MINERÍA DE DATOS: cómo hallar una aguja en un pajar

Gilberto Lorenzo Martínez Luna

Recientemente la computación ha creado la *minería de datos*. Este sistema tiene como finalidad prevenir a los directivos de empresas sobre situaciones interesantes, anomalías, e incluso peligros no detectados con anticipación. Los llamados "mineros" son auxiliares indispensables para el ejecutivo de cualquier empresa bien organizada.

Las instituciones y empresas privadas coleccionan bastante información (ventas, clientes, cobros, pacientes, tratamientos, estudiantes, calificaciones, fenómenos meteorológicos, etcétera, según su giro), aprovechando que las computadoras y los discos de almacenamiento se han abaratado, y las comunicaciones son también baratas y confiables. Esta información reside en *bases de datos operacionales*, llamadas así porque con ellas se lleva a cabo la *labor sustantiva* de la empresa: envío de mercancía a clientes, registro de estudiantes, tratamiento a pacientes, cobranza, entre otros.

Posteriormente la información se depura y sumariza (resume) para transferirla a bases de datos conocidas como *bodegas de datos*. Son "fotografías" periódicas (trimestrales, digamos) del estado de la empresa. Aquí se lleva a cabo la *labor estratégica* de la misma: averiguar qué pasa en ella. ¿Qué productos se venden significativamente menos? ¿Ha habido un auge inesperado de deserciones de las carreras en las ciencias sociales? ¿El aumento de la inversión en perforación de nuevos pozos no guarda proporción con la disminución de las reservas probables y probadas de hidrocarburos? Ésta es la zona de las decisiones estratégicas, y los sistemas usados para ellas se conocen como Sistemas de Apoyo a la Toma de Decisiones. Estos sistemas muestran al funcionario los indicadores principales del estado de la empresa (en el último bimestre, digamos). El funcionario indaga o averigua situaciones que él cree son de interés o preocupación. El sistema contesta con datos y

gráficas para que aquél pueda tomar decisiones. Aunque el directivo o gerente tiene la experiencia necesaria, a menudo (por falta de tiempo, o porque no se le ocurrió) no mira situaciones que están tomando rumbos interesantes, peligrosos quizá. Así, ciertas decisiones importantes pueden ser soslayadas, ignoradas, o tomarse ya muy tarde. Se pueden así desperdiciar oportunidades o admitirse riesgos indeseables.

Recientemente, la computación ha inventado la *minería de datos*, en auxilio del directivo que toma decisiones. En las bodegas de datos se colocan "mineros", algoritmos que buscan tendencias, anomalías, desviaciones o situaciones interesantes pero desconocidas, y otros eventos importantes. Estos mineros auxilian al directivo al mando del timón de la institución a mantener el mejor rumbo posible. Utilizan, además de las bases de datos, la *inteligencia artificial* (procedimientos para hallar grupos en situaciones similares, clasificar eventos nuevos en categorías conocidas, etcétera) y la estadística. Pero a diferencia de esta última, que toma una muestra de los datos y la estudia, la minería de datos estudia *todos* los datos. Mientras más datos se analicen, más precisa es, y su poder de detección y predicción aumenta.

En este artículo hablaremos de los mineros. En un mundo globalizado, donde es importante saber lo que ocurre en el entorno de la institución, en su contexto, los mineros son auxiliares indispensables para el ejecutivo de una empresa bien organizada. Para que los mineros trabajen bien, la empresa debe: a) tener registros operacionales que apoyen sus trabajos cotidianos, sus *funciones sustantivas*; b) "fotografiar" periódicamente estos registros, resumiéndolos (sumarizándolos), en "instantáneas trimestrales" que forman parte de la *bodega de datos*; y c) crear y depurar sus *mineros de datos*, haciéndolos trabajar exhaustivamente sobre la bodega de datos. En los primeros tres apartados de este artículo abordaremos estos aspectos. Finalmente, en el cuarto y final, daremos ejemplos de mineros creados y usados en México.

I. La operación cotidiana de la empresa

¿De dónde proviene el mar de datos?

Todas las organizaciones y empresas coleccionan y administran datos de su interés relacionados con personas, procesos u otro tipo de actividades para las cuales fueron creadas. Los más comunes son los relacionados con ventas de productos o servicios, empleados, pacientes o con clientes, o tan sofisticados como los que usa una organización dedicada a pronosticar el clima, o en actividades muy especializadas, como la detección de fraudes en el consumo de energía eléctrica.

