INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

No. 3 **Serie:** VERDE **Fecha:** Marzo 98

Estado del Arte y de la Práctica en Minería de Datos, Análisis y Crítica

Dr. Adolfo Guzmán Arenas♠

RESUMEN

Los mineros de datos son programas que de manera automática y sin intervención humana encuentran similaridades, situaciones interesantes y desviaciones en una base de datos. Se presenta una taxonomía de los sistemas de minería de datos, procurando indicar sus similitudes, diferencias, usos y técnicas de procesamiento. También se exponen : (a) los principales problemas a resolver, y (b) un sistema específico construido por el autor.

Palabras clave: Minería de datos, cubo de valores, clasificadores supervisados, casaderos de expresiones, información ejecutiva, base de datos relacionales.

♠Director del Centro de Investigación en Computación - I.P.N.

ADVERTENCIA

"Este reporte contiene información desarrollada por el Centro de Investigación en Computación del Instituto Politécnico Nacional a partir de datos y documentos con derechos de propiedad y por lo tanto su uso queda restringido a las aplicaciones que explícitamente se convenga.

La aplicación no convenida exime al Centro de su responsabilidad técnica y da lugar a las consecuencias legales que para tal efecto se determinen.

Información adicional sobre este reporte podrá obtenerse recurriendo a la Unidad de Publicaciones y Reportes Técnicos del centro de Investigación en Computación del I.P.N. Av. Juan de Dios Bátiz s/n, teléfono 729-60-00 ext. 56500, 56608 y 56610".

Estado del Arte y de la Práctica en Minería de Datos, Análisis y Crítica

ADOLFO GUZMÁN ARENAS

1. ANTECEDENTES Y OBJETIVOS

1.1 Antecedentes

El abaratamiento del disco de cabeza móvil y de las comunicaciones hace posible que ahora existan colecciones sistemáticas de datos periódicos de la actividad de la empresa u organización, ordenados por ejemplo por producto ; por expendio o centro de trabajo ; por día o fecha de realización ; y por vendedor, operario o responsable. Este mar de datos generalmente reposa en una base de datos relacional, y se explota mediante herramientas de generación de reportes, visualización y graficación (incluyendo sistemas de información ejecutiva), con las que los gerentes o responsables toman decisiones. Es decir, *la información se emplea para su análisis manual o visual* por parte de expertos.

Los inconvenientes de este análisis manual son las limitaciones de tiempo y procesamiento de parte de las personas involucradas, además de que no fácilmente se detectan anomalías o desviaciones pequeñas. Conviene complementar este análisis con otro automático, hecho por programas que de manera sistemática, y fuera de las horas pico, escudriñan la base de datos en busca de relaciones interesantes, mismas que al ser halladas se mostrarán al usuario o, más comúnmente, se guardarán en un archivo para su revisión manual posterior.

1.1 OBJETIVO

Generalmente se acepta el siguiente: Un sistema de minería de datos busca situaciones interesantes, desviaciones, tendencias y anomalías, en un mar de datos.

Por lo general la búsqueda es automática (la máquina efectúa los hallazgos sin intervención humana) sobre datos numéricos que yacen en una base de datos relacional. El siguiente capítulo muestra variaciones; con ✓ se señalan las variantes principales (aceptadas generalmente como "típicas" o propias de un minero), con ✗ aquellas que generalmente no se aceptan como típicas o propias, con 🕮 las variantes que constituyen retos técnicos, de construcción o campos de investigación hoy, y con ⇨ aquellas que forman parte de "MineDatos".[1]

2. TAXONOMÍA DE SISTEMAS DE MINEROS

Podemos clasificar a los sistemas de minería de datos de varias maneras, las más útiles son según el tipo de datos a procesar (§2.1), según la forma en que están guardados (§2.1.1), según el criterio de qué constituye una región interesante (§2.2), y según las técnicas computacionales usadas para la búsqueda (§2.3).

