Minería de Datos con Búsqueda de Patrones de Comportamiento

M.en C. Gilberto Lorenzo Martínez Luna,
Centro de Investigación en Computación (CIC)
Instituto Politécnico Nacional (IPN)
lluna@pollux.cic.ipn.mx

Dr. Adolfo Guzmán Arenas
Centro de Investigación en Computación (CIC)
Instituto Politécnico Nacional (IPN)
aguzman@pollux.cic.ipn.mx

Resumen.
En este trabajo se presenta una forma de realizar el descubrimiento de conocimiento o Minería de Datos a partir de una base de datos, la técnica utilizada es la generalización y sumarización de datos en cubos de datos. Se utiliza una herramienta construida en el CIC que permite definir y utilizar los cubos, elegir las regiones de interés de estudio y definir los patrones de comportamiento o situaciones anómalas a localizar en estas regiones. Nuestra herramientas permite programar los procesos de extracción y análisis de datos en horarios nocturnos para aprovechar los recursos computacionales. En la presentación de los resultados de las búsquedas, se busca que esta sea sencilla de revisar e interpretar para descubrir las tendencias o relaciones entre los datos, y así generar conocimiento validado. En este artículo se describe este desarrollo como una implantación a la tecnología de análisis automático.

1. Introducción

Una definición de Minería de Datos es "*el descubrimiento eficiente de información valiosa, no-obvia de una gran colección de datos*" [1], cuyo objetivo "es *ayudar a buscar situaciones interesantes con los criterios correctos, complementar una labor que hasta ahora se ha considerado "intelectual" y de alto nivel, privativa de los gerentes, planificadores y administradores. Además, de realizar la búsqueda fuera de horas pico, usando tiempos de máquina excedentes*" [4]. En general, el proceso de minería se puede ver en la figura 1.

La utilidad de la Minería de Datos ya no se pone a discusión [1][5], por lo cual está tecnología esta siendo aplicada por muchas herramientas de *software*. Las técnicas de aplicación varían de acuerdo a la herramienta, algunas la instrumentan haciendo uso de redes neuronales (*SPSS Neural Connection*), otras con generación de reglas (*Data Logic*) o Arboles de Decisión [*XpertRule Profiler*]. En [6] puede verse una clasificación de las herramientas para desarrollar minería, de acuerdo a su técnica de aplicación.

En el Laboratorio de Sistemas de Información del CIC-IPN, se desarrolla una herramienta que forma parte del proyecto ANASIN[1], con la cual la Minería de Datos se realiza utilizando la técnica que construye cubos de n-dimensiones, conocida como *generalización y sumarización en cubos de datos* [5], técnica implantanda en una base de datos relacional. La generalización de los datos se puede desarrollar en los niveles que se considere necesario usar y así realizar análisis a diferentes niveles de conceptos. En los cubos formados, la herramienta permite definir *regiones* de interés en las cuales se buscan *patrones de comportamiento* [3], al término de la ejecución de las búsquedas los resultados se muestran en reportes de tipo texto y gráficas.

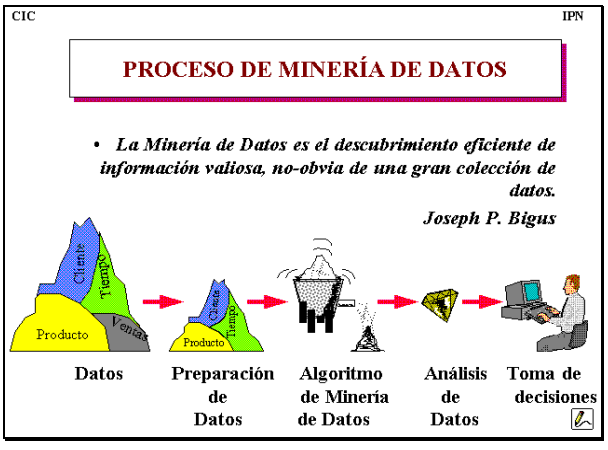


Figura 1

En esta parte de la ejecución del proceso de minería se pueden distinguir dos tipos de programas, los que extraen la región de interés de la base de minería, llamados *extractores*; y los programas que realizan la búsqueda de

[1] ANASIN se trabaja en el Laboratorio de Sistemas de Información y Bases de Datos del C.I.C., donde la actualización es apoyada por el CIC-IPN, *SOFTWARE PRO INTERNATIONAL* e I.D.A.S.A. Los módulos que componen el proyecto son: Bitácoras, Tabla Maestra, Generador de Formatos, Instalador Automático de Sistemas, Recolector de Datos, Interrogador, Reporteador, Graficador, Despliegue Geográfico, Minería de Datos, Clasificador de Entidades y Arbol Semántico de Conceptos.

patrones, que se les llama *mineros* [4]. Tanto la actividad de extracción como la de búsqueda de patrones generalmente pueden consumir demasiado tiempo por la gran cantidad de datos para formar las regiones y los numerosos calculos a realizar, por lo que estas actividades se delegan a programas que las realizan en forma autónoma y nocturna y así lograr aprovechar los recursos computacionales.