Las colecciones se pueden almacenar en discos de gran capacidad, que es ya posible comprar y tener en el hogar, y que pueden ser del tamaño de la palma de la mano o menos. Para darnos una idea de su capacidad, pueden almacenar el número del Registro Federal de Causantes (RFC) y la edad de cada uno de los habitantes de la República Mexicana, para lo cual basta un disco con capacidad de almacenamiento de un *terabyte* (mil *gigabytes*, equivalentes a un millón de *megabytes*, o 10^{12} *bytes*, es decir, 10^{12} caracteres).

El uso del mar de datos que surte al océano de datos

En general estas colecciones tienen dos principales tipos de usos o aplicaciones:

a) El primer uso es en aplicaciones conocidas como *procesamiento de transacciones en línea* (OLTP, por sus siglas en inglés). En este tipo de aplicaciones, las tran-

sacciones son para adicionar más información, realizando operaciones sobre uno o algunos datos de su interés, datos que también pueden ser borrados o modificados. Estas transacciones se llevan a cabo regularmente todos los días (ver artículo "La información es poder… sobre todo si está en una base de datos", de Hugo César Coyote, en este mismo número de *Ciencia*). Ejemplos de adición de nuevos datos es el registro de nuevas ventas o nuevos clientes; ejemplos de modificaciones a ellos es la disminución del saldo de las deudas por pago de los deudores, o cuando se incrementa la deuda por compras con tarjeta de crédito; y ejemplos de borrado es cuando ya no es necesario almacenar datos de clientes que ya no compran, de deudas ya pagadas, de calificaciones de alumnos que ya terminaron sus estudios en una escuela, de inventarios de años anteriores, o de ventas diarias de años anteriores, entre otras situaciones.

Como muestra, en la Tabla 1 se indican números aproximados de transacciones que administran algunas empresas a nivel nacional en México.

II. Las bodegas de datos se usan para tomar decisiones estratégicas

Al paso del tiempo, los datos de las aplicaciones OLTP se transfieren, con una serie de procesos conocidos como *extracción*, *transformación* y *limpieza* a colecciones llamadas *bodegas de datos*, donde su segundo uso es el análisis; ya sea con el *procesamiento analítico en línea* (OLAP, por sus siglas en inglés, *OnLine Analytical Processing*), o la *minería de datos*. Ambos análisis se caracterizan por utilizar un gran número de datos de interés (caso contrario de las OLTP) que se generaron a través de varios días, meses o años, de acuerdo con el interés de la organización.

En la Tabla 2 se dan valores aproximados del número de datos que se almacenan por varios años en una bodega de datos.

¿Cómo trabaja el análisis OLAP?

En las bodegas, los datos se organizan en lo que se conoce como *cubo de datos*, cuyos componentes principales son las variables de análisis conocidas como *dimensiones*, y la variable numérica a revisar llamada *hecho* o *medida*.

Un ejemplo de un cubo de datos con cuatro dimensiones y una medida a analizar puede verse en la Tabla 3, y la Figura 1 muestra una representación gráfica.

Las operaciones que aquí se realizan son principalmente *conteos* de datos, *sumas* de sus ventas o su producción y otras operaciones como saber el *máximo* o *mínimo* o *promedio* en un periodo de tiempo. Cuando se hace lo anterior, se dice que se desarrolla el análisis OLAP, y el resultado sirve como base para tomar decisiones, pues se revisa el comportamiento de interés.

Los análisis se visualizan en gráficas, en las que se pueden inferir situaciones de interés. Por ejemplo, en un *conteo* de pérdidas en varios meses, una gráfica podría mostrar que es una tendencia a crecer; otra gráfica

Tabla 1. Ejemplo de transacciones que almacenan algunas bases de datos

Empresa	*Dato de interés*	*Año*	*Transacciones mensuales (año 2010), en millones*	*Transacciones anuales (millones)*	*Otras transacciones*
Comisión Federal de Electricidad (CFE)	Clientes 34.2 millones	2010	Consumo-pago, 68.4	820.8	
Teléfonos de México (Telmex)	Líneas 15.6 millones	2010	Servicio-pago, 31.2	374.4	Llamadas en un trimestre 4 900 millones
Comercial Mexicana	Productos 70 000	2008	Compras, 22	264	
Instituto Mexicano del Seguro Social (IMSS)	Pacientes 44 693 474	2010	Consultas, 10.2 (hasta noviembre)	120	
Instituto Politécnico Nacional (IPN)	Estudiantes 160 000	2010	Calificaciones de cuatro materias 0.64	(Seis evaluaciones) 3.84	

Tabla 2. Ejemplo de historial de datos almacenados en una bodega de datos

Empresa	*Transacciones anuales*	*10 años*
CFE	820.8 millones consumos y pagos	8 208 millones
Telmex	374.4 millones de servicios y pagos	3 744 millones
Comercial Mexicana	264 millones de compras	2 640 millones
IMSS	120 millones de consultas	1 200 millones
IPN	3.84 millones de calificaciones	38.4 millones

