Conferencia Mundial
Sobre
Tecnologias de la Información
y
Comunicaciones
2008
Pachuca, Hidalgo, Mexico
25, 26 y 27 de Noviembre 2008

Antecumem, Prototipo de Herramienta para el Análisis de Datos con Cubos en Memoria Principal

G. L. Martínez [(1)], A. Guzmán[(2)]
[(1)]*Centro de Investigación en Computación, Ciudad de México, Distrito Federal, México*
[(2)]*Centro de Investigación en Computación, Ciudad de México, Distrito Federal, México*
e-mail: lluna@cic.ipn.mx

Resumen: *Se describe una herramienta llamada Antecumem que se utiliza para desarrollar análisis en bases de datos almacenadas en memoria principal. La descripción abarca una lista de preguntas de negocio y el almacén de la base de datos. El almacén es una estructura de datos con arreglos y que están ligados entre sí, llamada Arblis, con lo cual no se buscan los datos en disco, lo que reduce el tiempo en la búsqueda de datos. Arblis almacena la base de datos, que es modelada como una base multi-dimensional (cubos de datos). Este modelo, permite definir operaciones con los cubos, operaciones con un interés sobre sucesos a través del tiempo, pero que también pueden ser en cualquier otra dimensión. Una operación con los datos de interés a analizar, puede ser el porcentaje de incremento de un período a otro. Arblis permite responder a la lista de preguntas de negocios aquí planteada.*

Abstract: *It describes a tool called Antecumem which is used for analysis in databases stored in main memory. The description includes a list of questions from business and store the database. The warehouse is a data structure and arrangements that are linked to each other, call Arblis, which does not seek data on disk, which reduces the time in the search for data. Arblis stores the database, which is modeled as a multi-dimensional (data cube). This model lets you define operations in the data cubes, oprations with an interest in events over time, but may also be in any other dimension. An operation with the data of interest to analyze, may be the percentage increase from one period to another. Arblis responding to the list of business questions raised here.*

Keywords: Data Analysis, Data Mining, Database, Multidimensional Database, Data Cube.

1. Introducción. El Tiempo de Respuesta en el Proceso del Análisis de Datos

1.1 Preguntas de Negocio en el Análisis de Datos

El análisis de datos o las preguntas de negocio son indagaciones a las bases de datos de una empresa o institución con el fin de hallar datos valiosos o verificar situaciones que están relacionadas con diferentes tipos de decisiones como: **a)** la operación diaria, **b)** el resolver situaciones de corto plazo o **c)** para la planeación a largo plazo.

Un ejemplo de una pregunta de negocio para tomar una decisión del tipo **c),** es "Localizar los productos de la *temporada* de *octubre a diciembre* que han mantenido una demanda en los últimos *seis años atrás*, y que se han mantenido en los *primeros 10 lugares de ventas*, en las diferentes tiendas del país de una empresa departamental". A través del análisis de las preguntas de negocios que aparecen en varios artículos o libros relacionados al análisis de datos, como [2], [21] y [22] y que son las preguntas más recurrentes a resolver en las empresas o instituciones, se definió una lista de 7 tipos de preguntas y que se listan en la **Tabla 1**. El valor de las preguntas se relaciona con el objetivo de la organización, dueña de los datos y por ser preguntas para el apoyo a la toma de decisiones, se requiere disminuir su tiempo de obtención de las respuestas y en especial el de extracción de los datos que se requieren para obtener las respuestas.

1.2 Objetivo, Agilizar el Análisis de Datos

Para agilizar el análisis de datos, en la parte de reducir el tiempo de obtención de los datos y así obtener las respuestas a las preguntas de la **Tabla 1**, en este trabajo los datos se almacenan en memoria principal, sin sacrificar el tamaño o volumen de ellos; y también aquí

en RAM se analizan los datos; con esto se utiliza una alternativa al "*cache*" en disco. Para esto se emplean estructuras para representar una base de datos multi-dimensional o cubos de datos, los cuales se utilizan por un prototipo de *software* llamado *Antecumem (Análisis Temporal con Cubos en Memoria)*.