2.1 Tipos de datos

A) ✓⇨Datos numéricos. Los datos sobre los que trabaja un minero son generalmente números (por ejemplo, ventas en pesos), organizados a lo largo de ejes que pueden ser numéricos (fechas) o simbólicos (tipo de producto que se vendió; zona geográfica donde la venta se originó). Aún cuando los datos primarios no sean números (accidentes, enfermedades) se procura llevarlos a forma numérica (número de accidentes, número de enfermos), por la mayor disponibilidad de herramientas numéricas para su análisis.

B) ✗ Bloques de datos en binario. Por ejemplo, imágenes, sonidos, y otros objetos representados por bloques de información binaria. Un problema de investigación (🕮) es el de establecer predicados y funciones de similaridad para estos datos. Por ejemplo, en imágenes de rostros, nos interesa poder discriminar, digamos, entre hombres y mujeres, lo que no es trivial y requiere de información del dominio de aplicación.

[1] Construido por el autor en SoftwarePro International.

C) ✗ Objetos. Se puede hacer minería sobre objetos (en el sentido de C++), pero el problema se complica porque los objetos tienen, además de datos como su peso o color, *funciones* o métodos escritos en un lenguaje de programación.

D) ✗ 🕮 Datos simbólicos. La problemática es similar a 2.1.C, aunque aquí se disponen de mejores herramientas para analizar las propiedades numéricas y simbólicas de estos objetos. [ref. Objetos simbólicos] [ref. Libro Vol. I]

E1) Valores simbólicos, que no tienen significado numérico, como "protestante", "católico", "musulmán", para la variable "religión".

E2) Valores cualitativos, que admiten un orden de precedencia, como "muy frío", "frío", "tibio", "caliente", "muy caliente", para la variable "temperatura".

Hasta donde el autor sabe, no existen mineros sobre datos simbólicos, siendo esta un área importante de investigación (🕮).

F) ✗ 🕮 Texto. El hallazgo de situaciones interesantes en documentos textuales e información estructurada en secciones, párrafos y oraciones en español u otro lenguaje natural, requiere de funciones de similitud que de alguna manera "entiendan" lo escrito. Un enfoque inicial pero un tanto superficial es contar las palabras clave o los *conceptos* [ref. Clasitex] y así poder comparar dos documentos, por ejemplo, según el número de conceptos comunes o similares (se puede medir la similaridad entre dos conceptos si éstos se organizan en un árbol o taxonomía [ref. Clasitex]).

2.1.1 Forma de almacenamiento

Los datos anteriores pueden estar guardados en memoria secundaria en alguna de las siguientes formas.

A) ✓⇨Base de datos relacional. Es práctica común almacenar grandes volúmenes de datos en una base de datos relacional, donde los ejes o llaves son atributos importantes de esos datos, por ejemplo, la fecha, el lugar, el tipo de producto vendido, etc. Una ventaja importante de una base de datos relacional es poder usar su lenguaje estándar de consulta (SQL).

B) Archivos. Existen colecciones de datos guardados en archivos (no en tablas de bases de datos), con índices o llaves similares a las de §2.1.1.A. Trabajar sobre archivos es más pesado o primitivo, ya que no se disponen de las funciones propias de las bases de datos. Se puede mitigar esto en algo al usar *extractores de información* [ref. Anasin] y/o técnicas como el *descriptor de archivos* [ref. Descriptor de archivos] para poder accesar archivos de formato arbitrario pero conocido.

C) Base de objetos. Una base de objetos como [Ontologics] permite guardar en memoria secundaria los objetos. Se tienen las ventajas de §2.1.1.A y las desventajas de §2.1.C.

2.1.2 Organización de los valores

Es conveniente organizar los valores de las variables en una jerarquía o árbol, por ejemplo, la fecha en año-mes-semana-día, la localidad en país-estado-municipio-ciudad, etc. Así es posible agrupar los datos en "totalizadores" más útiles, por ejemplo, "total de ventas en el Estado de Guanajuato en el mes de septiembre".