Tabla 3. Ejemplo de cubo de datos para analizar consumos de energía

Dimensión/ valor	*Descripción*	*Valores por dimensión*
1. Medidor		$34*10^6$
2. Tarifa	Tipos de tarifas en la República Mexicana	Aproximadamente más de 100
3. División	División geográfica propia de CFE	13
4. Mes	12 por año	12
Medición: Consumo	Consumo	Más de $34 \times 10^6 \times 100 \times 13 \times 12$ consumos en un año

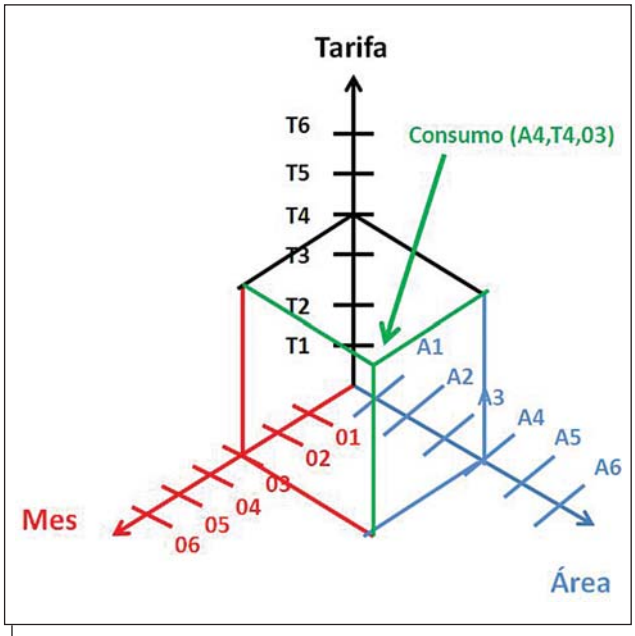


Figura 1. Representación gráfica del cubo con sólo tres dimensiones para analizar consumos de energía.

que muestre *sumatorias* (sumas) de producción de derivados de petróleo en dos años podría indicar si la producción se mantiene en los dos años; otra gráfica con las sumatorias de nacimientos contra muertes por año en un periodo de 55 años podría indicar cuándo habrá una coincidencia de ambas (muertes y nacimientos).

El análisis OLAP, con el historial de las actividades que han realizado los generadores de los datos, se realiza de manera manual, y dirigida por quien está al frente de la computadora revisando los cubos.

III. La minería de datos al auxilio al alto ejecutivo

La minería de datos se especializa en realizar estas tareas con ayuda de una computadora, apoyándose en un modelo de trabajo o proceso que se ha construido con la secuencia que se indica en la Figura 2. En esta sección nos concentraremos en la etapa de *minería de datos*.

¿Cómo trabaja la minería de datos?

Para detectar situaciones interesantes y anomalías (desviaciones de lo previsto), el *software* que lleva a cabo minería de datos se vale de varias técnicas y procedimientos ("algoritmos"). Algunos son:

- *Umbrales*: si tenemos un registro periódico (diario, semanal, etc.) de alguna variable de interés (las ventas de cierto producto, digamos) podemos fijarles un máximo "tolerado", arriba del cual nos interesa detectar excesos, y un mínimo "permitido", aba-

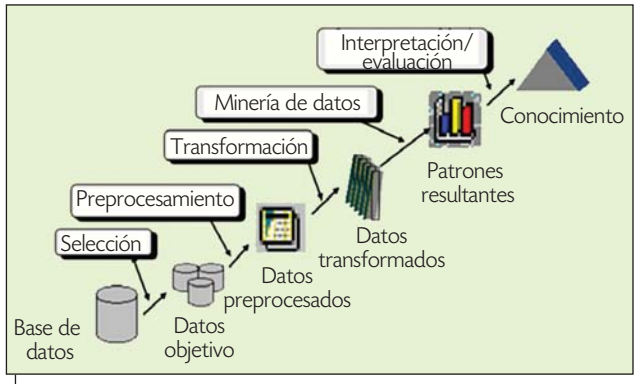
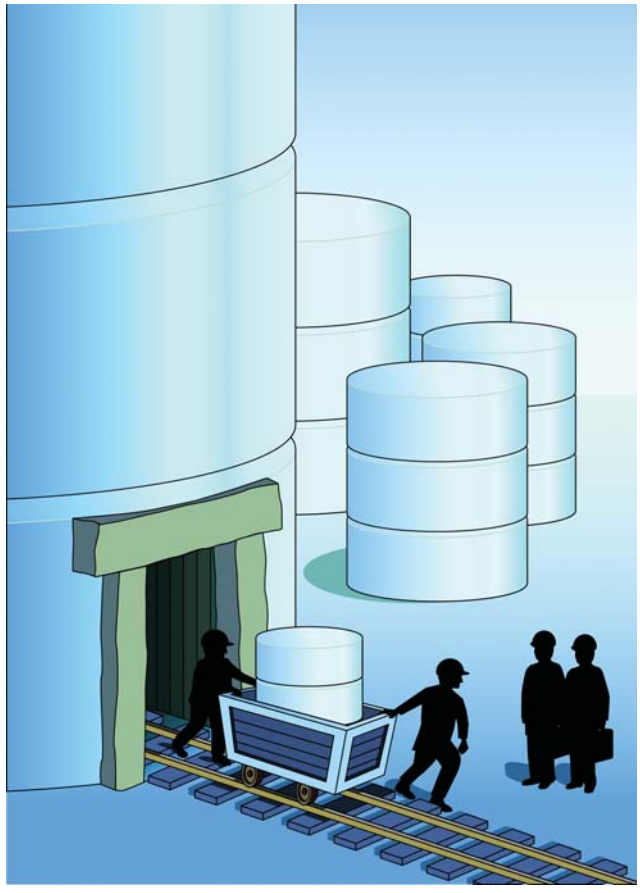