Tabla 1. Preguntas de Negocios

No, Nombre de Tipo de Pregunta y Descripción
1. Puntual. Revisar el valor o hecho de interés, de un elemento, en un momento, en un lugar, de un cliente, etcétera.
2. Rango. Revisar el acumulado de hechos, acotado por los rangos en las variables de interés.
3. Eficiencia. En base a los hechos de interés, calcular su porcentaje de incremento o decremento en *dos períodos* con rangos en todas las variables (¿Qué "tanto" mejoramos?).
4. Eficiencia Grupal. En base a los hechos de interés con su porcentaje de incremento o decremento, indicar los n "mejores elementos en *dos períodos* (¿En donde "mejoramos"?)**.**
5. Conservación/Perdida. En base a los hechos y *una variable de interés,* en *dos períodos de tiempo*, observar los elementos, ya sea que permanecen o que desaparecen o sea revisar los n "buenos" elementos, en un período y otro.
6. Temporalidad. En base a los hechos y *una variable de interés*, observar los n "mejores" elementos que permanecen, en dos o más *períodos de tiempo (temporadas)*. Períodos de tiempo que puede ser anual, mensual u otro período.
7. Tendencias. En base a los *hechos y una variable de interés*, buscar en *un rango de tiempo*, los **n** elementos de interés que mantienen una tendencia en ***p*** lapsos de tiempo.

1.3 Una Posible Solución

El acceso a *memoria principal* o RAM es del orden de 10^{-7} *segundos* (100 nano-segundos), con lo cual vemos que el tiempo de extracción se puede reducir, al ya no realizar la extracción a partir del disco. Este valor teórico es lo que motiva en gran parte este proyecto. La limitación principal para esta solución es la cantidad de RAM disponible. Por ejemplo, para una PC personal, su tamaño es 50 veces menor que la del disco (disco de 100 GB, RAM de 2 GB), pero aún así, es razonable el utilizar esta cantidad de RAM y caracterizar los problemas que se pueden resolver utilizando esta cantidad de memoria.

1.4 Una Característica del problema

Las ***preguntas de negocios*** listadas en la **Tabla 1**, y que aquí se resuelven, pueden ser expresadas por medio de expresiones del lenguaje de consulta estructurado (***SQL***, ***Structured Query Lenguaje***), esto en caso de tener una base de datos relacional**.** *Antecumem* ha organizado una manera especial de capturar los parámetros de estas preguntas para luego acceder a la estructura *Arblis.*

También *Antecumem* puede capturar un número de expresiones o consultas que son posibles de construir con la Forma General de la **Figura 1** y que depende del número de dimensiones y de las jerarquías en las dimensiones, como se describe en [1]**.** El resultado de responder las expresiones o consultas se le conoce como *vistas* o cubos de datos, además dependiendo del modelado de datos para responder las preguntas, como en [2] se pueden crear operadores para trabajar con las vistas o los cubos de datos

SELECT	A, S(a(A))
FROM	D
GROUP BY	A

Donde:
- A es un subconjunto de atributos de las relaciones en **D={d_1, .., dn}**, y que forman la base de datos.
- S(a(A)) es un agregado de interés sobre algunos de los atributos en A, pero de tipo numérico, como SUM, MAX, MIN, etc.

- Las **d_i** , **i=1,…,n** se conocen como dimensiones.

Fig. 1. Forma General de Consultas que Responden a una Pregunta de Negocio en Bases relacionales

2 Soluciones a la Reducción de Tiempo en la Extracción de Datos

El tema de agilizar el tiempo de respuesta, es un tópico y un área de investigación en las Base de Datos [3] que lleva más de 40 años en desarrollo, pero continua siendo actual por el constante crecimiento de las bases de datos, las recientes y variadas tecnologías de adquisición, almacenamiento y dispersión de datos, además de la importante necesidad de realizar constantes análisis a estos datos. Análisis que se realiza con los procesos de Minería de Datos y el Procesamiento Analítico en Línea (OLAP, ***Analytical Processing On-Line***).