Supongamos que tenemos una tabla de ventas organizada como sigue :

Ciudad	Fecha	Producto	Venta
Salina Cruz	20 marzo 1997	Pan Bimbo blancogrande	$432.

Es conveniente en el ejemplo organizar los datos de acuerdo a tres ejes : el eje temporal, en años, meses, semanas y días ; el eje geográfico ; y el eje de productos. Se ha formado un cubo tridimensional, denominado *cubo de*

valores, con dimensiones (fecha, localidad, producto), cuyas celdas son totalizadores. Por ejemplo, la celda (20 marzo 1997, Salina Cruz, Pan Bimbo blanco grande) contiene el valor 432.

Nótese que en este cubo habrá otras celdas o totalizadores que contendrán valores agregados (resúmenes o subtotales de ventas), por ejemplo, la celda (Marzo 1997, Salina Cruz, Pan Bimbo blanco grande), la celda (Marzo 1997, Oaxaca, Pan Bimbo blanco grande), la celda (Marzo 1997, Oaxaca, Panes y pasteles). Cuando se obtienen los datos de ventas (cuando el cubo se llena), la mayoría de sus celdas están vacías, pues solo contienen datos aquéllas con las coordenadas más pormenorizadas[2] (ya que se registran las ventas diarias de marzo, pero las ventas de marzo habrá que calcularlas mediante sumas), en nuestro ejemplo, son las celdas del tipo (fecha es día, localidad es ciudad, producto es detalle), por ejemplo, (20 marzo 1997, Salina Cruz, Pan Bimbo blanco grande). Este cubo es ralo, pero conforme los mineros trabajen, se irán calculando los totalizadores que ahora yacen vacíos. ¿Conviene realmente formar el cubo ? ¿Conviene llenarlo ? Existen tres posibilidades:

1) ✓⇨ Cubo real. Sí, fórmese y llénese el cubo. Este camino siguió la primera versión de MineDatos. Muchos productos comerciales exigen que se forme una base de datos (el cubo) específica para los mineros, facilitando la labor de los mineros (ventaja), pero duplicando la información ya existente en la base de datos real (desventaja). [Las empresas que tienen bases de datos regionales o corporativas generalmente las tienen en forma de tablas con índices y llaves foráneas, formando tablas que no se parecen al cubo de datos].

2) Cubo inexistente. No forme el cubo. El minero tendrá que accesar las tablas reales de la base de datos ya existente. Para esto, hay que transformar las expresiones (en SQL, típicamente) que accesan el cubo (inexistente), a expresiones equivalentes que accesen los mismos datos o regiones deseadas, pero en la base de datos real. El minero todavía "tiene la sensación" de accesar el cubo, mas sin saber, lo que accesa es la base de datos real. Este camino lo sigue Hugo de la Rosa [ref. Hugo de la Rosa] en su trabajo de tesis doctoral (🕮), que apenas inicia. La ventaja es que no se duplica la información (conservando espacio en disco). Pronto veremos (3.B.) que de todas maneras conviene conservar aunque sea algunos de los totalizadores calculados.

3) ⇨ Forme incrementalmente el cubo. Este camino siguió la segunda versión de MineDatos. Para cada celda del cubo se especifica una de estas tres variantes :

3.A) No se llene este totalizador. No se guarde en el cubo su valor, cuando éste se calcule. Si se vuelve a necesitar este valor, habrá que recalcularlo.

3.B) Llenado de totalizadores por adelantado. Motivados por los datos («data driven», en inglés). Llénese esta celda del cubo tan pronto como sea posible, tan pronto como los datos de las celdas hijas estén disponibles.

3.C) Llenado de totalizadores bajo demanda. Motivados por la demanda. («demand driven»). Llénese esta celda en el cubo solo si alguna vez se demanda o requiere su valor.