Figura 2. Fases del proceso de Descubrimiento en Bases de Datos. *Tomada de Usama Fayyad y Evangelos Simoudis.*

jo del cual deseamos que el minero nos informe. El algoritmo observa las ventas conforme pasan los días, y cuando detecta un valor más allá de los límites o umbrales fijados, nos avisa. Para no distraernos con "picos" pasajeros, podemos programar al minero para que nos avise si hay más de tres picos consecutivos (en tres semanas seguidas, por ejemplo).

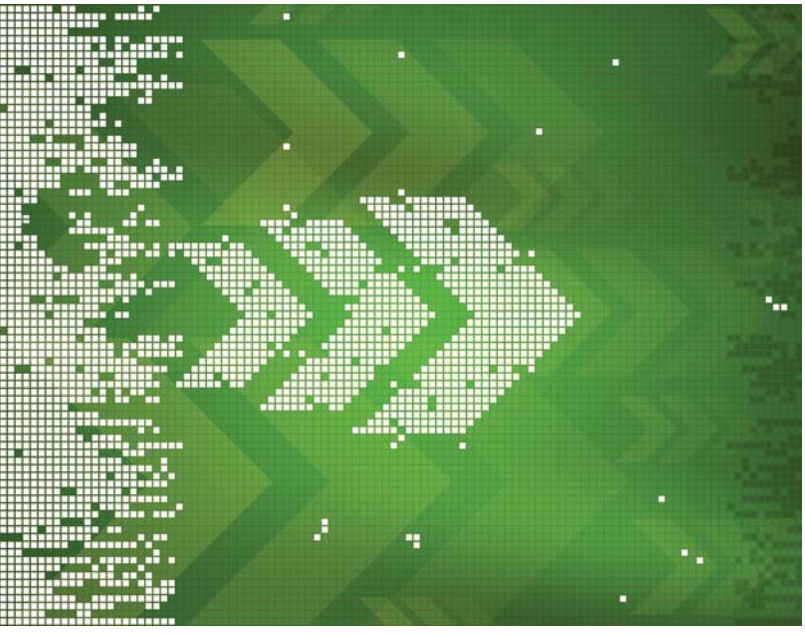
- *Tendencias*: este algoritmo observa si de una semana a la siguiente la variable observada (las ventas, en nuestro ejemplo) tiene un crecimiento o disminución considerable (del 15 por ciento o más, digamos). Nos avisa de oportunidades que hay que aprovechar, o de problemas que debemos resolver. También se le puede pedir que sólo nos avise de los aumentos que ocurren en tres periodos de tiempo consecutivos, o si estos aumentos ocurren en establecimientos geográficamente cercanos (lo que significa que la tendencia se observa en toda una zona).
- *Franja de normalidad*: como a menudo la variable que estamos observando tiene un comportamiento estacional (por ejemplo, en época de frío se vende menos helado que en la de calor), en vez de establecer cotas superiores e inferiores, podemos decirle al minero que nos informe cuando la variable de interés se salga de una "franja de normalidad" establecida, tomando en cuenta, digamos, cómo se comportó esa variable (ese fenómeno que estamos observando) durante el año pasado.
- *Comportamiento errático*: quizá nos interese que el minero nos informe de épocas (o de zonas del territorio, o de productos) en que el comportamiento no siga una tendencia definida, es decir, registre tumbos, suba o baje. En este caso, el minero comparará varios valores semanales consecutivos.
- *Máximos*: ¿qué productos se venden más?, ¿en qué temporadas se venden más productos de ferretería?, ¿en qué zonas se venden más desodorantes para hombre? Un minero que sistemáticamente barra las ventas y detecte máximos podrá contestar preguntas de este tipo. Igualmente sucede con los valores mínimos: algo que se venda poco, una carrera en un instituto que tenga pocos egresados, una enfermedad que ya casi no ocurre, etcétera.
- *Patrones frecuentes*: "cada vez que alguien compra leche, compra pan"; es una regla que, de ser cierta,



