Modelo de Minería Incremental con Ajuste de Curvas

Martínez-Luna G. L*., Guzmán-Arenas A.*, Alexandrov-Aronovich M. **
Laboratorio de Sistemas de Información y Bases de Datos*
Laboratorio de Lenguaje Natural**
Centro de Investigación en Computación del I.P.N.
07738 México, México
5729-6000 ext 56600 Fax 5586-2936
lluna@cic.ipn.mx, a.guzman@acm.org, dyner@cic.ipn.mx

Resumen.

En este documento se presenta la propuesta de un Modelo de Minería Incremental donde se incluyen variables que ayudan a crear perfiles de interés y de utilidad a las solicitudes de Minería de Datos en una área de aplicación. El modelo se aplicara con la herramienta ANASIN Minería de Datos V.2.1, usando la técnica ajuste de curvas, se analizaran las solicitudes de minería anteriores y sus respuestas para contestar o mejorar la respuesta a una nueva solicitud. Considera los nuevos bloques de datos que llegan a la base de datos de minería y que afectan los análisis ya realizados, para completar o mejorar la respuesta (éxitos totales, parciales y posibles). Este modelo ayuda a disminuir la participación del usuario, diseñando al *software* para ejecutar tareas mecánicas y repetitivas. El *software* planeara el desarrollo de la minería al detener o continuar su ejecución, y si es necesario modificar su ambiente de trabajo, ya sea con una extracción de la base de datos de minería, o un diferencial de datos, o cambiar los parámetros que le permitan generar conclusiones (localizar registros y analizarlos), entre otras afectaciones.

Palabras claves. Minería de datos, Minería incremental, Ajuste de Curvas, Bases de Datos Relacionales, Bases de Datos Multidimensionales, Cubos de Datos.

1. Estado del Arte. Tiempo y Utilidad de Respuestas en la Minería de Datos.

En estos momentos el proceso de Minería de datos [1, pág. 10], se desarrolla en forma "manual" [2, pág. 2, 3, 6] y [3, Cap. 1, pág. 2], es decir el usuario tiene que indicar "situaciones interesantes de valor y desconocidas" que él "intuye", que existen en su base de datos y que el detectarlas pueden ayudar a la toma de decisiones. Las búsquedas y corroboraciones a estas situaciones, las realizan los programas de sistemas de minería; las cuales pueden tardar ya sea minutos, horas y días [4, pág. 1], [5, pág. 1], puesto que los análisis se desarrollan sobre bases de datos con grandes cantidades de registros, generalmente medidos en terabytes [1, pág. 2], [6, pág 333] y [7].

En algunos casos las búsquedas pueden "fracasar" al no hallar situaciones que se le solicitan y tardar en dar su respuesta, que posiblemente si se obtiene no sea útil (muy pocos registros a los esperados o regresar registros de una situación ya conocida).

Es claro el planteamiento la necesidad de resolver el problema de agilizar y hacer eficientes las búsquedas a las respuestas en este proceso por su relación con la toma decisiones. Este problema se ha atacado en varias formas; en especial en sistemas que permiten el uso de *OLAP* [siglas en inglés, ***On-line Analytical Processing***] y que utilizan el almacenamiento de datos llamado cubo de datos, como se indica en [8], [9], [10]; está estructura (Anasin V.2.1 la utiliza), es una forma de agilizar la respuesta, y se conoce como materialización de las consultas o precalculado de los valores de interés [4], [5] y [9]. Generalmente se materializan los obtenidos de las funciones de agregación del SQL [Structured Query Language] y otras que se definan y se requieran. El trabajo de almacenar y recuperar se puede dejar al Sistema Manejador de Bases de Datos Multidimensionales [SMBDM] como Arbor Essbase [11], o dejar a un Sistema Manejador de Bases De datos Relacional [SMBDR] lo realice, como lo realiza Dbminer [12, pág 2], apoyandose en SQL Server y Excel. Otra forma de agilizar la respuesta es con el uso de operadores especiales como en [8], [9] y [13] donde se definen estos operadores sobre los cubos de datos y por ser idonéos para esta organización se pueden eficientizar las búsquedas; otra forma es compactar en forma especial los cubos [14]; una forma más es cuando se organizan en forma especial para agilizar o tratar mejor las jerarquías y las dimensiones [4]; otros trabajos logran avances en este problema al utilizar avances en las tecnologías de redes, memoria, y uso de procesadores como se puede ver en [15, pág 1] y [16, pág 1].

Los problemas descritos podemos resumirlos como: i) A pesar de trabajos buscando la automatización [25], aún no hay herramientas automáticas de Minería de Datos, ii) El tiempo y la utilidad de la respuesta a una pregunta deben mejorarse, iii) Y el saber trabajar con los incrementos [5], [18] en los volúmenes de datos (decidir si analizar toda la base o el diferencial nuevo de datos), para identificar los

J. Díaz de León, G. González, J. Figueroa (Eds.): Avances en Ciencias de la Computación, pp. 73-82, 2003.
© IPN, México 2003.

cambios en las tendencias en los datos de la base y los cambios en las peticiones del usuario.