Nótese que algunas celdas nunca se llenarán, otras lo harán solo si se requieren (3.C), y otras en cuanto se pueda (3.B); el cubo tiene los tres tipos de celdas. De esta manera el usuario (o el administrador del cubo) puede cambiar espacio en disco por tiempo de cómputo, según se requiera.

2.2 Criterios de "interesante": qué buscar

Esta sección se refiere a las maneras en que se indica qué buscar, o cómo definir algo "interesantes" digno de ser reportado por el minero.

A) ⇨Formas específicas sobre una variable (sobre un flujo de datos). A menudo la irregularidad o "región de interés" puede tener una forma conocida de antemano. Ejemplos: (1) un ascenso continuo en los valores de la variable; (2) ídem. descenso; (3) un mínimo en una sucesión de valores; (4) ídem. un máximo; (5) una banda (límite superior e inferior), afuera de la cual es adecuado reportar una anormalidad o región de interés; (6) un

[2] Denominaremos *celdas básicas* a aquéllas cuyas coordenadas están todas lo más pormenorizadas posible, o sea, que ninguna coordenada puede descomponerse en otras más detalladas.

salto importante en el valor medio de la variable. En estos casos el problema se reduce a un problema de *ajuste de curvas*. El minero busca regiones donde haya un ascenso continuo, un mínimo, etc.

En MineDatos se sigue este enfoque, y existen mineros predefinidos que buscan las curvas del tipo 2.2.A.1 a 2.2.A.6. Estas curvas están descritas en un cierto lenguaje (ver 2.2.B), por lo que es posible definir otros tipos de curvas para su búsqueda automática.

B) ⟐Fórmulas y predicados definidos por el usuario. El usuario puede indicar al minero (en un cierto lenguaje (al que denominaremos L_u) para describir expresiones numéricas) el predicado que una región deba cumplir para ser reportada como interesante. Este enfoque se utiliza en MineDatos. Definir adecuadamente estos lenguajes es actualmente un campo fértil de investigación (🕮). Un problema es que el usuario suele expresarse en forma informal ("todas las tiendas de Guanajuato donde haya ..."), que hay que traducir a la estructura "apta para la búsqueda" del cubo (Guanajuato** en L_u significa "los nietos geográficos de Guanajuato", o sea, las tiendas de Silao, Pénjamo, León, etc.), que hay que traducir a una expresión en SQL que accese al cubo, misma que hay que traducir a otra expresión equivalente en SQL que accese la base de datos real (según §2.1.2.2). En la primer versión de MineDatos se inventó un L_u que se interpretaba por un evaluador de formas postfijas, mismo que accesaba el cubo real.

C) Parecidos o similaridades entre partes de un mismo flujo. En vez de señalar anticipadamente (como en §2.2.A) la forma a buscar, podemos buscar dos secciones de un flujo de datos que se parezcan entre sí, por ejemplo, utilizando la función de autocorrelación. Un área de investigación (🕮) es disminuir la gran cantidad de cómputo que se requiere para comparar pares de trozos de un flujo de datos en búsqueda de similaridad. Una técnica usada es reemplazar un flujo de datos numérico por expresiones simbólicas de tipo (AAEED) que significa (ascendente, ascendente, estacionario, estacionario, descendente), y trabajar con las cadenas simbólicas, mucho más cortas. Nótese que AAEED bien puede señalar un máximo en el flujo.

D) 🕮 Relaciones entre varias variables. Se pueden buscar relaciones entre dos o más flujos de datos, por ejemplo, hallar trozos de los flujos a, b y c en donde se cumpla que a = b + c, o donde se cumpla que b y c suben mientras a baja. El problema aquí no es predefinir la fórmula (a = b + c, por ejemplo) utilizando el lenguaje de §2.2.B, y buscar trozos de flujo donde tal fórmula se cumpla. El problema es más complejo, pues se trata de *determinar la fórmula* (que de alguna manera resulte interesante) y los trozos de flujo donde se cumpla. Un área de investigación es disminuir la gran cantidad de cómputo requerida.