El presente documento esta organizado cómo sigue. En la sección 2 se describe en detalle el proceso de minería de datos como lo desarrolla la herramienta construida; en la sección 3 se describe algunos problemas por resolver y hacer más completa la herramienta; en la sección 4 se indican algunas líneas de investigación que se estan trabajando; en la sección 5, algunas conclusiones, al avance de nuestro trabajo; para terminar con características del software .

2. Descripción del Proceso de Minería

La herramienta desarrollada, se llama *Módulo Minería de Datos – ANASIN* y tiene el modelo de trabajo *Cliente/Servidor*, donde, se distinguen 4 actividaes principales:

- Definición del cubo y configuración de los niveles de búsqueda.
- Realizar solicitudes de minería, en una estación de trabajo o cliente
- El proceso príncipal de minería; generación de la región y búsqueda de un patrón determinado, en el Servidor
- La visualización de resultados de la búsqueda en el Cliente.

Para realizar la minería con el módulo, se siguen los pasos:

- Definir el cubo de datos o espacio de búsqueda de mineros
- Generalización o definición de los niveles de búsqueda en cada una de las dimensiones del cubo
- Generar los datos y cargar el cubo de datos
- Definir los horarios de trabajo de los procesos de minería
- Generar las preguntas (definir región y patrón a buscar)
- Solicitar ejecución del proceso de extracción y análisis.
- Ejecución de la extracción de la región solicitada y la búsqueda del patrón
- Revisar e Interpretar los resultados

Una descripción de los pasos que debe realizar un usuario se amplia a continuación.

2.1 Definición de cubo de datos.

Un usuario selecciona las variables de sus bases de datos de sus sistemas operacionales, en las que desea buscar patrones o comportamientos de interés. Según el número de variables seleccionadas es la dimensión del cubo formado, a cada variable se le denomina una *dimensión* o un eje del cubo (ver figura 2 y Figura 3).

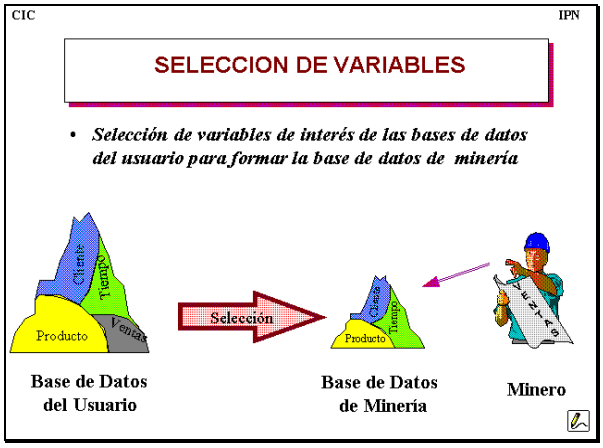


Figura 2

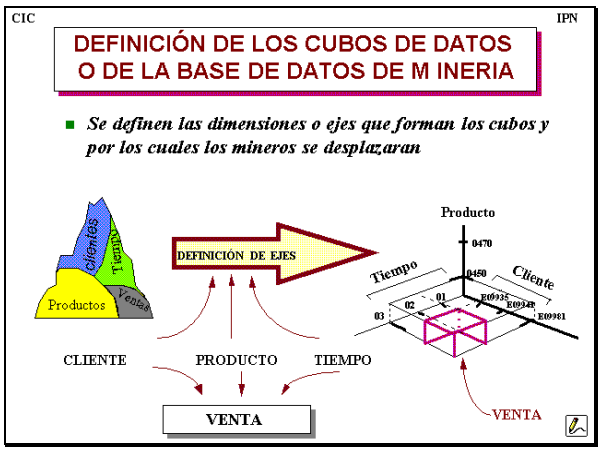


Figura 3

Ejemplo. Generalmente se manejan una gran cantidad de variables en una base de datos, pero solo pueden interesar tres variables cuya intersección es el valor de interés a analizar, esto da como resultado un cubo de datos con tres dimensiones, más una dimensión que puede contener los valores en los que se realizará la búsqueda del patrón. Un usuario puede elegir la relación *venta(producto, cliente, tiempo)*, definir como primer eje al *producto*, como segundo eje al *cliente* y como tercer eje al *tiempo*. La

intersección es la *venta* de un *producto* para un *cliente* en un momento definido en el *tiempo*.

2.2 Generalización o niveles en las dimensiones del cubo

El usuario define los niveles de análisis en cada una de sus dimensiones. *Ejemplo*. En la tabla 1 se pueden apreciar varios niveles de generalización o niveles de análisis, con respecto a variables con datos geográficos, datos de tiempo, entre otras

Nivel de Análisis	Geográfico	Producto	Servicio a Autómovil	Tiempo
1	Continente	Clase	Tipo	Año
2	País	Subclase	Frecuencia	Mes
3	Estado	Identificador		Quincena
4	Municipio			Semana
5	Ciudad.			Día

Tabla 1.

La hacer esta definición, internamente la herramienta utilizá la *función de agregación sumar*, la cual se combina con el lenguaje de especificación de regiones (ver punto 2.5.1), permitiendo realizar análisis a los diferentes niveles de abstracción definidos en cada uno de los ejes. *Ejemplo*. Los análisis pueden ser a nivel de ciudad, o sumarizar los datos de ciudad para hacer el análisis a nivel municipio, o sumarizar los datos nuevamente para hacer el análisis a nivel estado, y así consecutivamente.