En estos procesos, primero con ayuda de los programas llamados extractores de datos se forma el espacio de búsqueda; y después la corroboración de hipótesis se realiza con programas analizadores. Todo este trabajo puede tardar mucho tiempo [2] y [4], ya que se realizan en bases de datos que contienen grandes cantidades de registros. En www.wintercorp.com existe

un registro de las bases de datos más grandes para la toma de decisiones u ***OLAP***, del orden de 100 *terabytes.* El cubo de datos [1] y [2] puede formar parte de las siguientes soluciones para agilizar el tiempo de respuesta:

1) La materialización de vistas en disco que se detalla en [1] y [5], con el correspondiente calculo de su tamaño en [6].
2) El tratar la similaridad en consultas, enfoque dado en [7], [8] y [9], para crear una función de distancia entre los objetos de la base de datos.
3) Crear estructuras para resolver consultas de tipo rango, ya sea sumas o máximos. Esto se puede ver en [1], [2], [10] y [11].
4) Crear funciones de densidad en datos continuos, otra forma es compactar en forma especial los cubos [12].
5) Aprovechar el orden en la estructura de Lattice que almacena en sus nodos las vistas en disco como se indica en [4].
6) Crea Sistemas Administradores de Bases de Datos (SMBD) en Memoria Principal, como se diseña e implementa un *SMBD* en [13] en memoria con todas los componentes de un *SMBD* en disco, pero en él, su fin es realizar transacciones o trabajar un sistema de tipo ***OLTP*** (***Online Transaction Processing***).
7) Datos en Memoria Principal, existen trabajos como [14], que también carga una base de datos en memoria principal, su objetivo es apoyar la toma de decisiones, con las premisas de "dado un dato hallarlo en el menor número de accesos a la estructura en memoria donde se almacena la base de datos".

Algunas de estas soluciones son para disminuir el tiempo de respuesta en accesos a bases de datos y en su análisis. Las soluciones que acceden disco, son alternas a nuestra solución, y las que cargan en memoria principal bases de datos tienen otro objetivo: transacciones o responder accesos directos a registros, más no responder preguntas de negocio de las que aquí se plantean.

3. Análisis y Diseño de la Solución

El análisis y diseño de *Antecumem*, podemos resumirlo en: **1.** Localizar en las fases de los procesos de *OLAP* y Minería de Datos donde se puede reducir el tiempo **(sección 3.1)**; **2.** Diseñar la estructura *Arblis*, almacén de datos en memoria **(sección 3.2)**; **3.** Modelar cubos o bases de datos multidimensionales con *arblis* **(no se presenta en este artículo),** y **4.** Modelar las preguntas de negocio que pueden ser resueltas con cubos de datos y los algoritmos que responden a estas preguntas **(no se presenta en este artículo)**.

3.1 Las Fases de *OLAP* y Minería de Datos

Los procesos de *OLAP* y de Minería de Datos, coinciden en dos fases como se describen en las modelaciones realizadas en [15], [16], [17], [18] y [19]; las cuales son: 1) extracción y 2) el análisis. Las fases de estos procesos se ilustran en la **Figura 2**. En la fase de extracción, ahora se evitara el acceso a disco, al cargar en RAM la base de datos para responder cualquier consulta de la forma ya mencionada, con lo cual se elimina el trabajo del *SMBD* a disco, es decir ahora tenemos las dos fases en memoria principal y realizadas por un mismo programa, como se observa en la **Figura 3**.