2. Propuesta de Solución.

En nuestra forma particular de atacar estos problemas, los programas que realizan los análisis, deben tener cierto nivel de autonomía, esto acorde a [1, pág. 2, 22] y [2, pág ixx, 2, 3, 6], Una nueva generación de herramientas de análisis, con autonomía en cuanto a tomar decisiones:

- Que ayuden a realizar estos análisis en tiempos más cortos [ver uso de directrices de tiempo, punto 4.2] o aseguren dar una respuesta útil [ver uso de directrices de espacio de búsqueda, número de registros aceptables, punto 4.2]. Aquí el usuario puede ayudar al indicar cuanto puede tardar a lo más la búsqueda de la respuesta y en observar el porcentaje de avance después de un tiempo definido.
- Autonomía con la capacidad de modificar los parámetros que ayuden a regresar respuestas no vacías [ver uso de similaridad con otras preguntas, punto 4.3]. El modificar los parámetros en el lenguaje que se le definen sus búsquedas, es afectar expresiones, ya sea en expresiones textuales que representen fórmulas [17], extensiones a consultas SQL, expresiones matemáticas [17] u otro lenguaje de minado de datos como el ***DMQL*** [***Data Mining Query Language***] de Dbminer [12, pág. 1] y el propuesto por [5, pág. 2], llamado ***MineSQL***.
- También, es deseable que estos programas tengan memoria o aprendizaje para recordar análisis que fracasan en su cometido [ver uso de directrices para almacenar y recordar resultados , punto 4.2], y mejor consuman recursos (tiempo, acceso a disco, procesador, memoria) en explorar nuevas alternativas que incrementen su eficiencia [ver uso de similaridad en preguntas, punto 4.3 y ver avances en 19].
- Y también puede que sea necesario cambiar las estructuras que almacenan los datos, las cuales facilitan los procesos para buscar y recuperar de datos (bases de datos multidimensionales o cubos, bajo el esquema estrella) [4], [6, Cap. 9].

La autonomía, el sensar y modificar el ambiente de trabajo, son algunas habilidades que se mencionan en trabajos sobre agentes, en especial a los programas que se llaman agentes de *software* [20], [21],[22], [23, cap 8]; y que es la tecnología que puede ayudar a modelar y automatizar este proceso; en nuestro caso se propone en punto 4.1 para:

- Dar autonomía en tomar la decisión al modificar su ambiente (expresión y espacio de búsqueda)
- Tener sensores para verificar si existe posibilidad de éxito o fracaso.
- Y tener efectores para realizar la modificación a la pregunta y a su espacio de trabajo.

En este trabajo se presenta la propuesta del Modelo de Minería Incremental con las innovaciones mencionadas en los dos parrafos anteriores y que se pretenden incorporar a la herramienta Anasin Minería de Datos V.2.1, e incluyen los datos que también se comentarón en los parrafos anteriores. Datos como la tendencia o curva búscada, el espacio de búsqueda, la variable tiempo, los éxitos a alcanzar, éxitos parciales, posibles éxitos, número de registros a minar, los nuevos volúmenes de datos y otras variables que se identifiquen.

La descripción del modelo, la iniciaremos con el **Modelo Actual de Anasin Minería de Datos V.2.1** en la sección 3, para continuar con la **Propuesta de Arquitectura para el Modelo de Minería Incremental** en la sección 4, analizaremos las **Situaciones a Resolver con el Modelo de Minería Incremental** en la sección 5 y terminaremos con las **Conclusiones** en la sección 6.

3. Modelo Actual de Anasin 2.1.

Este documento es la continuación de [24], por lo cuál resaltaremos elementos básicos de la herramienta diseñada y desarrollada por el Dr. Guzmán Arenas, ya descrita en [24], que busca curvas en datos históricos y los presenta para su interpretación. **Los elementos básicos** son **i)** la organización de los datos, **ii)** la manera de preguntar o expresar las tendencias, **iii)** la definición de un espacio de búsqueda, **iv)** la forma en que se localizan las curvas y **v)** el desplegado de los resultados para su análisis.

3.1 Arquitectura y Proceso de Anasin 2.1

La arquitectura de Anasin V.2.1 se puede ver en Fig. 1. El detalle de cómo utilizar esta herramienta esta en [17]. En breve su modelo es cliente-servidor, con horario de trabajo programado y etapas para llevar a cabo el proceso:

1. Definición del área y espacio de trabajo (**Módulo configurador**).
2. Creación y carga del cubo donde realizar minería (etapa no realizada por esta herramienta).
3. Actualización cíclica al cubo de datos (etapa no realizada por esta herramienta).
4. Definición de la pregunta; patrón a buscar y región de análisis o espacio de búsqueda (**Módulo Generador de Preguntas**).

5. Aplicación de la Minería (**Módulo Extractor y Módulo de Minería**).
6. Visualización de los Resultados (**Módulo Visualizador de Resultados**).

3. Arquitectura de ANASIN Minería de Datos 2.1

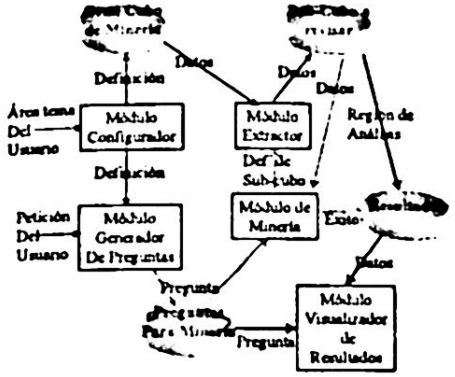

Figura 1.