2.3 Generar los datos y cargar el cubo de datos

Las especificaciones que se describen en las variables de interés y el nivel de concepto de cada una de ellas se usan para construir el cubo de datos. Estás especificaciones ayudan a definir los procesos de extracción a partir de las bases originales y contenedoras de los datos de las variables (bases de usuario).

Estas tareas de extracción pueden ser sencillas o complejas dependiendo de la organización de la base fuente de datos. Si la esta base puede ser accesada con instrucciones de manipulación de datos como *SQL* o *4GL* (sistemas manejadores de bases de datos como *INFORMIX*, *ORACLE*, entre otros), la extracción se podría reducir a usar este tipo de instrucciones para extrer los datos, pero si esta es una base donde no se puede accesar con un lenguaje de manipulación, es necesario desarrollar programas con un lenguaje que permita accesar y extraer los datos (archivos de *RM-COBOL*, archivos de *BASIC*, entre otros)

La complejidad de generar los datos para la base de minería (ver sección 3) generalmente se incrementa, dado que en las bases de los sistemas generadores o capturadores de datos, la organización no esta estructurada para realizar operaciones de minería. Estas bases se definen para llevar a cabo operaciones relacionadas a la dinámica diaria o periódica de consultas, reportes y otras actividades y no para realizar análisis históricos, búsqueda de patrones de comportamiento, clasificación de objetos, pronósticos, entre otros análisis de datos en grandes volumenes.

Ejemplo. Una base con una estructura que dificulte la extracción para formar el cubo, es la que almacena las ventas mensuales de gasolina con clave GA01 y se representa en las tablas 2, ,3 y 4.

AÑO	01	02	03	04	05	06
1990	890	0		650	540	650
1991	890	850	700	660	550	600
1992	900	850	710	680	570	610
1993	910	850	720	660	580	610

AÑO	07	08	09	10	11	12
1990	670	789	770	770	800	890
1991	620	690	760	760	820	890
1992	650	690	775	775	810	890
1993	650	690	790	790	850	890

Tabla 2. Cliente 02 en Chihuahua

AÑO	01	02	03	04	05	06
1991	880	0		640	510	610
1992	890	820	710	620	550	610
1993	880	810	700	710	540	600
1994	890	820	720	730	560	620

AÑO	07	08	09	10	11	12
1991	620	790	720	720	780	850
1992	620	640	640	650	770	800
1993	710	660	650	650	780	810
1994	750	660	650	650	800	850

Tabla 3. Cliente 03 en Chihuahua

AÑO	01	02	03	04	05	06
1992	810	0		610	500	610
1993	810	840	660	610	540	600
1994	810	840	660	610	540	600

1995	810	840	660	610	540	600

AÑO	07	08	09	10	11	12
1992	670	790	770	770	800	890
1993	620	690	700	700	750	760
1994	620	690	700	700	750	760
1995	620	690	700	700	750	760

Tabla 4. Cliente 01 en Sonora

Para poder utilizar los datos en los procesos de búsqueda de patrones se requiere estructurar como se muestra en las tablas 5, 6, 7 y 8

Llave1	Descripción
AC01	Aceite número 1
AC01	Aceite número 2
GA01	Gasolina número 1
GA02	Gasólina número 2

Tabla 5. Dimensión 1 *Producto*, con 2 niveles, clase e identificador.

Llave2	Descripción
MXSO 01	Cliente de México, estado de Sonora, identificador 01
MXSO 02	Cliente de México, estado de Sonora, identificador 02
MXCH 02	Cliente de México, estado de Chihuahua, identificador 02
MXCH 03	Cliente de México, estado de Chihuahua, identificador 03

Tabla 6. Dimensión 2, *Cliente,* con 3 niveles; país, estado e identificador.

Llave3	Descripción
199001	Enero de 1990
199002	Febrero de 1990
199003	Marzo de 1990
199004	Abril de 1990
199005	Mayo de 1990
...	...
...	...
199511	Noviembre de 1995
199512	Diciembre de 1995

Tabla 7. Dimensión 3, *Tiempo* con 2 niveles; año y mes.

Llave1	Llave2	Llave3	Valor
AC01	MXSO01	199001	810
AC01	MXSO01	199002	0
AC01	MXSO01	199003	
AC01	MXSO01	199004	610

Tabla 8. Dimensión 4, *venta*, dimensión donde buscar el patrón de comportamiento interesante.

Este cambio de estructura en los datos, se refleja en los datos que forman parte de la *carga inicial* (gran volumen de datos) a la base de minería, actividad que abarca analizar, diseñar, construir e implantar los algortimos de extracción o generación de datos. En nuestro caso, la carga inicial es la extracción y generación de archivos de tipo ASCII los cuales se cargan a la base de datos de minería.

Otra actividad posterior y necesaria para el proceso de minería, es la definición de *cargas de actualización* (menor volumen de datos que la carga inicial) a la base de minería. La actualización se debe realizar en períodos que se consideren conveniente para mantener vigentes los resultados de la minería. Está actividad también contiene la complejidad que se describe en la carga inicial, además de que se debe de considerar el no cargar datos duplicados.

Cumplidos los procesos de crear la base de minería y cargarla de datos se procede a los siguientes pasos.