2.3 Técnicas de búsqueda, herramientas usadas

Esta sección muestra las diferentes técnicas o modelos que los sistemas emplean para llevar a cabo el hallazgo de situaciones interesantes.

A) ⟐Ajuste de curvas. Los mineros del §2.2.A y §2.2.B pueden realizar su cometido utilizando un ajuste de curvas. La curva a ajustarse, ya sea predefinida (§2.2.A) o definida por el usuario (§2.2.B) se va comparando contra el flujo de datos que se analiza; un ajuste apretado (bueno) motivará que se señale un hallazgo interesante : se ha encontrado una región o trozo de interés.

B) Correlación, métodos estadísticos.

1. Se utiliza para comparar dos trozos de flujo (§2.2.C).

1. También para buscar fórmulas entre varias variables (§2.2.D), como sigue: Supongamos que tratamos de verificar si la fórmula a = b + c se cumple en un cierto trozo. Entonces, calcúlese el nuevo flujo b + c y véase si se correlaciona con el flujo a. [En este caso podríamos más eficientemente calcular el nuevo flujo b + c - a y ver si es casi siempre cercano a 0]. A menudo las fórmulas tienen un retraso en el tiempo: $a_{i+10} = b + c$, es decir, *a* dentro de diez días será igual a *b* + *c* de hoy.

B) Casaderos de expresiones, comparadores. Útiles para buscar relaciones entre varias variables (§2.2.D), pues comparan cadenas de símbolos o caracteres. Ejemplo de uso: supongamos que estamos viendo en cuál mes, en la tienda "Cantarranas" de Salina Cruz, en el producto Leche Nestlé condensada de medio litro, las ventas aumentaron, las devoluciones disminuyeron y los saldos de crédito (cantidades fiadas) también aumentaron (esto

puede originarse por una promoción muy efectiva del producto lácteo). El flujo de ventas pudo ser 123_V 135_V 147_V 193_V 186_V 93_V y similarmente para devoluciones y crédito. Reemplácense los flujos numéricos por otros simbólicos que pudieran ser, por ejemplo, $A_VA_VA_VE_VD_V$ para ventas, $D_DD_DE_DE_DD_D$ para devoluciones, y $A_CE_CE_CE_CE_C$ para crédito, y úsese un casador o aparejador de expresiones o patrones [ref. Convert] para hallar el patrón $A_VD_DA_C$, que literalmente dice : "aumentaron las ventas, disminuyeron las devoluciones y aumentaron los créditos".

C) Redes neuronales. Por medio del ajuste de pesos de una red neuronal, es posible aprender algunas propiedades sencillas de los datos.

D) Algoritmos genéticos. Mediante técnicas de intercambio de cromosomas, es posible entender las diferentes fórmulas entre los datos, como en §2.2.C.

E) ✗ Clasificadores supervisados. Colecciónense varios objetos en una matriz de aprendizaje MA, donde se indica la clase a la que cada uno pertenece. Úsese esta matriz para enseñar al algoritmo a identificar las clases. Una vez que el algoritmo "aprende", úsese éste sobre una matriz de control MC, que contiene objetos pertenecientes a clases conocidas, pero desconocidas al algoritmo que está aprendiendo. Con MC determínese el porcentaje de aciertos. Si éste es adecuado (grande), úsese ya el algoritmo para clasificar objetos desconocidos. Los clasificadores supervisados no son propiamente mineros, puesto que no buscan "situaciones interesantes". Como el tema de minería de datos está de moda, algunos productos que son clasificadores ahora toman el nuevo nombre.

F) ✗ Clasificadores no supervisados. Estos algoritmos agrupan un conjunto grande de objetos en "nubes" (subgrupos, llamados clases) tales que entre miembros de cada clase hay cierta afinidad o parecido. Parten un conjunto en clases "naturales" ; hallan tales clases. Los clasificadores no supervisados no son propiamente mineros.