3.2 Tres Elementos de Anasin 2.1.

La definición del **i) espacio de trabajo o las dimensiones, ii) los hechos a evaluar y iii) las jerarquías** se realiza con el **Módulo Configurador**, que tiene como entrada la definición de un esquema de base de datos (tablas) escrito en *SQL*. En las dimensiones y los hechos es donde se busca localizar las curvas o tendencias interesantes en el área tema de trabajo. La salida de este módulo es nuevamente un archivo *SQL*, pero con el esquema de la Base de Datos de Minería, con la definición de las dimensiones, hechos y jerarquías.

Cada dimensión se define con la creación de una entidad que almacena las claves concatenadas para formar un identificador o llave de cada dimensión, más un campo con su descripción:

Llave de una Dimensión... Descripción

$Clave_1$	$Clave_2$	...	$Clave_n$	Descripción
Valor	Valor	...	Valor	Descripción

La definición de los **hechos a evaluar** se realiza con la creación de una entidad que almacena las llaves de todas las dimensiones y el correspondiente valor para cada combinación de una instancia de las llaves y es de la forma:

Nombre-Hechos.

$Llave_1$	$Llave_2$	...	$Llave_k$	Hecho
Valor 1	Valor 2	...	Valor k	Valor

Esta última entidad, la podemos ver como **S el Espacio Total de Búsqueda**, que esta formado por puntos de dimensión k, con la forma:

$s=(s_1,s_2,...,s_k)$ ..... (1)

La jerarquía en cada una de las dimensiones [4, pág 7] de nuestro cubo se define con la creación una entidad que almacena la llave o identificador de una dimensión, los niveles en una jerarquía y los tamaños de cada de las claves concatenadas en una dimensión.

Niveles de Jerarquía.

Llave	No. De Claves concatenadas	$Tamaño_1$ $Clave_1$	...	$Tamaño_n$ $Clave_n$
Nombre Dimensión	n	$Tamaño_1$	...	$Tamaño_n$

La suma de los n tamaños de la claves concatenadas debe ser el tamaño de la llave completa o identificador de una dimensión.

Ejemplo 1. En las siguientes 3 tablas se puede ver un ejemplo de cada una de las 3 definiciones anteriores

TABLA 1. Dimensión Origen-Ingreso

In-greso	Edo.	Origen IPN/ Otro	Fuente	Forma	Descripción
1	00	0	0	0	Ingreso
1	09	1	0	0	Vocacional
1	09	1	0	1	Vocacional con examen
1	09	2	0	0	No Vocacio.
1	09	2	1	0	CECYT
1	09	2	1	1	CECYT con examen.
1	09	2	2	0	Bachilleres
1	09	2	2	1	Bachilleres con examen
1	09	5		00	Sup.Del IPN
1	09	5	0	1	Sup.con examen
1	09	6	0	0	No Superior del IPN
1	09	6	1	0	Facultad UNAM
1	09	6	1	1	Facultad UNAM examen

TABLA 2. Evaluación (Hechos).

Llaves Hecho	Registro 1	Registro 2	Registro 3
$Llave_1$	109101	109101	109101

$Llave_2$	000	000	000
$Llave_3$	00912	00912	00912
$Llave_4$	1994009001	1994009001	1994009001
$Llave_5$	001994000	001995000	001996000
Hecho	7.0	7.0	8.0

TABLA 3. Niveles de Jerarquía.

Tabla	Niveles	Tamaño subclaves
Origen-Ingreso	5	1,2,1,1,1
Egreso	1	1,1,1
Escuela-Carrera	2	2,3

La **Tabla 1** de una dimensión se tienen 5 subclaves; en la **Tabla 2**, en total se guardan 5 dimensiones y en la **Tabla 3** hay varias jerarquías en las 3 dimensiones que ahí se describen.

3.3 Una Pregunta en Anasin 2.1.

La formulación de una cuestión o pregunta se realiza con el **Módulo Generador de Preguntas.** La definición es denotada por Q_i con la forma:

$Q_i=(T_i, S_i)$.....(2),

Donde:

- T_i = Tendencia = curva = patrón = expresión
- S_i = **El sub espacio o cubo de búsqueda** y que sus puntos tienen la forma de (1) en 3.2.

La Tendencia T es de la forma:

$(v_1)or_1(v_2)\ [ol\ (v_n)or_2(v_m)]^*$.....(3)

Donde:

- **v** es una variable o expresión aritmética y el subíndice {1,2,n,m} son para diferenciar las variables.
- **or** es un operador de comparación {<, <=, =, >, >=} y el subíndice {1,2} son para diferenciar estos operadores de comparación.
- **ol** es un operador lógico {Y, O} y el subíndice {1,2} son para diferenciar estos operadores lógicos.
- * puede repetirse cero, una o más veces lo que esta entre [ ].