2.4 Definir los horarios de trabajo de los procesos de minería

Dependiendo de la carga de trabajo de los servidores donde se almacena la base de datos de minería, se procede a programar el inicio y fin de las ejecuciones a las solicitudes de búsqueda de patrones o comportamientos interesantes.

Esta programación para aprovechar los recursos de cómputo se recomienda sea definida en horarios nocturnos, cuando los servidores generalmente estan libres de procesamientos de datos. *Ejemplo.* iniciar los procesos a las 22:00 y terminarlos a las 08:00.

Si se desea esta programación de inicio o fin de minería más autónoma, se pueden desarrollar agentes de *software* (programas) [1] que detecten la no actividad en la computadora y si ellos consideran conveniente iniciar la tarea de extracción y análisis de datos.

2.5 Generar las preguntas (definir región y patrón a buscar)

Generalmente el cubo de minería donde se realizaran las búsquedas, almacena una gran cantidad de datos históricos, pero al usuario solo le interesa analizar en una región más pequeña, región donde buscar su patrón de comportamiento de interés. La definición de la pregunta se divide en dos partes a describir:

- La región que se desea a analizar
- El patron a buscar en la región.

2.5.1 Región a Analizar

Para definir una región en el Minería de Datos-ANASIN, la herramienta proporciona un lenguaje de especificación de "*tramos*" o intervalos de búsqueda en cada uno de los ejes, los operadores usados en lenguaje se pueden observar en la tabla 9.

Símbolo	Significado	Ejemplo
.	Buscar dentro de un intervalo de claves, donde el símbolo se coloca en la clave inicial y final	AC01.AC10; en el eje de producto buscar en los que se tienen las siguientes claves, AC01, AC02, AC03,...,AC09 y AC10
|	Buscar en claves específicas, donde cada una de las claves a buscar esta separada por el símbolo |	AC01|AC02|AC03|AC06, en el eje de producto, solo buscar en los productos AC01, AC02, AC03 y AC06
*	Buscar en todos las claves que esten definidos para esa dimensión o eje. También puede funcionar como un comodín , que indica que se desea la búsqueda de todos los elementos de un nivel inferior	MX*; en eje de cliente, buscar en los clientes cuyas claves inicien con MX o sea MXSO01.MXSO99 y MXCH01 ... MXCH99 (en los que existan).

Tabla 9. Operadores del Lenguaje de Especificación.

Un ejemplo con los operadores para definir una región es:
producto = AC*
cliente = MXSO01|MXSO02|MXCH03|MXCH06
tiempo = 199501.199512
La región son todos los productos cuya clave inicie con AC; los clientes cuya clave es MXSO01, MXSO02, MXCH03 y MXCH06; y en el año de 1995, desde el mes 01 al mes 12.

Si el nivel del producto va de AC01 a AC99, Esto define un cubo de posiblemente 99 * 4 * 12 < 4800 registros a analizar.

2.5.2 Patron de comportamiento o situación interesante

El usuario define un patrón de comportamiento (ver figura 4) a buscar en la región que el defina (ver figura 5)

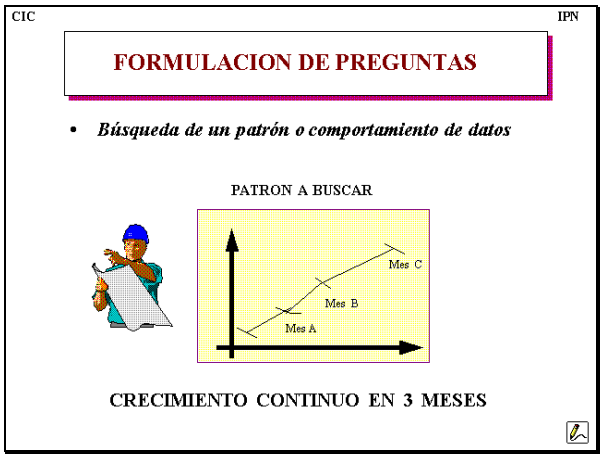


Figura 4

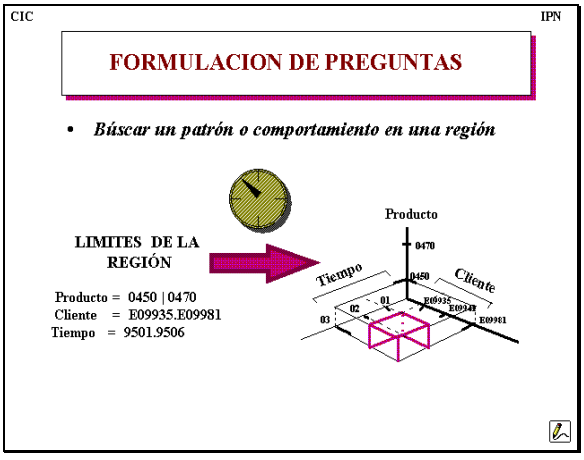


Figura 5

Minería de Datos-ANASIN, cuenta con un catálogo de patrones de comportamiento bien identificados a través de *n momentos o puntos del tiempo* (los puntos pueden ser

dias, semanas, meses, años u otra unidad de tiempo), catálogo que ayuda al usuario a buscar un conocimiento oculto en sus datos. Entre los comportamientos se hallan:

- Crecimientos
- Crecimientos con valles
- Crecimientos con crestas
- Decrecimientos
- Decrecimientos con valles
- Decrecimientos con crestas
- Crecimientos escalonados
- Decrecimientos escalonados
- Zig Zag
- Constantes

Cada uno de los patrones anteriores tiene su expresión matemática que lo define.