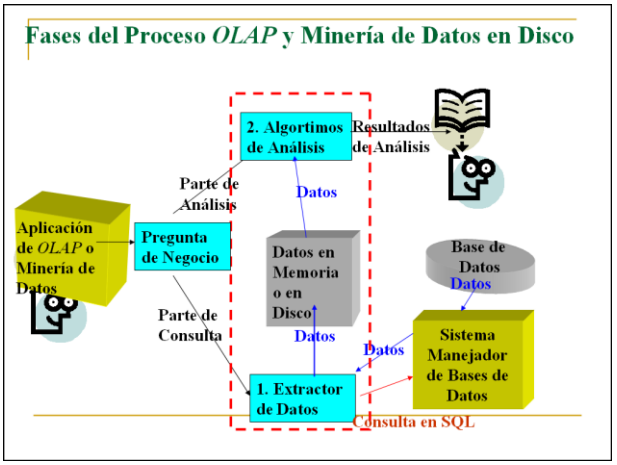


Fig. 2. Fases de Extracción y Análisis

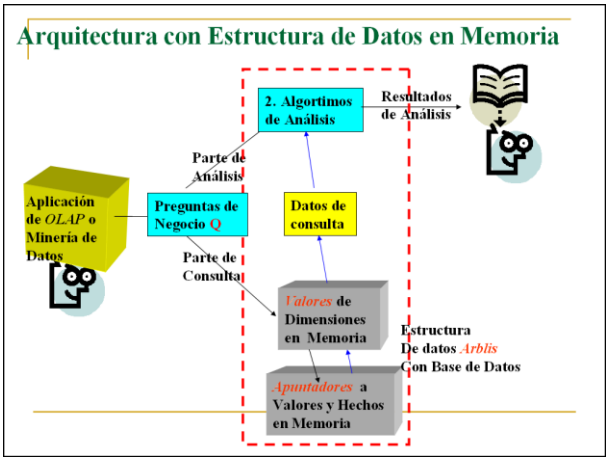


Fig- 3. Arquitectura con Datos en RAM.

La estructura *Arblis* almacena la Base de Datos Multi-Dimensional (***BDMD***) o el cubo de datos original, el cual se trabajara a partir de sus variables de interés. Como ya se menciono en la sección 1.4, las variables se denominan dimensiones, y los valores a sumarizar o agregar, se conocen como hechos. De igual forma, cualquier resultado que se obtenga a partir de una consulta a la *BDMD* o *Arblis* se llamara cubo de datos, con lo cual podemos obtener varios cubos que participen al resolver una pregunta.

3.2 *Arblis*, la Estructura en Memoria Principal para Almacenar los Datos a Analizar

Para explicar la organización de la estructura en memoria, se hará con los datos de la **Tabla 2**, que tiene 18 hechos (columna *sales*). La estructura se aprecia en el primer renglón, que es además de los hechos, tres dimensiones y que forman la base *autos(Model, Year, Color, Sales)*, que puede ser definida como $C(d_1, d_2, d_3)$. Se tienen las siguientes características en esta base:

- Son tres dimensiones y los hechos.
- Dimensión d_1 con dos valores *{Chevy, Ford}*.
- Dimensión d_2 con 3 valores *{1990,1991,1992} y*
- Dimensión d_3 con 3 valores *{blue, red, white}*.

Tabla 2. Bases de Datos "*autos*" con 18 hechos.

Registro	*Model*	*Year*	*Color*	*Sales*
0	Chevy	1990	Blue	5
1	Chevy	1990	Red	87
2	Chevy	1990	white	62
3	Chevy	1991	Blue	54
4	Chevy	1991	Red	95
5	Chevy	1991	white	49
6	Chevy	1992	Blue	31
7	Chevy	1992	Red	54
8	Chevy	1992	white	71
9	Ford	1990	blue	64
10	Ford	1990	red	62
11	Ford	1990	white	63
12	Ford	1991	blue	52
13	Ford	1991	red	9
14	Ford	1991	white	55
15	Ford	1992	blue	27
16	Ford	1992	red	62
17	Ford	1992	white	39

La reorganización de la **Tabla 2**, se observa en la **Tabla 3**, que es en dos arreglos, donde los valores de las dimensiones están en el *arreglo 1* y los apuntadores que se agregan están en el *arreglo 2*. Brevemente los valores están ordenados en la forma siguiente:

- Dimensión d_1 *model*, segmento 1, posiciones 0-1.
- Dimensión d_2 *year*, segmento 2, posiciones 2-7 y
- Dimensión d_3 *color*, segmento 3, posiciones 8-25.
- Las posiciones 0 y 1 están ligadas a las posiciones 2 y 5 respectivamente.
- Las posiciones entre 2 y 7 están ligadas a las posiciones 8, 11, 14, 17, 20 y 23 respectivamente.
- En cada una de las entradas de la 8 a la 25, están almacenados los hechos.

Aquí, 2, 6 y 18, son los tamaños de los segmentos pertenecientes a las dimensiones en el arreglo 1 que almacena los valores de las dimensiones.

- Tanto el arreglo que almacena los valores de las dimensiones como el arreglo que almacena los apuntadores son de tamaño 26.