El sub espacio S_i es donde se cree existe la tendencia T_i, es un sub-cubo, donde los valores de cada eje o dimensión se definen con los operadores:

- '.' = Un rango.
- '|' = Valores discretos, el operador.
- '*' = La jerarquía

La sintaxis es:

- $S_1Valor_0.S_1Valor_1$, que define un rango
- $S_2Valor_0\ [|S_2Valor_m]^m$, que define valores discretos y m indica que se puede repetir m veces
- S_3Valor^*, se toman los valores bajo una jerarquía.

Ejemplo 2. La Tendencia de crecimiento en 4 puntos del tiempo o mantenerse constante es:

(a)<=(b) Y (b)<=(c) Y (c)<=(d).

Puede significar, buscar alumnos que durante los años sus promedios fueron mejorando o mantuvieron.

Ejemplo 3. El definir el subespacio en un cubo de dimensiones se realiza de la siguiente forma:

- **Eje 1.** * = Seleccionar todo tipo de origen ingreso al I.P.N.
- **Eje 2.** * = Seleccionar sin importar si ya o terminaron.
- **Eje 3.** 00912|00913 = Seleccionar solo de escuela 009 y las carreras 12 y 13.
- **Eje 4.** * = seleccionar todos las matrículas.
- **Eje 5.** 00199400000.00199700000 = Seleccionar solo los promedios desde el Año 1994 al año 1997.

3.4 El Ajuste de Curvas en Anasin 2.1.

El Módulo de Minería es el que busca responder preguntas. El proceso de responder una pregunta, $Q_i=(T_i, S_i)$ como se definió en la sección 3.3, consta de 4 pasos, y son:

Paso 1) De la pregunta Q_i, se toma la definición sub-cubo o sub-espacio S_i y con ayuda **Módulo Extractor**, se extrae de S todos puntos que cumplen las condiciones de S_i. decir cada para cada eje o dimensión se busca el rango, valor o valores de una jerarquía para reunir la serie de puntos que consta S_i (cada punto tienen la forma $s_i = (s_{i,1}, s_{i,2}, ..., s_{i,k})$.

Paso 2) Se organizan los puntos de S_i tal que pueda buscar la tendencia T_i, generalmente organizados con respecto al tiempo.

Paso 3) Se guarda el espacio analizado ó S_i, con nueva organización de análisis.

Paso 4) Se guardan los puntos s_i donde se cumple tendencia T_i, en especial donde inicia esta los siguientes puntos que completan tendencia, que podríamos denotarlo por E_i.

Ejemplo 4. Si la tendencia T_i es buscar crecimiento en 3 puntos, y el sub-espacio incluye los puntos de la **Tabla 2**, parte resultado con éxito E_i serán los puntos de columnas 2,3 y 4 de esta tabla.

3.5 Presentación de Resultados.

Nuestra presentación de resultados se realiza con el **Módulo Visualizador de Resultados** y es mostrando los resultados ordenados como se organizan en el paso 2 de **3.4**, tanto en forma de texto como gráfica, como se ilustra en [17] y [24].

4. Propuesta del Modelo de Minería Incremental.

4.1 Funcionalidades del Nuevo Modelo.

Basándose en la sección 2, para automatizar este proceso de minería de datos se plantean las siguientes 4 premisas:
Premisa 1. "Regresar respuesta útil".
Premisa 2. "El usuario siempre puede indicar directrices o restricciones":

- Para saber o intuir si existe la situación interesante.
- Si los resultados serán útiles,
  - Acotando el tiempo de búsqueda.
  - Indicando una relación entre registros a buscar y los éxitos"

Premisa 3 "Aprovechar resultados anteriores" = memoria + estadísticas.
Premisa 4 "Utilizar el diferencial de datos" = Nuevos arribos de datos.

Lo que nos lleva a modificar la arquitectura de Anasin V.2.1 agregandole un nuevo módulo y varias bases de datos de apoyo. El nuevo módulo es el **Afinador** que monitorea el proceso. Un primer esquema de la propuesta de trabajo se puede ver en la Fig. 2.

4. Arquitectura Propuesta

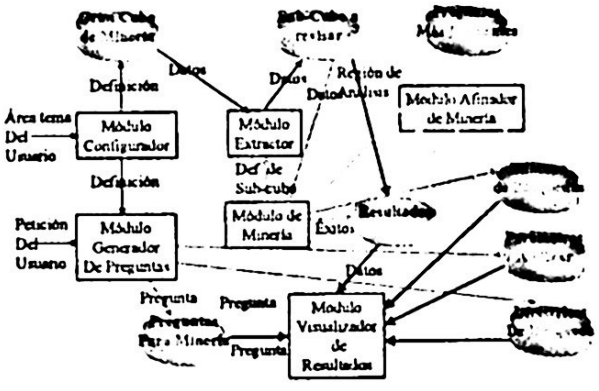

Figura 2.