Ejemplos:

a) Crecimiento en 4 puntos:
(mesa)<(mesb) y
(mesb)<(mesc) y
(mesc)<(mesd)

b) Crecimiento con cresta en 5 puntos:
(mesa)<(mesb) y
(mesb)<(mesc) y
(mesc)<(mesd) y
(mesd)>(mese) y
(mesc)<(mese)

c) Decrecimiento en 4 puntos:
(mesa)>(mesb) y
(mesb)>(mesc) y
(mesc)>(mesd)

d) Crecimiento con cresta en 5 puntos:
(mesa)<(mesb) y
(mesb)<(mesc) y
(mesc)<(mesd) y
(mesd)>(mese) y
(mesc)<(mese)

e) Crecimiento escalonado en 4 puntos:
(mesa)=(mesb) y
(mesb)<(mesc) y
(mesc)=(mesd)

f) Crecimiento escalonado en 6 puntos:
(mesa)=(mesb) y
(mesb)<(mesc) y
(mesc)=(mesd) y
(mesd)<(mese) y
(mese)=(mesf)

Estos patrones en la herramienta (figuras 6 y 7) se presentan en gráficas que ayudan a visualizar los patrones de comportamiento a buscar.

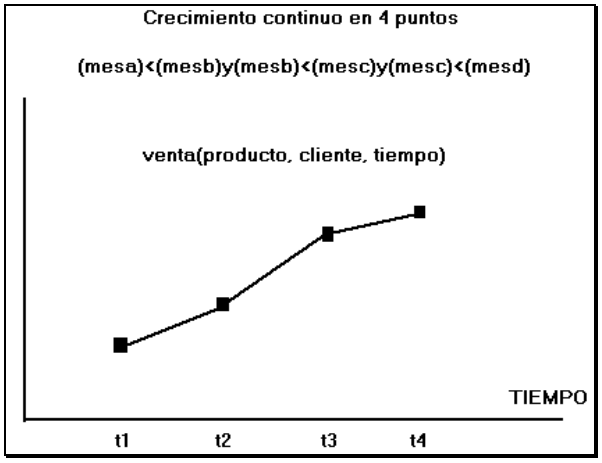


Figura 6.

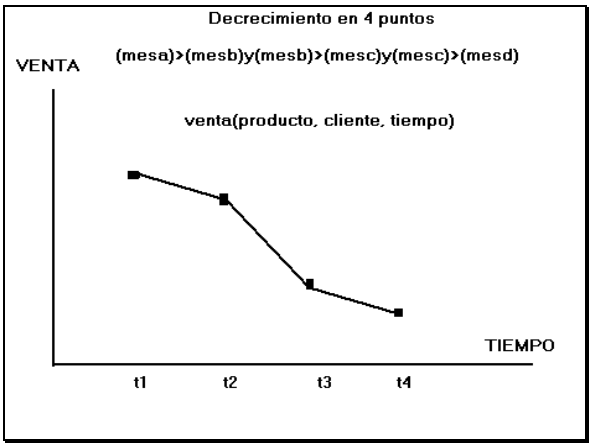


Figura 7.

Con los dos componentes de las preguntas descritos anteriormente, se pueden realizar las siguientes búsquedas:

- ¿ Con que clientes la venta de gasolina tiene un crecimiento continuo en 4 meses?

¿ En que clientes se ha mantenido una venta o un consumo en 3 meses ?

¿ Cuáles son los clientes en los que se ha mantenido un porcentaje de variación mínima de 1 %, sin importar cuantas unidades de tiempo se ha sostenido la variación ?

Preguntas que se pueden armar con un lenguaje de consulta (SQL o 4GL), pero su construcción puede resultar difícil,

2.6 Solicitud de proceso de extracción y ánalisis

La solicitud de la ejecución a responder la pregunta, se registra en una base de datos que contiene todos los procesos de extracción y análisis o búsqueda de patrones.

2.7 Ejecución de Análisis

La ejecución de los procesos de extracción de datos y búsqueda de patrones (análisis) se realizaran de acuerdo a los horarios programados, donde el programa extractor y minero utilizaran los recursos del servidor donde se almacena la base de minería.

2.8 Resultados del Análisis

El éxito en los procesos de Minería de Datos o de análisis automático en un gran cúmulo de datos, se deberá a la interpretación correcta de los resultados. Para lo cual ayuda mucho la forma en que se presenten al usuario los resultados. Está debe ser sencilla y entendible para una fácil interpretación (ver figura 8).