Tabla 3. Reorganización de Tabla 2 en la Estructura *Arblis*

		Arre-glo 1	Arre-glo 2		
Dimen-Sión	Posi-ción	Valor	Apun-tador	Seg-mento	Apun-ta a
Model	0	Chevy	2	1	Seg-mento 2
	1	Ford	5		
Year	2	1990	8	2	Seg-mento 3
	3	1991	11		
	4	1992	14		
	5	1990	17		
	6	1991	20		
	7	1992	23		
Color	8	Blue	5	3	Hechos
	9	Red	87		
	10	White	62		
	11	Blue	54		
	12	Red	95		
	13	White	49		
	14	Blue	31		
	15	Red	54		
	16	White	71		
	17	Blue	64		
	18	Red	62		
	19	White	63		
	20	Blue	52		
	21	Red	9		
	22	white	55		
	23	Blue	27		
	24	Red	62		
	25	White	39		

3.3 Navegación Resumida en *Arblis*

Por la forma en que se describe el arreglo 1, se puede ver que el dominio de valores $D(d_i)$, donde d_i es la i-ésima dimensión de la *BDMD* o cubo, están representados en el arreglo 1 y además se encuentran ordenados tal que se que permite agilizar su recorrido. Todas las entradas o valores $x_i \in d_i$ de las tuplas (x_1, x_2, x_3), si es que existen en el cubo, se pueden hallar recorriendo el segmento correspondiente del arreglo 1 a cada una de las dimensiones.

El valor x_1 que pertenece a la dimensión d_1, se buscaría en el primer segmento del arreglo 1, luego tomar el valor x_2 de la dimensión d_2 y buscarlo en el segundo segmento del arreglo 1; así hasta realizar la búsqueda del valor x_n en su correspondiente segmento (el último) y finalmente tomar el valor del hecho en el arreglo 3, la navegación entre las dimensiones se realiza con los apuntadores del arreglo 2. Es decir, se formaría la coordenada con los valores y el hecho correspondiente de la forma $(x_1, x_2, x_3, v_{1,2,\ldots,n})$, donde $v_{1,2,\ldots,n}$ esta en función de los valores (x_1, x_2, x_3).

3.4 Las Preguntas de Negocio a Resolver con la Estructura

La estructura es útil, ya que permite agilizar el acceso a datos y realizar el análisis de preguntas de la **Tabla 1**, donde ya tiene una clasificación y la descripción de la pregunta. Por restricciones de espacio no se describen los elementos formales que ayudan a definir el tipo de pregunta y los algoritmos en función de los cubos que son necesarios para resolver estas preguntas.

En esta sección se indico la parte del proceso que se lleva de disco a RAM tanto del proceso de OLAP y Minería de Datos. Se describió en forma resumida la estructura *Arblis* que representa la base de datos en memoria; cómo la estructura representa la unidad básica a tratar, un cubo de datos o la vista en detalle; y cómo Arblis o los cubos pueden responder una pregunta de negocio de la **Tabla 1**.

4 Pruebas y Resultados

Las pruebas que se hicieron para mostrar el trabajo de *Antecumem*, son con una base de datos de prueba que se llama *SH* que fue obtenida de la instalación del *SMBD Oracle 9.2i* [www.oracle.com]. Las dimensiones en la base de datos están en la *Tabla 4*.

Tabla 4. Base de Datos *SH*, con 4 dimensiones.

Dimension (d_i)	1	2	3	4	Hechos
Nombre	*Product*	*Customer*	*Promotion*	*Time*	*Sales*
Valores	*10,000*	50,000	501	1,461	1,016,271

Las pruebas aquí descritas se llevaron a cabo con:

- Una PC compatible con procesador XEON a 1.7 MH y RAM con 1.5 GB.
- El SMBD *MySQL* 4.0.2 para NT y programación JAVA con JKD 1.5.
- Sistema Operativo Windows 2000 Profesional.

4.1 Prueba de Velocidad de Acceso a Diferentes Volúmenes de Datos

La **Tabla 5** muestra el comportamiento de tiempos en segundos de una *Pregunta de Tipo Rango* en la base de datos *SH* con incrementos de 1,016,271 registros**.** Se observa el *SMBD MySQL* versión 4.0 resulta más rápido que el programa desarrollado con el número inicial de registros de la base de prueba, pero pasado un límite, *Antecumem* en memoria muestra mayor rapidez.