En este análisis y diseño se modelan los cambios necesarios para lograr las premisas 1-4. La nueva arquitectura, nos muestra donde se realizarán los cambios, en principio en los módulos **Configurador** y **Minero**, además de crearse el módulo **Afinador** y que llevara una carga fuerte de trabajo para esta nueva organización. Esto se observa por la relación con las nuevas bases de datos a crearse, que son **4** y que son **Preguntas más frecuentes, Bitácora de Búsquedas, Parámetros a afinar y Directrices de Búsquedas.**

Las actividades detectadas para el **Afinador** son:

- Actualizar los resultados de las cuestiones Q_i con los diferenciales de datos (nuevos datos) D_k.
- Para una Q_j identificar las Q_i **"similares" (ver sección 4.3)**, para saber si utilizar resultados R_i.
  - **"Similar"** en cuanto la tendencia T_j con la Tendencia almacenda T_i.
  - **"Similar"** en cuanto el espacio de búsqueda S_j con espacio de búsqueda S_i.
  - Es decir en grado esta contenida (o hay intersección) de la respuesta a Q_j o en que grado puede ayudar otra Q_i a resolver Q_j.
- Indicar lo anterior al **Módulo Minero**, para realizar el proceso de ajuste de curvas o minería.
- Usar las directrices de tiempo permitido de ejecución o búsqueda e indicarselo al Módulo de Minería.
- Usar las directrices de porcentaje de avance, para ser válidos los resultados e indicarselo al Módulo de Minería.
- Actualizar la mayoría de las bases de datos que se observan en la nueva arquitectura.

En este **Módulo Afinador**, es donde se observa que puede ser necesario construir **Agentes de *Software***, por las actividades de monitoreo y toma de decisión a continuar o no con las búsquedas en turno, llevadas a cabo por el **Módulo de Minería.**

Las tareas nuevas identificadas para el **Módulo Configurador** son:

- Capturar las directrices de tiempo de ejecución.
- Capturar las directrices de avance en registros.
- Capturar directrices para decidir que resultados almacenar.
- Y Si usar los resultados anteriores.

Y una nueva tarea identificada para el **Módulo de Minería** es almacenar los resultados y estadísticas.

4.2 Redefinición de una Pregunta para la Minería Incremental.

Revisemos los elementos enunciados en las nuevas actividades y escribamos nuestra nueva modelación. **S** sigue siendo nuestro Espacio de Búsqueda, pero Ahora una cuestión o pregunta se define como:

$$Q_i = (T_i, S_i, C_i, M_i, U_i) \quad .....(4)$$

Donde T_i, S_i son como se definio en (2) de 3.3, pero **los nuevos parámetros son:**

- C_i, las directrices o restricciones en la búsqueda en tiempo y espacio de búsqueda.
- M_i, que resultados almacenar para recordar como memoria
- U_i, indicará si se utilizan los resultados anteriores, se revisa todo el subcubo u otra combinación.

Veamos ahora que valores puede dar el usuario a guardar en las directrices. En las directrices de búsqueda en $C_i = \{t_i, u_i, st_i, s_i, e_i, p_i, c_i\}$ se definen como:

- t_i = tiempo permitido de ejecución.
- u_i = unidades del tiempo {s, m, h, d} = {segundos, minutos, horas, días}.
- st_i = {0, 1}, 1 se toma como referencia el número de registros del total del espacio.
- s_i = {0, 1}, 1 se toma como referencia el número de registros del espacio en que se buscara.
- e_i = [dddd] o [0-100]:
  - Si se tiene 0, dddd debe ser un número menor que la cardinalidad de S o un número menor que la cardinalidad de S_i.
  - En porcentaje si se tiene 1 en p_i.
- p_j = {0,1}:
  - 0 si se desea manejar número de registros en e_j con respecto a S o S_j.
  - 1 si se desea manejar porcentaje en la relación de e_i con S o con S_i.
- c_i = {0,1}:
  - 0 si se desea que se detenga cuando se cumpla e_i en st_i o s_i en el tiempo t_i según p_i.
  - 1 si se desea que se continue respondiendo la cuestión Q_j cuando se cumpla e_i en st_i o s_i en el tiempo t_i según p_i.
  - Si no se cumple e_i en st_i o s_i en el tiempo t_i según p_i el proceso se detiene.

Las indicaciones de que guardar en memoria o que resultados almacenar se indican en $M_j = \{m_i, ri_i, rf_i\}$ con:

- m_i,= {0,1}, con 1 se guardan los resultados, tanto éxitos totales, parciales y posibles que se indican en ri_i, rf_i.
- ri_i = [0, n], donde n es un número menor que el tamaño de T_i (número de puntos en el tiempo).
- rf_i = [0, m], donde m es un número mayor que el tamaño de T_i (número de puntos en el tiempo).

Y los resultados a utilizar se indica en $U_i = \{r_i, d_i\}$ con:

r_i = {0, 1}, 1 indica que se usen los resultados anteriores de otras preguntas ya contestadas.

d_i = {0, 1}, 1 indica que se use el D_k con los resultados anteriores de otras preguntas ya contestadas. El uso de d_i no se aplicará en esta propuesta, queda pendiente, para otro documento.

Como trabajan las combinaciones de las directrices de tiempo y porcentaje de avance, se puede observar en la Tabla 4:

Tabla 4.