G) ✗ Visualización, búsqueda manual. Aquí se le pide al usuario que él (o ella) identifique las diferentes situaciones interesantes o desviaciones, utilizando gráficas, reportes tabulares, diagramas de pastel, etc. El hallazgo de los hechos interesantes lo hace una persona. Pienso que no es típico de un sistema de minería de datos, en donde todo es automático y ocurre sin intervención humana.

H) ✗ Sistemas interactivos, intervención humana. Un híbrido entre mineros automáticos y visualizadores (§2.3.H), estos algoritmos buscan hechos interesantes, pero ocurren con frecuencia al usuario, ya para mostrar lo encontrado, ya para que se corrijan los parámetros de operación. Al estar involucrado el usuario, le dedica un tiempo considerable a la minería (desventaja).

I) ✗ Proceso de aplicaciones en línea ("olap" en inglés). Cuando se está modificando en línea un registro, generalmente la única aplicación que se ejecuta es una de validación de que los datos nuevos estén dentro de lo normal. Al disponerse de mayores velocidades de procesamiento, resulta útil ejecutar alguna aplicación más compleja conforme los datos se están introduciendo, es decir, la aplicación se procesa en línea. Las aplicaciones en línea no son mineros, aunque algunos mineros pueden trabajar en línea (ser una aplicación que se procesa en línea).

3. TÉCNICAS PARA MANEJAR GRANDES VOLÚMENES; PROCESO INCREMENTAL

Muchos algoritmos consumen demasiado tiempo cuando procesan grandes volúmenes de datos. Por ejemplo, un clasificador no supervisado (§2.3.G) que clasifica un millón de clientes. En estos casos, resulta conveniente hacer una versión *incremental* del algoritmo, de manera que, por ejemplo, cuando lleguen 50,000 clientes nuevos, no haya que procesar de nuevo 1,050,000 clientes.

Estos algoritmos emplean algunas de las siguientes modalidades, como por ejemplo :

• Guárdense los parámetros que se usaron durante el cálculo inicial de la función (el centro de gravedad, por ejemplo), de manera que se pueda recalcular cuando llegan nuevos objetos. Por ejemplo, al calcular el promedio

de n números, guárdese no solo tal promedio, sino también "n". De esta suerte, cuando lleguen dos números más, recalcúlese el promedio como

$$p_n = (p_v * n + num_1 + num_2)/ n+2.$$

4. ESTRUCTURA DE UN MINERO: MINEDATOS

Esta sección describe cómo está formado y cómo funciona MineDatos, un minero diseñado y construido por el autor.

Las partes del minero son las siguientes :