El módulo de Minería de Datos ANASIN, presenta tanto reportes como gráficas de los resultados de la extracción y búsqueda de patrones. Los resultados son los siguientes:

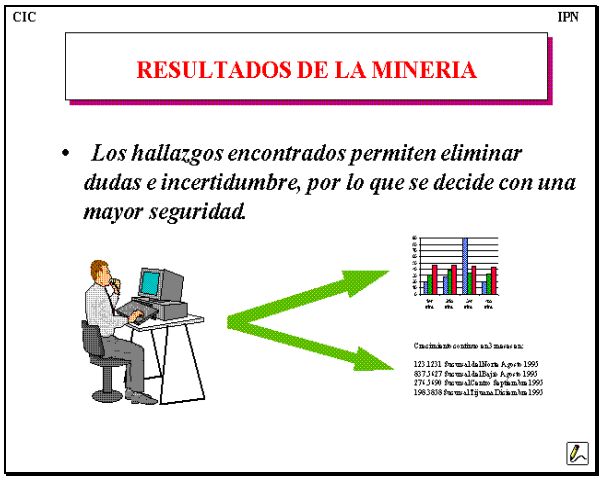


Figura 8

- Reporte que muestra el resultado o la indicación del tiempo y lugar donde se cumplió el patrón buscado
- La extracción que contiene los lugares donde se cumplio el criterio que define la región y en los cuales se busco el patrón (con éxito o sin éxito)
- Gráficas para visualizar los lugares donde se busco el patrón
- Gráficas para visualizar donde tuvo éxito la búsqueda del patrón, es decir, donde se encontró el patrón definido en la pregunta.

Ejemplo de esto son las gráficas 9, 10, 11 y 12.

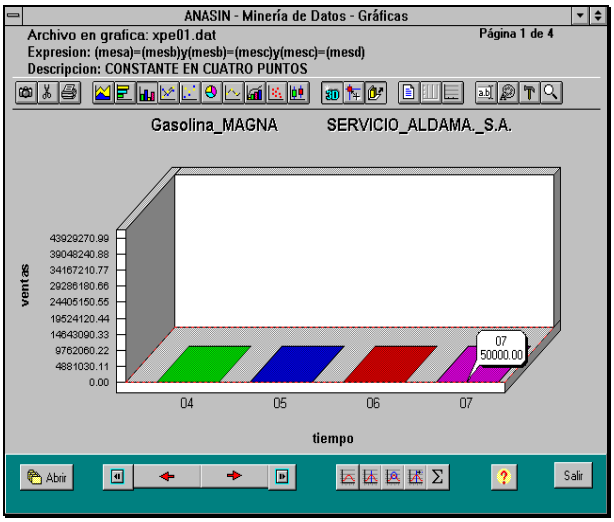


Figura 9. Región de éxito para el patrón de búsqueda “Constante en 4 puntos”

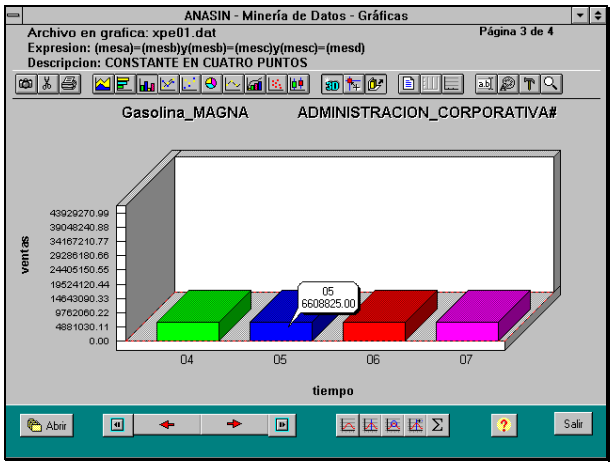
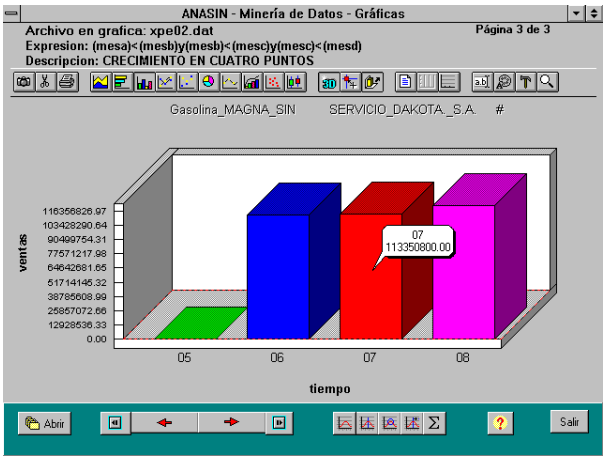
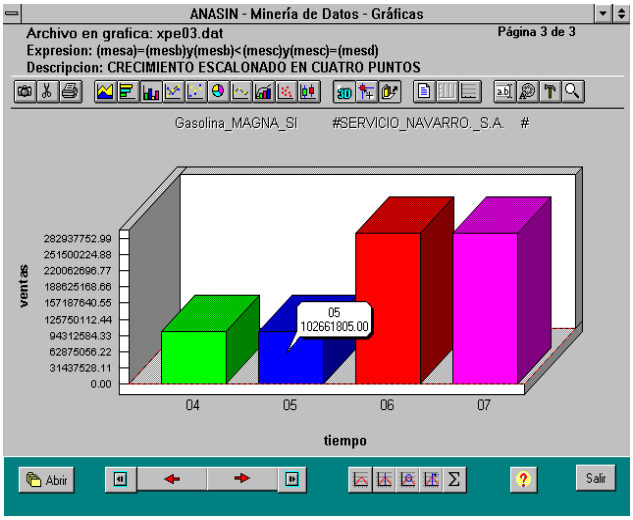