No.	t_i	u_i	st_i	s_i	e_i	p_i	c_i	Descripción de cómo trabaja el Módulo de Minería
1	X	{s, m, h, d}						Solo trabaja el tiempo t_i
2	X	{s, m, h, d}			X	0	0	Solo trabaja tiempo t_i, o se detiene al hallar los éxitos indicados en e_i
3	X	{s, m, h, d}			X	0	1	Solo trabaja tiempo t_i, y continuará si halla los éxitos indicados en e_i, en el tiempo t_i
4	X	{s, m, h, d}	1		X	1	0	Solo trabaja tiempo t_i, o se detiene al hallar un porcentaje de éxitos con respecto a S, porcentaje indicado en e_i en el tiempo t_i
5	X	{s, m, h, d}	1		X	1	1	Solo trabaja tiempo t_i, y continuará al hallar un porcentaje de éxitos con respecto a porcentaje indicado en e_i en el tiempo t_i
6	X	{s, m, h, d}		1	X	1	0	Solo trabaja tiempo t_i, o se detiene al hallar un porcentaje de éxitos con respecto a porcentaje indicado en e_i en el tiempo t_i
7	X	{s, m, h, d}		1	X	1	1	Solo trabaja tiempo t_i, y continuará al hallar un porcentaje de éxitos con respecto a S_i, porcentaje indicado en e_i en el tiempo t_i

En esta Tabla 4 no importa que valores tengan opciones que estan en blanco. Se buscará cumplan cualquiera de estas 7 combinaciones y tratara de resolver en la búsqueda de la respuesta la cuestión. La manera que indican las directrices para la memoria estan en la Tabla 5.

Tabla 5.

No.	m_i	ri_i	rf_i	Descripción
1	0			No se almacenan resultados.
2	1	N	m	Se almacenan los éxitos y los resultados parciales entre [n,m].

Para nuestro modelado, se supone que no hay problemas con el espacio en disco.

$Q = \{Q_1, Q_2, ..., Q_n\}$ = Conjunto de preguntas en S.
T_i = Expresión que define una tendencia (definida anteriormente).
S_i = Sub-espacio o sub-cubo donde el usuario definió buscar a T_i.
R_i = Resultados de buscar T_i en S_i, $R_i=(X_i, Y_i, Z_i, t_i, V_i, P_i)$.
X_i = Puntos de dimensión k, donde se cumple exactamente la tendencia buscada.
Y_i = Puntos de dimensión k, donde se cumple **parcialmente** y posiblemente nunca se cumpla la tendencia buscada. **Ejemplo 5:** se busca tendencia de crecimiento con 4 puntos, pero se hallo un crecimiento con 3 puntos, pero ya no se modificará este crecimiento en la base de análisis.
Z_i = Puntos de dimensión k, donde hay **posibilidad** que se cumpla la tendencia buscada, con el arribo de nuevos datos (D_k). **Ejemplo 6:** se hallo un crecimiento con 3 puntos, pero con un arribo de nuevos datos, es posible que se complete un crecimiento de 4 puntos.
t_i = Tiempo en obtener los resultados.
V_i = Total de puntos del Espacio.
P_i = Total de puntos del espacio de búsqueda o sub-cubo a revisar.

Ejemplo 7. Una pregunta pueden tener las siguientes directrices:

No.	t_i	u_i	st_i	s_i	e_i	p_i	c_i	Descripción de cómo trabaja el Módulo de Minería
1	10	M						Trabaja 10 min.
2	10	M			100	0	0	Trabaja máximo 10 min., o se detiene al hallar 100 éxitos.
3.	10	M			100	0	1	Trabaja 10 min., o continuará si halla 100 éxitos antes de 10 min.
4	10	M	1		10	1	0	Trabaja máximo 10 min., o se detiene al hallar un 10% de éxitos con respecto al total de registros en S.
5	10	M	1		10	1	1	Trabaja 10 min, o continuará si halla un 10% de éxitos con respecto al total de registros en S.
6	10	m		1	10	1	0	Trabaja máximo 10 min., o se detiene al hallar un 10% de éxitos con respecto al total de registros en S_i.
7	10	m		1	10	1	1	Trabaja 10 min, o continuará si halla un 10% de éxitos con respecto al total de registros en S_i.

4.3 Similaridad en Preguntas.

La **igualdad y similaridad** entre preguntas se puede definir de la siguiente forma:

Dos preguntas son **iguales** si dadas Q_i y Q_j que pertenecen a Q, con $Q_i = (T_i, S_i, C_i, M_i, U_i)$ y $Q_j = (T_j, S_j, C_j, M_j, U_j)$, y se cumple que $T_i = T_j$, $S_i = S_j$; por lo tanto los resultados $R_i = R_j$

- Esto último puede ser erróneo, si el espacio ha sido afectado con una alimentación de datos o diferencial de datos por un nuevo período de tiempo. No se modelará este caso.
- Aquí se debe analizar que tanto se pueden usar los resultados de una pregunta Q_i para responder otra pregunta Q_j.

Dadas Q_i y Q_j que pertenecen a Q, $Q_i = (T_i, S_i, C_i, M_i, U_i)$ y $Q_j = (T_j, S_j, C_j, M_j, U_j)$, Q_i y Q_j son **similares**, si:

- En alguna sección dos variables y un operador relacional de las expresiones que definen a T_i y T_j, estas coinciden con el relacional.
- Existe una intersección en las definiciones de los subcubos S_i y S_j, como:
  - $S_i < S_j$ = el i-ésimo cubo esta contenido en el j-ésimo.
  - $S_i > S_j$ = el i-ésimo cubo contiene al j-ésimo.
  - $S_i \text{ Int } S_j <> 0$ = no se contienen, pero su intersección no es vacía.