Tabla 5. Tiempos de Acceso a Millones de Registros

No. de Prueba	Registros	*SMBD*	*Antecumem*
1	1,016,271	0.8	1:016
2	2,032,542	1.52	1:172
3	3,048,813	2.22	1:334
4	4,065,084	2.98	1:516
5	5,081,355	3.95	1:714
6	6,097,626	4.67	1:953
7	7,113,897	5.44	2:234
8	8,130,168	6.13	2:488
9	9,146,439	6.75	2:563
10	10,162,710	9:23	2:735

4.2 Resumen de Pruebas

En la Tabla 6 se tiene un resumen de los resultados en segundos en las diferentes preguntas de negocio de la **Tabla 1**, con diferentes volúmenes de registros, utilizando el SMBD y el *software* desarrollado.

Tabla 6. Tiempos de Respuesta, *SMBD* (S) vs *Antecumem (Antcmm)*

No	*SM-BD*	*Ant-cmm*	*SM-BD*	*Ant-cmm*	*SM-BD*	*Ant-cmm*	*SM-BD*	*Ante-cmm*
	1,016,271		10,162,710		12,195,252		15,244,065	
1	.31 .31	0	(24) 3	0	(22) 3	0	(28) 4	0
2	(5) 0.8	1.390	(25) 9	3.438	(21) 11	4.062	(28) 13	4:875
3	(6) 1.6	2,641	(32) 17	5.094	(32) 20	5.390	(40) 24	5:953
4	(4) 3	0.704	(34) 34	1.16	(56) 41	1.0	(72) 51	1:141
5	(3) 3	1.359	(32) 12	2.125	(30) 15	2.266	(38) 17	2:469
6	(9) 2	2.15	(15) 15	2.390	(34) 19	2.453	(44) 22	2:703

() Primera medición con el SMBD, la segunda generalmente es menor.
* Faltan operaciones para regresar lo que se desea, acorde a la operación de negocios.

4.3 Variante en la Prueba de Temporalidad

Se diseño otra variante de la prueba de temporalidad para analizar más el desempeño de la herramienta. La variante consiste en el aumentar el número de meses (cubos) de análisis y ver el desempeño en la pregunta de temporalidad. Los resultados en segundos están en la **Tabla 7**.

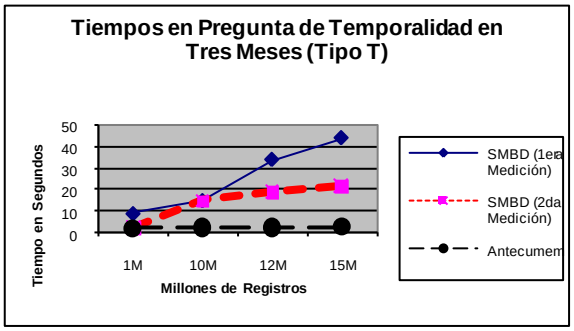
Tabla 7. Tiempos en Pregunta de Temporalidad

Me-ses	SM-BD	*Ant-cmm*	SM-BD	*Ant-cmm*	SM-BD	*Ant-cmm*	SM-BD	*Ant-cmm*
	1,016,271		10,162,710		12,195,252		15,244,065	
3	(9) 1	2.015	(35) 15	2.313	(33) 17	2.344	(44) 22	2:562
4	(10) 2	2.672	(40) 20	3.110	(39) 24	3.125	(51) 30	3:453
6	(9) 3	4.015	(49) 29	4.859	(52) 35	4.891	(67) 50	5:575
12	(11) 6	8.140	(79) 59	10.59	(86) 71	10.92	(109) 89	11:50

() 1era. medición con el SMBD, la segunda generalmente es menor.

4.4 Desempeño Empíricamente Estable

Al graficar el desempeño de la herramienta en cada una de las preguntas contra los diferentes volúmenes de pruebas se observa que es empíricamente estable. Como ejemplo se puede ver **Gráfica 1**.