establece que (leche, pan) es un patrón frecuente. Para que un patrón sea frecuente, sus componentes deben serlo (si *pan* es un producto poco comprado, entonces no puede ser miembro de ningún par de productos frecuentes). Los patrones frecuentes deben tener un *soporte* (el porcentaje de comprobantes de compra del supermercado donde se compró leche y pan) mínimo, digamos 6 por ciento de los comprobantes. Podría ser que el patrón frecuente (leche, pan) fuera parte de otro *patrón frecuente* más extenso, digamos (leche, pan, arroz). Para determinar los patrones frecuentes, el minero comienza examinando todos los comprobantes para saber cuáles son los *ítems* (productos individuales) frecuentes. Como a menudo los datos a examinar son voluminosos, no caben en la memoria principal de la computadora, y es necesario que el minero maneje cuidadosamente los accesos (lecturas) al disco, para no desperdiciar tiempo. Una vez detectados los patrones frecuentes, es relativamente fácil detec-

tar los *pares* de patrones frecuentes, y de ellos ver cuáles son los *tríos* de patrones frecuentes, etcétera.

- *Reglas de asociación*: una vez determinado un patrón frecuente, por ejemplo (leche, pan, arroz), sería interesante para el minero descubrir cuál producto causa que los otros sean comprados. Por ejemplo, ¿quien compra leche, compra también pan y arroz? En este caso, leche → pan, arroz. Pero pudiera ser que quien compra arroz y leche compra también pan. En este caso, arroz, leche → pan. Éstas se llaman *reglas de asociación*, útiles para determinar causa y efecto. Para que una regla de asociación sea establecida como tal, se requiere que la regla rebase cierta *confianza mínima*. Por ejemplo, la *confianza* de la regla leche → pan, arroz es el porcentaje de los clientes que, habiendo comprado leche, efectivamente también compraron pan y arroz. Como hay muchas reglas posibles a ensayar, el minero tiene que efectuar esos ensayos en un orden cuidadosamente establecido, a fin de no desperdiciar tiempo de máquina.
- *Cúmulos* (clusters): dados todos los clientes de una cadena de establecimientos (o todos los pacientes de un conjunto de hospitales), usando *técnicas de agrupación* se pueden agrupar o clasificar a los clientes en, digamos, seis categorías o cúmulos, que nos representan a clientes con propiedades parecidas entre sí, pero distintas a los pertenecientes a otros cúmulos.



Hay otros métodos, omitidos aquí por brevedad. Así, usando la estadística, las bases de datos y la inteligencia artificial, los mineros van descubriendo automáticamente situaciones interesantes en un mar de datos. A diferencia de la estadística, que examina una muestra (una pequeña porción) de los datos para inferir características de todos los datos, el minero examina *todos* los datos. Éstos a menudo son muchos, por lo que, como hemos dicho, debe efectuar sus lecturas de disco y sus procedimientos en memoria con cierto orden, a fin de no desperdiciar tiempo.

El análisis mediante minería de datos se lleva a cabo con dos actividades para obtener *conocimiento no conocido*: a) describir en detalle a los generadores de datos, y b) predecir su comportamiento en su entorno; todo esto utilizando la historia almacenada en la bodega de datos.

La descripción en detalle se hace a partir de una revisión exhaustiva de toda la información disponible, revisión que también permite conocer a los generadores de datos en cada momento. Y conocer el comportamiento de los generadores puede ayudar a las personas que toman decisiones a identificar futuras situaciones deseadas o no deseadas, aun con datos faltantes, y poder indicar el valor de éstos con cierta certidumbre.

El conocimiento obtenido puede ayudar a los ejecutivos en objetivos como los siguientes:

- *Mejorar los servicios o productos que se ofrecen*. Esto es posible si se registra en la bodega el detalle de la respuesta a la compra por parte de los clientes al haber cambios en los productos o servicios, en cuanto a si se incrementa o se disminuye la venta. De estos resultados se puede aprender.
- *Evitar situaciones no deseadas*, como la de perder clientes en servicios contratados. Estas situaciones se pueden prevenir, ya que se tiene el historial de la facturación de un servicio contratado, como el teléfono, al igual que los clientes que tienen el antecedente de que se han quejado por el servicio, los periodos de tiempo en que su número de llamadas

decrece, y los que han cancelado su contrato en condiciones similares. También se debe tener datos de clientes que se han logrado retener y con qué estrategias se logró, al igual que el costo de cada estrategia. Se busca retener clientes, dado que es más barato mantenerlos que ganar nuevos clientes.