Con lo cual los resultados de una pregunta pueden servir para responder la otra pregunta, es decir:

- $R_i < R_j$ = los i-ésimos resultados están contenidos en los j-ésimos.
- $R_i > R_j$ = los i-ésimos resultados contienen a los j-ésimos.
- $R_i \text{ Int } R_j <> 0$ = no se contienen, pero su intersección no es vacía.

Ejemplo 8. Las expresiones que buscan crecimiento en 2 puntos y la que busca crecimiento en 3 puntos coinciden en 2 variables y un operador de relación.

- (e<f) Crecimiento en 2 puntos.
- (a<b)Y(b<c) Crecimiento en 3 puntos.

Ejemplo 9. Hay una intersección o contención en dos cubos de búsqueda, si al comparar los rangos de los ejes de los cubos en algunos de los ejes, en estos hay una intersección, sin importar que suceda en los otros ejes.

- Eje n de S_i < Eje n de S_j.
- Eje n de S_i > Eje n de S_j.
- Eje n de S_i int Eje n de S_j <> 0.

5. Situaciones a Resolver con el Modelo de Minería Incremental.

Con el detalle de la sección 4, todas las situaciones posibles a tratar por los mineros se encuentran en la Tabla 6, pero antes de mostrar esta tabla, definamos algunos términos más, para comprender mejor esta matriz:

D_k es el diferencial de datos que arribaron. Existe una base de datos de diferenciales de datos, organizados por fecha, no se usarán en este modelado.

SD_k son diferentes versiones de espacio de total búsqueda, secuencialmente afectadas por D_k, existe una base de diferentes versiones de espacio de búsquedas, organizados por fecha.

Q_i será una pregunta ya contestada.

Q_j es una nueva pregunta a contestar.

Los resultados de las preguntas Q_i se actualizan cuando llega un nuevo diferencial de datos D_k, pero se conservan los resultados de la Q_i cuando se realizo anteriormente (por fecha, indica cuando se contesto). Las preguntas van acompañadas por las directrices definidas en **4.2.**

Para las tendencias estan las siguientes situaciones:

$T_j = T_i$, significa tendencias iguales.

$T_j < T_i$, significa tendencia T_j más corta que T_i.

$T_j > T_i$, significa tendencia T_i más corta que T_j.

$T_j <> T_i$, significa tendencia T_i diferente que T_j.

C_j = en todos los casos aplicar las directrices si se tienen, para reducir el tiempo de búsqueda y regresar resultados útiles.

R_j = en todos los casos guardar resultados.

Las situaciones que se tienen identificadas, su acción a realizar y su posible ahorro están en la Tabla 6, con la siguiente notación por columnas:

- **No.** = Número de caso.
- **Ten.** = Relación entre tendencias.
- **Sub.** = Relación entre subespacios de búsqueda.
- **Acción** = Realizar proceso o usar resultado anterior R_i.

Tabla 6.

No.	Ten.	Subespacios	Acción
1	$T_j = T_i$	$S_j = S_i$	Usar S_i, 0 ahorro.
2	$T_j = T_i$	$S_j = S_i$	Regresar R_i, ahorro tiempo t_i.
3	$T_j = T_i$	$S_j < S_i$	Usar S_i, 0 ahorro.
4	$T_j = T_i$	$S_j < S_i$	Usar R_i, debe calcularse el ahorro$_i$.
5	$T_j = T_i$	$S_j > S_i$	Usar S_i, 0 ahorro.
6	$T_j = T_i$	$S_j > S_i$	Usar $R_i + D_k$, debe calcularse el ahorro$_i$.
7	$T_j = T_i$	S_j int S_i no es vacia	Usar S_j, 0 ahorro.
8	$T_j = T_i$	S_j int S_i no es vacia	Usar R_i, debe calcularse el ahorro$_i$.
9	$T_j = T_i$	S_j int S_i es vacia	Usar S_j, 0 ahorro.
10	$T_j < T_i$	$S_j = S_i$	Usar S_i, 0 ahorro.
11	$T_j < T_i$	$S_j = S_i$	Usar R_i, debe calcularse el ahorro$_i$.
12	$T_j < T_i$	$S_j < S_i$	Usar S_i, 0 ahorro.
13	$T_j < T_i$	$S_j < S_i$	Usar R_i, debe calcularse el ahorro$_i$.
14	$T_j < T_i$	$S_j > S_i$	Usar S_i, 0 ahorro.
15	$T_j < T_i$	$S_j > S_i$	Usar $R_i + D_k$, debe calcularse el ahorro$_i$.
16	$T_j < T_i$	S_j int S_i no es vacia	Usar S_j, 0 ahorro.
17	$T_j < T_i$	S_j int S_i no es vacia	Usar R_i, debe calcularse el ahorro$_i$.
18	$T_j < T_i$	S_j int S_i es vacia	Usar S_j, 0 ahorro.
19	$T_j > T_i$	$S_j = S_i$	Usar S_i, 0 ahorro.
20	$T_j > T_i$	$S_j = S_i$	Usar R_i, debe calcularse el ahorro$_i$.
21	$T_j > T_i$	$S_j < S_i$	Usar S_i, 0 ahorro.
22	$T_j > T_i$	$S_j < S_i$	Usar R_i, debe calcularse el ahorro$_i$.
23	$T_j > T_i$	$S_j > S_i$	Usar S_i, 0 ahorro.
24	$T_j > T_i$	$S_j > S_i$	Usar $R_i + D_k$, debe calcularse el ahorro$_i$.
25	$T_j > T_i$	S_j int S_i no es vacia	Usar S_j, 0 ahorro.
26	$T_j > T_i$	S_j int S_i no es vacia	Usar R_i, debe calcularse el ahorro$_i$.
27	$T_j > T_i$	S_j int S_i es vacia	Usar S_j, 0 ahorro.
28	$T_j <> T_i$	Cualquier combinación	Usar S_j, 0 ahorro.