1. Selección de variables. Las variables más importantes se escogen de acuerdo al especialista, y se incluyen en el cubo de datos del punto 2.
2. Cubo de datos. Se linearizan los árboles de cada variable del punto 1. Por ejemplo, la jerarquía de fechas (año, mes, semana, día) se coloca sobre un eje. Lo mismo para la jerarquía geográfica, para la jerarquía de productos, o de enfermedades, etc. Es un hipercubo, pues en general tiene más de tres dimensiones. En cada celda se coloca el totalizador respectivo. Solo los totalizadores más pormenorizados tienen datos inicialmente (es decir, existe un dato de venta para Zapatos Bostonianos cafés talla 7 en Salina Cruz el 18 de septiembre de 1997, pero no para esa semana, ni para Oaxaca). Cada celda tiene padres geográficos, de producto, temporales, etc. El cubo de MineDatos es *real*, es decir, todas sus celdas existen. Primero, existen vacías en su mayoría, y después, son llenadas por adelantado. En una segunda versión, las celdas son llenadas bajo demanda (solo si se necesitan).
3. Se definen los mineros como clientes del servidor de datos, que está en Informix. Los clientes son programas en C que evalúan una expresión (predicado) que regresa V ó F sobre cada celda. Los mineros efectúan un barrido sobre las celdas ; aquéllas que producen V originarán una "situación interesante" que será registrada en un archivo de salida. La expresión a evaluar está en forma normal posfija, y *no* se convierte a SQL, sino que se interpreta. Es decir, el minero analiza su predicado y extrae del cubo real un sub-cubo de datos, trayéndolo a la memoria de la máquina cliente para su procesado. La mayor parte del tiempo los mineros están haciendo sumas para calcular los totalizadores que no existen en el cubo. Esta arquitectura aprovecha el poderío de las varias máquinas conectadas al servidor ; éste se utiliza solo para repartir bloques de datos a las máquinas que son las que hacen el trabajo pesado. (Muchos mineros comerciales efectúan este trabajo en el servidor, sobrecargándolo).
4. Cada minero tiene una agenda de trabajos pendientes (regiones a escudriñar). Cuando los trabajos prioritarios aumentan, el minero cesa de trabajar, reanudando sus labores posteriormente.
5. Se les puede cambiar a los mineros :
   a) Las fórmulas de los predicados. Inicialmente, buscan por máximos, mínimos, pendientes positivas o negativas "considerables", y otras cuantas "curvas" interesantes. En este sentido, los mineros hacen ajuste de curvas.
   b) La región de búsqueda. Por ejemplo, restringirlo solo a la zona sur del país.
   c) El orden de búsqueda. Por ejemplo, mirar primero tal o cual producto o enfermedad.
   d) Interacción diferida con el usuario. Si el usuario manualmente nota que una celda está cerca de ser interesante, puede fabricar un minero ad hoc que la vigilará, avisándole cuando sea interesante.
6. La interacción que un sistema de mineros brinde al usuario debe ser lo más amigable posible, pues él normalmente no tiene los conocimientos requeridos de estadística, informática, de cómo está el cubo, etc.

CONCLUSIONES

El campo de la minería de datos es nuevo y ha despegado debido al abaratamiento de la capacidad de cómputo de las computadoras personales y de la de almacenamiento del disco de cabeza móvil. En general, es un campo donde el estado de la práctica (programas y sistemas comerciales) está dominado por consideraciones de eficiencia y utilidad, en tanto que el estado del arte (publicaciones e investigación) está desvinculado de los practicantes, por no saber quiénes son. Por consiguiente, es un campo muy fructífero para instituciones o grupos donde ambos aspectos se combinen o complementen.

Algunos desarrollos ya conocidos, como clasificadores, visualización, e incluso selección de variables, son ofrecidos como parte de un sistema de minería de datos, aunque nosotros reservamos este nombre para la búsqueda automática de situaciones interesantes, desviaciones y anomalías.

Las aplicaciones que se procesan en línea («olap» en inglés) no son necesariamente mineros, aunque algunos mineros son aplicaciones en línea.

REFERENCIAS

[Clasitex] A. Guzmán. Reporte en preparación. SoftwarePro International, Austin, Texas, 1997.

[Objetos simbólicos]. José Ruiz Shulcloper. Reporte en preparación. Centro de Investigación en Computación (C. I. C.), Instituto Politécnico Nacional. 1997

[Ontologics] Base de objetos. Producto comercializado por Ontologics.

[Anasin] Producto comercializado por SoftwarePro International, Austin, Texas. 1995. Contiene extractores de información, sumarizador-transmisor, clasificadores, mineros.

[Descriptor de archivos]. A. Guzmán. Comunicación técnica de la Sección de Computación, CINVESTAV-IPN, 1983.

[MineDatos] A. Guzmán. Mineros de datos, en Anasin.

[Hugo de la Rosa]. Tesis doctoral. Trabajo en progreso. Centro de Investigación en Computación (C. I. C.), Instituto Politécnico Nacional. 1997