- *No manufacturar productos que en un futuro ya no se venderán*. Se pueden predecir cambios en los gustos de los consumidores, dado que con el historial de ventas se detectan las características de los productos que se dejan de vender.
- *Detectar productos de temporada*. Una tienda comercial vende sus productos y registra la fecha de venta. Al revisar sus ventas por largos periodos, puede saber con precisión el intervalo de fechas en que algunos de estos productos tienen un alto volumen de ventas, y con esta información tomar una serie de decisiones alrededor de este comportamiento: cuáles productos comprar y ofrecer, cuándo pedir los productos para tenerlos disponibles, qué cantidad solicitar y almacenar para esas ventas con el fin de no tener sobrantes, realizar la publicidad apropiada para su venta, y en qué lugares ofrecer los productos o servicios.
- *Conocer productos o servicios que se pueden vender en forma conjunta*. Al revisar el historial de las ventas se identificarán los productos que coinciden en su venta conjunta, y con las estadísticas se seleccionarán los conjuntos de productos que coinciden en alto porcentaje, definido por el usuario interesado.

IV. Ejemplos de mineros y sus aplicaciones

Conviene dar algunos ejemplos que nos ilustren para qué sirven y cómo pueden ayudar los mineros a la toma de decisiones estratégicas y a mediano plazo. Usaremos trabajos realizados en el Centro de Investigación en Computación.

Localizar tendencias de consumo a través del tiempo

Tomando como ejemplos a Pemex y la Comisión Federal de Electricidad (CFE), en estas empresas es importante saber cómo se realiza el consumo de derivados del petróleo o de energía eléctrica a través del tiempo en el país.

Para Pemex, en qué lugares se tiene un consumo similar de cierto derivado a través del tiempo, y así planear la distribución de este hidrocarburo.

Para CFE, saber esto le servirá para preparar la fuente generadora de energía con tiempo, generalmente con ayuda del agua de ríos o presas, dado que la energía hidroeléctrica es más barata que la generada por otros medios, como la termoeléctrica o la nuclear.

El Centro de Investigación en Computación (CIC) del IPN construyó una herramienta llamada Sistema de Minería de Datos, módulo de ANASIN (conjunto de herramientas para realizar análisis), que puede tomar como fuente los consumos del derivado de gasolina por centro de distribución, en qué periodos se realizaron, o los consumos de energía eléctrica por zonas, con mediciones mensuales a través de varios años para reconocer algunos patrones o tendencias de consumo de energía.

Con este sistema se puede seleccionar un patrón o tendencia (crecimiento, decrecimiento, constante o variada) con los valores de interés (consumos, en este ejemplo) a través de varios lapsos (días, semanas, meses, entre otros).

Los programas del módulo ANASIN revisan en forma exhaustiva el cubo de datos, como el de la Figu-

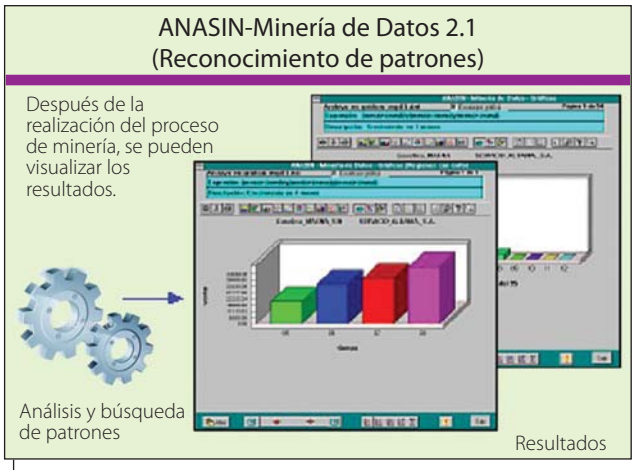


Figura 3. Presentación de patrones solicitados y localizados.

ra 1, y terminan su trabajo regresando ya sea un reporte o una serie de gráficas con los espacios de tiempo donde se cumple el tipo de tendencia buscado. Por ejemplo, las gasolineras con los periodos donde hay un crecimiento cuatrimestral continuo en su consumo del derivado (Figura 3). El conocimiento de las características de las áreas con el tipo consumo localizado las deduce el usuario (las del sur de la República, o las del norte, por ejemplo).