Gráfica 1. Pregunta de Temporalidad

5 Conclusiones, Ventajas y Trabajo Futuro

5.1 Conclusiones

Las conclusiones de este trabajo son:

1. La estructura *Arblis* está diseñada para buscar en orden en las dimensiones y con los ciclos predefinidos dado un valor en cada dimensión, que se refleja en su construcción y que acelera los tiempos de respuesta las consultas del tipo de la **Tabla 1.**
2. El ***modelo de datos*** permite identificar las partes de las preguntas de la **Tabla 1**.
3. El ***modelo de datos*** permite realizar los algoritmos para responder las preguntas de la **Tabla 1** usando la estructura ***Arblis***. (la demostración formal de estos tres puntos no forma parte de este artículo)

5.2 Ventajas y Desventajas del Prototipo

Ventajas.
a) Agiliza tiempo de respuesta sobre disco, se alcanzo hasta más de 8 veces el tiempo menor de respuesta, dependiendo del tipo de pregunta.
b) La estructura que almacena la base de datos multidimensional no tiene celdas vacías.
c) Puede adaptarse a otras preguntas de negocios que se definan y que trabajan con cubos de datos.

Desventajas.
a) Es un prototipo aún, esto significa que es necesario realizar más código para dejarlo como una herramienta (pero, esto no afecta el desempeño de la herramienta).
b) No tiene una interfaz amigable.
c) Tarda mucho en cargar y formar la estructura, con un millón de registros tarda 2 minutos, con diez millones tarda 120 minutos y aún más en quince millones de registros, aunque esto ya se ha resuelto parcialmente, al

guardar la estructura *Arblis* en disco y leerla, con lo cual la carga se reduce a decenas de minutos.

5.3 Trabajos Futuros

Algunos trabajos que se consideran interesantes son:

1) Acelerar la Carga de la Base a la Estructura. Como se menciono, una de las desventajas es su tiempo de carga. Para acelerar la carga, se puede programar a *Antecumem*, si no hay cambios en los datos y la estructura este almacenada en disco en disco leerla, en otro caso se pregunta al analista si los datos que se tienen pueden ser útiles para realizar análisis, a pesar de que hay cambios, o se puede cuantificar la afectación de los cambios o el porcentaje de los cambios.

2) Mejorar la Interface de Captura y la Entrega de Resultados. Por el momento la captura de la "expresión" que define la pregunta de negocio no es sencilla para un analista de negocio y la respuesta que se entrega es en formato texto que dificulta su interpretación. Una forma de capturar la expresión es mostrando solo las áreas que se requieren para plantear el tipo de pregunta, pues el tipo de pregunta define los parámetros. Una mejora en la entrega de resultados, podría ser si estos se muestran en gráficas de tipo Excel.

6. Referencias

1. J. Gray, A. Bosworth, A. Layman, and H. Pirahesh, "Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals," *Data Mining and Knowledge Discover*, Vol. 1, No. 1, 1997, pp. 29-53, 1997.

2. R. Agrawal, A. Gupta, S. Sarawagi, "Modeling Multidimensional Databases", IBM Almaden Research Center, 1997.

3. Z. Chen, C. Li, J. Pei, "*Recent Progress on Selected Topics in Database Research* ", Microsoft Research, 2003.

4. V. Harinarayan, A. Rajaraman, J. Ullman. "*Implementing Data Cubes Efficiently*", Stanford University, 1996.

5. T. Morzy, M. Wojciechowsky, M. Zakrzewicz, "*Materialized Data Mining Views*", Poznan University of Technology, Institute of Computing Science, 2000.

6. A. Shukla, P. M. Deshpande, J. F. Naughton, K. Ramasamy, "Storage Estimation For Multidimentional Aggregates in The Presence of Hierarchies", Computer Sciences Department University of Wisconsin – Madison, 1996.

7. V. Ganti, J. Gehrke, R. Ramakrishnan, *"Mining Data Streams under Block Evolution",* SIGKDD Explorations, 3(2):1-10, 2002.

8. P. Ciaccia, M. Patella, "Approximate Similarity Queries: A survey", University of Bologna, Italy, 2000.

9. B. Braunmüler, M. Ester, H-P. Kriegel, J. Sander, "*Efficiently Supporting Multiple Similarity Queries for Mining in Metric Databases*", Institute for Computer Science, University of Munich, 2000.