Figura 10. Región de éxito para el patrón de búsqueda “Constante en 4 puntos”



Fígura 11. Región de éxito para el patrón de búsqueda “Crecimiento en 4 puntos”



Fígura 12. Región de éxito para el patrón de búsqueda “Crecimiento escalonado en 4 puntos”

3. Un Desafío en la Minería con Cubos de Datos

Para poder aplicar la técnica de cubos de datos, es necesario realizar una transformación a los datos del usuario para dejarlos listos como se requiere (figuras 2 y 3), la trasformación es originada como ya se dijo en el punto 2, por las fuentes de datos (sistemas operacionales) que no estan diseñadas para realizar minería, por lo cual es necesario su preparación.

Un caso en extremo puede plantearse en la siguiente forma, la estructura óptima de un sistema escolar podría ser la que se presenta en la tabla 10, que almacena datos históricos de totales por alumno inscritos en una institución con varios planteles.

No	Campo	Tipo	Inicio	Fin
1	Matrícula	char(8)	0	7
2	separador	char(1)	8	8
3	clave de plantel	char(3)	9	11
4	separador	char(1)	12	12
5	clave de carrera	char(2)	13	14
6	separador	char(1)	15	15
7	plan de estudios	char(2)	16	17
8	separador	char(1)	18	18
9	Ingreso	entero(3)	19	20
10	Separador	char(1)	21	21
11	causa ingreso	entero(3)	22	23
12	Separador	char(1)	24	24
13	causa egreso	entero(3,)	25	26
14	separador	char(1)	27	27
15	créditos acumulados	entero(4)	28	30
16	separador	char(1)	31	31
17	porcentaje acréditados	entero(4)	32	35
18	separador	char(1)	36	36
19	créditos obligatorios	entero(4)	37	39
20	Separador	char(1)	40	40
21	Porcentaje obligatorios	entero(4)	41	44
22	Separador	char(1)	45	45
23	créditos optativos	decimal(3, 0)	46	48
24	Separador	char(1)	49	49
25	Porcentaje optativos	decimal(4, 0)	50	53
26	Separador	char(1)	54	54
27	Aprobados en ordinario	decimal(2, 0)	55	56
28	Separador	char(1)	57	57
29	Reprobados en ordinario	decimal(2, 0)	58	59
30	Separador	char(1)	60	60
31	Aprobados en extraordinario	decimal(2, 0)	61	62
32	Separador	char(1)	63	63
33	reprobados en extraordinario	decimal(2, 0)	64	65
34	separador	char(1)	66	66
35	promedio	decimal(4, 0)	67	70
36	separador	char(1)	71	71
37	inicio de carrera	decimal(3,	73	74

		0)		
38	Separador	char(1)	75	75
39	Fin de carrera	decimal(3, 0)	76	78
40	Separador	char(1)	79	79
41	Sexo	char(1)	80	80
42	Separador	char(1)	81	81
43	Auxiliar	char(1)	82	82
44	Separador	char(1)	83	83
45	Fecha de nacimiento	char(6)	84	89
46	Separador	char(1)	90	90
47	Nacionalidad	char(1)	91	91

Tabla 10. Datos de Alumnos

Algunos patrones de interés a buscar a través de las generaciones podrián ser:

- La matrícula
- Por cada una de las causas de ingreso
- Por cada una de las causas de egreso
- Porcentaje de créditos aprobados en forma normal
- Porcentaje de créditos aprobados por extraordinario
- Porcentaje de créditos reprobados
- Promedio de calificaciones
- Edad

Los cuales se estudiarian con respecto:

- Plantel
- Carrera
- Sexo
- Causas de ingreso
- Causas de egreso
- Generación

La dimensión de nuestros cubos hasta aquí es de 6, pero como se mantiene un histórico por alumnos, es decir la matrícula se convierte en un eje más, la dimensión de nuestros cubos será finalmente de 7.

Si tuviesemos en total 13 valores diferentes que definen 13 posibles patrones a buscar, cada uno de nuestros registros de los alumnos sería necesario duplicarlo 13 veces (promedio, porcentaje aprobado, porcentaje reprobado, causa de ingreso, causa de egreso, entre otros). Las preguntas posibles a contestar serían:

- ¿ En que alumnos por tipo de ingreso crece su promedio ?
- ¿ En que alumnos por tipo de ingreso decrece su promedio ?
- ¿ En que alumnos por tipo de ingreso se incrementa su porcentaje de aprobadas ?
- ¿ En que alumnos por tipo de ingreso se incrementa su porcentaje de reprobadas ?
- ¿ En que alumnos por tipo de ingreso se incrementa su porcentaje de aprobadas obligatorias ?
- ¿ En que alumnos por tipo de ingreso se incrementa su porcentaje de reprobadas obligatorias?

Cada registro con una estructura similar salvo por el valor de una variable que define el tipo de patrón que almacena el registro.