Localizar medidores de consumo de energía clasificados como malos medidores

Para la tarea de identificar o clasificar malos medidores de energía se construyó un conjunto de programas con el nombre de "clasificadores", también del módulo de ANASIN, que pueden tomar como fuente las mediciones de los consumos mensuales de energía para realizar las siguientes tres fases:

1. Con un conjunto de medidores de energía eléctrica y sus características (tipo, edad, número de hilos, tipo de negocio, cantidad de consumo, tipo de medición, entre otras), donde se indica quiénes realizan tanto una mala medición (ya sea en forma intencionada o no) como quienes realizan una buena medición, los programas aprenden a reconocer estas situaciones, regresando varios resultados; entre ellos, una estadística similar a la de la Figura 4. La mala medición posiblemente es un fraude en el consumo.

2. Después, con otro conjunto de medidores y sus características, donde algunos realizan una mala y otros una buena medición del consumo, pero sin indicar a los programas la clasificación de la medición (buena o mala), estos programas, tomando como referencia la fase 1, deben indicar qué medidores realizaban una buena o una mala medición. Según el número de aciertos, se podía calificar la eficiencia de estos programas. El resultado de la eficiencia depende del conjunto dado en la fase 1, así que se puede mejorar ésta si se cambia el conjunto, hasta que el usuario quede satisfecho.

3. Ya con otro conjunto de medidores, sin saber si éstos realizan una mala o buena medición del consumo, y también tomando como referencia la fase 1, los programas producen una estadística de cuántos medidores realizan una buena o una mala medición, además del conocimiento para identificar los medidores (como se ilustra en la Figura 4). Esta identificación puede tomarse como referencia para que los empleados de la empresa corroboren la situación de posible fraude en el consumo de energía, visitando la instalación del medidor. Tener una herramienta con un menor error que la creencia humana al visitar un medidor que pudiera estar realizando malas mediciones se refleja en una menor inversión de tiempo, dinero y personas asociadas a esta tarea.

Como imaginará el lector, la utilidad de esta actividad es disminuir el esfuerzo y tiempo para detectar y clasificar estas situaciones, además de usar un menor

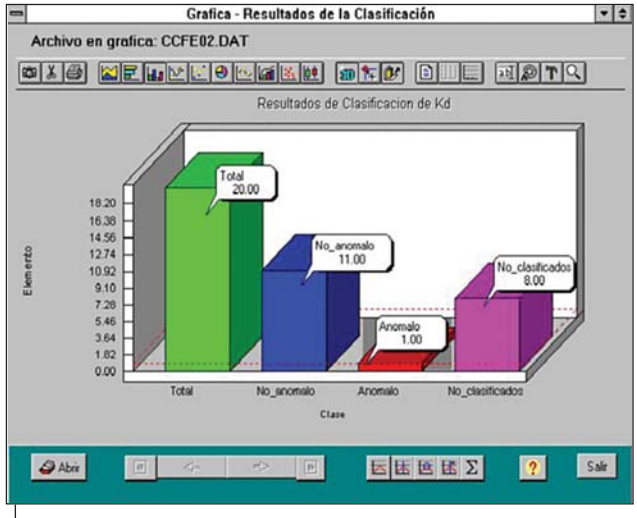


Figura 4. Resultados de clasificar un conjunto de objetos sin clases.

número de recursos físicos (personas, transporte y planeación de las visitas). Las decisiones de mantener o corregir esta situación dependían ya de la dueña de los datos.

Herramienta para localizar comportamientos complejos predefinidos

Otra herramienta construida es *Antecumem* (Análisis Temporal en Cubos de datos en Memoria), la cual permite localizar algunos análisis predefinidos en diferentes ambientes de datos. Los análisis predefinidos abarcan algunas de las consultas más frecuentes de operaciones en cubos de datos sin usar jerarquías; preguntas como "localizar los productos que más bajaron sus ventas en dos temporadas" (pregunta 4) o "localizar los productos de temporada en verano" (pregunta 5).

Aquí, el cubo de datos puede tener *n* dimensiones d_i, el valor numérico de interés con el agregado de sumar (ejemplo: ventas). Q es la consulta que define un subcubo, v_i es el valor en la *i*-ésima dimensión, R_i es un intervalo en la *i*-ésima dimensión, y S(C) es una suma de valores en el subcubo C.

1. *Pregunta puntual*: localizar el valor del hecho en valores por cada una de las d_i: $Q(v_1, v_2, \ldots, v_n)$.

2. *Pregunta sólo con rangos*: se tiene un subcubo de datos definido por rangos para cada una de las d_i, del cual se obtendrá una suma. $S(C)=Q(R_1, R_2, \ldots, R_n)$.