10. A. Gupta, V. Harinarayan, D. Quass, "Aggregate-Query Processing In Data Warehousing Enviroments", IBM Almaden Research Center, Stanford University, 1995.

11. Ch. Ho, R. Agrawal, N. Megiddo, R. Srikant, *"Range Queries in OLAP Data Cubes",* IBM Almaden Research Center, 650 Harry Road, San José, CA 95120, 1997.

12. J. Shanmugasundaram, U. M. Fayyad, P. S. Bradley, "Compressed Data Cubes for OLAP Aggregate Query Approximation on Continuous Dimensions", Microsoft Research, University of Wisconsin, Technical Report MSR-TR-99-13, 1999.

13. K. Jung, K. Lee, "Design and Implementation of Storage Manager in Main Memory Database", System ALTIBASE", Real-Time Tech Lab, ALTIBASE Co. Seoul, Korea, 2003.

14. J. Rao, "Cache Concious Indexing for Decision-Support in Main Memory", Columbia University, 1999.

15. G. L. Martínez, Y. Albores, C. Castillo, "Automatización del Proceso de Minería de Datos", Memoria del 2do. Foro "Computación de la Teoría a la Práctica", Canacintra CIC-IPN, 2001.

16. G.L. Martínez., A. Guzmán., M. Alexandrov, "Modelo de Minería Datos con Ajuste de Curvas", CIICC-2003.

18. Z. Chen, "Intelligent Data Warehousing, From Data Preparation to Data Mining", CRC PRESS, 2000.

19. A. Berson, S. Smith, Data Warehousing, Data Mining & OLAP, McGraw-Hill, 1997.

21. Ch. Li, X. S. Wang, "*A Data Model for Supporting On-Line Analytical Processing*", George Mason University, cli@isse.gmu.edu, 1996.

22. Harjinder S. Gill, The Official Guide to Data Warehousing, ISBN 07897-0714-4, Editorial QUE, 1996.

6. Biografías

	*Nombre del autor*, Licenciatura en Matemáticas Aplicadas en el Instituto Politécnico Nacional en 1985. Maestría en Ciencias en el CINVESTAV en 1998. Los estudios del Doctorado los termino en el CIC-IPN en 2007. Es profesor-investigador en el CIC-IPN en el Laboratorio de Bases de Datos y Tecnología de Software, con una experiencia de más de 25 años en el desarrollo de Sistemas de Información. Su línea de investigación son Bases de Datos, Minería de Datos Bodegas de Datos. Más información en http://148.204.45.159:9000/paginagil/
	El Dr. Adolfo Guzmán Arenas es Ingeniero en Comunicaciones y Electrónica de la Escuela Superior de Ingeniería y Mecánica y Eléctrica del Instituto Politécnico Nacional (IPN). Obtuvo su Maestría y su Doctorado en Ciencias de la Computación en el Instituto Tecnológico de Massachusetts (MIT), en Cambridge, Massachusetts, EE.UU. Se dedica a la computación, sobre todo diseñando soluciones a problemas recientes en procesamiento distribuido, sistemas de información y manejo de información no numérica, a menudo usando técnicas de Inteligencia Artificial. Fue profesor del Departamento de Ingeniería Eléctrica del MIT; del Departamento de Inteligencia Mecánica de la Universidad de Edimburgo; del Centro de Investigación y Estudios Avanzados del IPN (México), donde fundó la Maestría y Doctorado en Computación; del Instituto de Investigación en Matemáticas Aplicadas y Sistemas de la UNAM, donde fue Jefe del Departamento de Computación; y de la Unidad Interdisciplinaria (UPIICSA) del IPN. Fue Director del Centro Científico IBM para América Latina, IBM de México, S.A. Ha sido Investigador de la empresa MicroElectronics and Computer Corporation; Vicepresidente de Ingeniería en International Software Systems, y fundador y Presidente de SoftwarePro International, empresa en Austin, Texas, dedicada al desarrollo de paquetes comerciales y herramientas de Ingeniería de Software, el más reciente siendo BiblioDigital©, una biblioteca digital distribuida. Más información en http://alum.mit.edu/www/aguzman y en http://aguzman.blog.com