dos ejes) de totales (o valores) con el árbol de productos en el eje vertical y el árbol geográfico en el eje horizontal. Cada celda representa ventas de cierto producto en cierta región geográfica.

3.3.1. Descripción del árbol de cada eje

Los mineros necesitan una pequeña tabla donde se describa la estructura de cada eje (del eje geográfico; del eje temporal, ...), así como cuáles son los campos que tienen valor geográfico, valor temporal, etc. (Es decir, dónde se encuentran en la base de datos cada una de estas coordenadas). Esto es con el objeto de que sepan navegar de un nivel de agregación hacia otro.

4. CONSIDERACIONES PRÁCTICAS Y CONCLUSIONES

4.1. Criterio de "lo interesante"

Se describen a los mineros las regiones donde tendrán que buscar cosas interesantes, así como qué constituye una condición interesante a reportar (por ejemplo, una variación de más de 20% del valor actual contra el promedio de los cuatro valores anteriores) o anómala. Como es fácil confundir al usuario si le pedimos que describa conceptos demasiado abstractos ("número de desviaciones estándar de la muestra para considerarse significativa"), la definición o descripción de los criterios de "interés", "anomalía", etc., se llevarán a cabo a dos niveles posibles.

* un nivel simple, fácil de usar, donde el usuario define el tipo de búsqueda con menús conteniendo descripciones tales como "pesimista", "20% de desviación", "indicadores primarios", etc.

* un nivel experto, donde el usuario da parámetros como número de desviaciones estándar, niveles de confianza, relación de señal a ruido, etc. Este nivel es el adecuado para un usuario con conocimientos estadísticos o matemáticos.

4.2. Órden de búsqueda, criterio de "lo primero a buscar"

Se pueden vetar regiones donde no se considera que habrá situaciones de interés. Los mineros no buscarán en ellas.

Se pueden jerarquizar, mediante una lista que constituye un "esqueleto de agenda", el orden de búsqueda o la prioridad relativa de búsqueda: primero busca tales y cuales indicadores de productividad; luego busca ventas mayores que x; luego ....

Se podrá especificar el orden de búsqueda de una manera más global: busca a profundidad primero; busca a lo ancho primero.

En general, una agenda es perenne. Es decir, el minero busca en esta base de datos de acuerdo con la agenda actual. La próxima vez que haya nuevos datos (la próxima "tanda" de datos) usará también esta misma agenda, este mismo orden de búsqueda.

4.2.1. Almacenamiento de hallazgos

Las situaciones interesantes que se vayan encontrando se guardan en un archivo o tabla para que puedan ser analizados manualmente (visualmente) por el usuario.

4.3. Retroalimentación, enseñanza, adecuación

El usuario, al ver lo encontrado por el minero, querrá afinarlo, enseñarle, modificar su criterio. Esto se podrá hacer con las interfaces "para usuarios simples" o "para expertos".

4.4 .Algunos problemas a resolver

Mencionamos dos de los problemas que se presentarán en la explotación práctica de los mineros.

* Agujeros, carencia de datos. A menudo no hay valores para ciertos meses, o para ciertas agencias. No es que no haya habido ventas, es que no se sabe cuánto se vendió. Claramente, un minero desprevenido identificaría "cero ventas" como una situación muy interesante, cuando en realidad el 0 ahí significa "no sé". Solución: utilizar un valor tal como "?" para indicar "no sé"; por ejemplo, un acceso a una celda que regrese "llave inválida" significa "no se encontró el dato", o sea, "no sé su valor."

* Saltos y discontinuidades en las variables. A veces una variable de pronto cambia bruscamente, debido a que su definición cambió. Por ejemplo, las ventas de 1992 a 1993 se redujeron mil veces, debido a la introducción de los nuevos pesos. Para resolver este problema, el usuario tendrá que hacer un "cálculo al valor constante", introduciendo una fórmula que diga: v' = v si t > 1993, de lo contrario es v/1000.

4.5. Algunos productos comerciales de minería de datos

No hay productos comerciales generalizados para minería de datos. Generalmente son desarrollos recientes que hay que ajustar al cliente. Por ejemplo, IBM Almaden vende un "Data mining facility". En EEUU, SoftwarePro International comercializa un sistema llamado "Intelligent Synthesizer/Analyzer" que posee un módulo de minería de datos. En México, este producto lo comercializa (y lo ajusta a la medida) Informática Directiva Aplicada, S. A.

Otros productos que compiten o tratan de realizar la misma función son:

* Analizadores estadísticos, paquetes de estadística. Buenos, pero hay que saber programarlos, y quizá no funcionen de manera autónoma.

* Visualizadores. Producen representaciones (gráficas) a colores, que permiten descubrir de manera manual (por el usuario) la situación interesante.

* Productos para Bodegas de datos («Data Warehouses»). Realizan la recolección de datos, pero no necesariamente su búsqueda automática.

4.6. Conclusiones

Los mineros son productos recientes que trabajan sobre bases de datos relacionales conteniendo información extensa. Trabajan por lo general de manera asíncrona (como demonios o procesos autónomos), y buscan de manera exhaustiva datos anómalos, interesantes, desviaciones, tendencias o anomalías. Para detectar tendencias incipientes, utilizan la suma de muchos valores: es difícil detectar, viendo las ventas de clavos de 3 pulgadas, si hay algo interesante (la señal tiene mucho ruido), pero sumando las ventas de todos los tamaños de clavos, se puede detectar ya una caída significante de ventas.

Una ventaja de los mineros es que la búsqueda la hacen de manera autónoma y automática, de noche o en horas de poco proceso, convirtiéndose en ayudantes importantes que utilizan el mismo criterio que el tomador de decisiones (un gerente de producto, por ejemplo).

Otra ventaja es que no requieren hardware especial o dedicado. Trabajan en las redes de oficinas nacionales (o regionales), utilizando por las noches el servidor relacional de bases de datos, y las PCs o estaciones de trabajo (donde yacerán los revisores o verificadores) ya existentes. Es decir, trabajan sobre datos ya recolectados, en máquinas ya existentes, realizando labores útiles mientras los usuarios duermen.

SoftwarePro International

aguzman@www.cic.ipn.mx
Dr. Adolfo Guzmán Arenas