3. *Pregunta de eficiencia entre dos cubos*: calcula un porcentaje de incremento o decremento en dos subcubos de datos, $E=100*[(S(C_2) /S(C_1)) – 1]$.

4. *Pregunta de eficiencia grupal*: eficiencia de un conjunto de elementos de una dimensión entre dos subcubos de cada elemento, $E_i=100*[(S(C_{i,2}) /S(C_{i,1})) – 1]$, donde *i* son cada uno de los elementos de la dimensión de interés.

5. *Pregunta sobre conservación/pérdida*: permite localizar elementos en una dimensión entre dos subcubos que conservan o pierden una posición entre los mejores o peores; puede variarse el tiempo (para comparar periodos) u otra dimensión.

6. *Pregunta de temporalidad*: igual que la pregunta anterior, pero se trata de más de dos subcubos; si varían las unidades del tiempo, serán periodos de tiempo (días, semanas, meses, años…).

7. *Pregunta de búsqueda de tendencias en elementos de una dimensión*: localiza los elementos que tienen un comportamiento en un número de periodos o momentos continuos de tiempo.

A una pregunta de temporalidad como "se desea saber cuáles productos en el intervalo de [500-3000] fueron los mejores en el año de 1998 y se conservaron entre los 10 primeros en las ventas en el año de 1999 en todos los clientes y en todas las promociones", *Antecumem* responde indicando el tiempo que tardó, cuántos y qué productos se mantuvieron, y qué resultados numéricos contribuyeron a la respuesta.

Otra pregunta de tendencia como "se desea saber cuáles productos en el intervalo de [500-3000] fueron de los diez mejores durante tres meses consecutivos a partir de febrero en 1998, es decir, se conservaron entre los diez primeros en las ventas para todos los clientes y en todas las promociones". *Antecumem* responde nuevamente indicando el tiempo que tardó, y cuántos y qué productos se mantuvieron con la tendencia especificada. Por separado, se tendría que revisar los valores en esos lapsos para corroborar el resultado.

Al igual que las herramientas anteriores, la minería de datos realiza una revisión exhaustiva en los datos para hallar el conocimiento deseado, pero queda la tarea de que ésta sea validada por el usuario.

El futuro de estas herramientas está en tratar de facilitar los dos tipos de análisis de datos, pero agregando las técnicas del área de estudio conocida como "visualización de la información". Para mayor infor-

mación consultar *www.kdnuggets.com* y *http://conferences.computer.org/infovis/*. El lector puede consultar una amplia variedad de ejemplos de herramientas de minería de datos y de OLAP tanto comerciales como de acceso libre en la página *www.kdnuggets.com*

Gilberto Lorenzo Martínez Luna estudió en la Escuela Superior de Física y Matemáticas (ESFM) del Instituto Politécnico Nacional (IPN). Obtuvo la maestría en el Centro de Investigación y Estudios Avanzados (Cinvestav) y el doctorado en computación en el Centro de Investigación en Computación (CIC), ambos del IPN. Profesionalmente ha desarrollado sistemas de información desde 1982, tanto en el sector público como en el privado. Desde 1987 ha impartido cursos en licenciatura, maestría y doctorado en diversas escuelas y universidades. Ha participado en la organización de eventos de minería de datos en el IPN, desde 1998 hasta 2005. Actualmente es jefe de laboratorio, imparte cursos, dirige tesis de posgrado, y hace investigación sobre bases de datos y minería de datos.

llunameztli@gmail.com

Bibliografía

Chaomei, Chen (2006), *Visualization information, beyond the horizon*, Londres, Springer.

Chen, Z. (2001), *Intelligent data warehousing*, Boca Raton, CRC Press.

David J. Hand, Heikki Mannila y Padhraic Smyth (2001), *Principles of data mining*, Cambridge, Massachusetts, MIT Press.

Fayyad, U. M. y G. Piatetsky-Shapiro (1996), *Advances in knowledge discovery and data mining*, Menlo Park, California, AAAI Press.

Jiawei, Han y Micheline Kamber (2006), *Data mining: concepts and techniques*, 2ª ed., edición de Jim Gray, San Francisco, California, Morgan Kaufmann Publishers (The Morgan Kaufmann series in Data Management Systems).

Pang-Ning, Tan, Michael Steinbach y Vipin Kumar (2006), *Introduction to data mining*, Addison-Wesley.

Witten, Ian H., Frank Kaufmann y Morgan Kaufmann (2005), *Data mining: practical machine learning tools and techniques*, 2ª ed., edición de Jim Gray, San Francisco, California, Morgan Kaufmann Publishers (The Morgan Kaufmann series in Data Management Systems).