Es decir, si tuviesemos en total 8,000 registros, es necesario ahora tener 104,000 registros en la tabla intersección. Aunque solo tuviesemos 10 planteles, 5 carreras, 2 sexos, 10 generaciones y 100 causas de egreso y ingreso

En forma similar tenemos, si la estructura de almacenamiento de ventas de gasolina por cliente contiene por registro la venta de un año en doce items, es necesario cada registro duplicarlo 12 veces y registrarlo 12 veces en la tabla que permite almacenar los comportamientos. De aqui, si tenemos 10,000 clientes, 4 productos y almacenados 10 años, nuestra información en la tabla de intersección de los cubos, registraria al menos 10,000 * 4 * 10 *12 = 4,800,000 registros.

Aquí no se se toman en cuenta los registros que pueden facilitar el tiempo de respuesta y que son los precalculados que agilizan el tiempo de respuesta pero incrementan la cantidad de registros [7].

Como se observa de alguna forma nuestros datos se duplican y por lo tanto existe la necesidad de manejar lo que se conoce como *un cubo virtual*, el cual se generaría cada vez que se procesa una pregunta. Con solo la ventaja de no duplicar la información y no almacenar dicho cubo en disco; y las desventajas de buscar óptimizar los tiempos necesarios para su cálculo, que dependererían de la región de interés a analizar y el algoritmo que definiese el área de almacenamiento temporal.

Esto es un problema que se mantiene vigente y necesario de resolver [5]. Aunado a este desafio existen otros como el de generar en forma autómatica los algoritmos de los procesos de extracción y carga de datos a la base de datos que se usa para la minería (ver sección 2.3).

4. Trabajos Futuros.

Esta herramienta se utiliza como apoyo en varios proyectos que se estan planeado en conjunción con otros Laboratorios del C.I.C, entre los cuales destacan:

- Uso de Agentes [2] para la Minería de Datos en Sistemas Distribuidos y Cooperativos [8]
- Uso de Agentes para la Minería Distribuida
- Uso de Agentes para la Minería en Texto
- Generación de nuevos agentes (mineros que buscan variables con comportamientos similares, ya sea que crezcan, decrezcan o comportamientos contrarios) [4]

5. Conclusiones

Algo que complica el llevar a cabo el desarrollo de herramientas con tecnología nueva, es que implica varios tipos de conocimiento, entre los cuales tenemos:

- Algoritmos matemáticos
- Organización de las bases de datos
- Algoritmos de recuperación
- Diseño de interfaces de usuario
- Sistemas operativos

Una de las principales aportaciones de nuestro desarrollo, es el modelo de trabajo que permite aprovechar los recursos de cómputo y permite aún más buscar este aprovechamiento, además como ya se menciono en los trabajos futuros, permite buscar aplicar otras tecnologías cómo son la de agentes de software.

6. Características del Software

Modelo de trabajo. Cliente/servidor
Software en Servidor

- *UNIX SCO*
- *INFORMIX (Online y 4GL)*
- *SAMBA*
- Programas en C
- Programas en *Shell Scripts* de *UNIX SCO*

Software en Cliente

- *Windows NT* o *Windows 95*
- Ejecutable en *DELPHI*
- *INFORMIX*

7. Grupo de desarrollo de versión 2.1[2]

Integrantes del Laboratorio de Sistemas de Información en 1998:

- Lilia González Rodriguez (Analista / Programador, Documentador)
- Alfonso Garcia González (Analista / Programador, Documentador)
- Absalom Zamorano Castellanos (Analista / Programador)
- Karina Ortiz Nicolas (Analista / Programador)
- Rodolfo Rodríguez bando (Analista / Programador)
- Bertha Patricia Ramírez Arzate (*Beta Tester*)
- Ixcheel Martínez Castillo (*Beta Tester*)
- M. en C. Gilberto Lorenzo Martínez Luna (Líder de Proyecto)

Asesores

- Dr. Adolfo Guzmán Arenas
- M. en C. Guillermo Rafael Domínguez de León

Referencias

[1] Bigus Josep P. "*Data Mining With Neural Networks*", McGraw-ill 1996.

[2] Davidsson P., "*Autonomous Agents and the Concept of Concepts*" Departament of Computer Science, Lund University, Sweden 1996.

[3] Guzmán Arenas A, "Estado del Arte y de la Práctica en Minería de Datos, Análisis y Critica", Conferencia Magistral, Cuba, Marzo de 1996

[4] Guzmán Arenas A, "Uso y Diseño de Mineros de Datos", Soluciones Avanzadas, Junio de 1996

[5] Ming-Syan Chen, Jiawei Han, and Philip S. Yu, Fellow, "*Data Mining: a view from database perspective*", IEEE, Dic. 1996

[6] http: // www. kdnuggets. com / siftware. htm

[7] Harinayan v., Rajamaran a., Ullman j, " Implementing Data Cubes Efficiently", Stanford University

[8] P. Papazoglou Mike, K. Sellis Timos, " *International Journal of Intelligent & Cooperative Information Systems* ", IJICIS, Volume 1, Number 1, Queensland University of Technology Brisbane, Australia, University of maryland, College Park, USA, March 1992.

[2] (2). La versión 1.0 fué desarrolloda por las empresas *Software-Pro International* e *IDASA*, con apoyo del CONACYT para la Gerencia de Informática y Telecomunicaciones de la C.F.E. dirigida por el Ing. Enzo Molino e Ing. Ramón Soberón Kuri en 1994