6. Conclusiones.

Se reafirma lo mencionado en [19] que es:

- Este es un planteamiento especial para la minería dirigida, y en específico al tipo de ajuste de curvas o búsqueda de un comportamiento a través del tiempo y que puede funcionar para automatizar el proceso de minería.
- El usuario o el dueño del proceso de minería no se le puede eliminar, pero si reducir su participación en las primeras etapas del proceso y en especial cuando ya tienen realizadas "suficientes" preguntas.
- Una vez formada una base de datos de preguntas, un número considerable, las siguientes preguntas podrán ser monitoreadas por los resultados de las anteriores.
- El perfil de los tendencias interesantes se reflejara en la base de datos de preguntas y sus resultados; su monitoreo será de limitar y redireccionar la búsqueda con el fin de obtener los mejores resultados.
- Dependiendo del perfil y las frecuencia de similaridad entre las preguntas, será el ahorro en tiempo y la eficiencia de las respuestas.
- Se puede construir una herramienta planteada acorde a [1].
- Este es otro planteamiento para la Minería de Datos Incremental.

Los elementos importantes en esta propuesta se podrían considerar:

- Las directrices o restricciones o condiciones.
- La modificaciones o afinaciones (cambiar las preguntas).
- Los resultados a almacenar (memoria del modelo).
- El trabajo del **Módulo Afinador** que utiliza los resultados anteriores, así, como el empleo de diferenciales de nuevos registros en la base de minería.

Con esto se observa que se ha modelado para definir tendencias interesantes en una base de datos y usar los resultados que en función de la historia han sido los aceptados y útiles.

Referencias y Bibliografía.

[1] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, *"Advanced in Knowledge Discovery and Data Mining"*, American Association for Artificial Intelligence, The MIT Press, 1996.

[2] Ian H. Witten, Eibe Frank, *"Data Mining, Practical Machine, Learning Tools and Techniques with JAVA implementations"*, Morgan Kaufmann Publishers, 2000.

[3] Thuraisingham B., *"Data Mining, Technologies, Techniques, Tools, and Trends"*, CRC press, 1999

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

[5] TadeuszMorzy, Marek Wojciechowsky, Maciej Zakrzewicz, *"Materialized Data Mining Views"*, Poznan University of Technology, Institute of Computing Science.

[6] Berson A., Smith S., *"Data Warehousing, Data Mining, & OLAP"*, Mc Graw-Hill, 1997

[7] http://www.wintercorp.com/VLDB/2000_VLDB_Survey/winners/index.html

[8] Jim Gray, Adam Bosworth, Andrew Layman, Hamid Pirahesh, *"Data Cube: A Relational Aggregation Operator Generalizing Group By, Cross-Tab, and Sub-Totals"*, Technical Report MSR-TR-95-22, Microsoft Research, Advanced Technology Division, Microsoft Corporation.

[9] Rakesh Agrawal, Ashish Gupta, Sunita Sarawagi, *"Modeling Multidimensional Databases"*, IBM Almaden Research Center.

[10] Ching-Tien Ho, Rakesh Agrawal, Nimrod Megiddo, Ramakrishnan Srikant, *"Range Queries in OLAP Data Cubes"*, IBM Almaden Research Center, 650 Harry Road, San José, CA 95120.

[11] White paper "The Role of the OLAP Server in a Data Warehousing Solution", http://www.essbase.com/download_files/resource_library/white_papers/olap_in_a_data_warehousing_solution.pdf

[12] Han J., *"DBMiner: A System for Mining Knowledge in Large Relational Databases"*, Data Mining Research Group, Database Systems Research Laboratory, School of Computing Science, Simon Fraser University, British Columbia, Canada

[13] Gupta A., Harinarayan V., Quass D. *"Aggregate-Query Processing In Data Warehousing Enviroments"*, IBM Almaden Research Center, Stanford University

[14] Jayavvl Shanmugasundaram, Usama M. Fayyad, Paul S. Bradley, *"Compressed Data Cubes for OLAP Aggregate Query Approximation on Continuous Dimensions"*, Microsoft Research, University of Wisconsin, June 1999, Technical Report MSR-TR-99-13.

[15] Rakesh Agrawal, John C. Shafer, *"Parallel Mining of Association Rules"*, IBM Almaden Research Center, 650 Harry Road, San José, CA 95120.

[16] David W. Cheung, Vincet T. Ng, Ada W. Fu, Yongjian Fu, *"Efficient Mining of Association Rules in Distributed Databases"*, IEEE Transaction on Knowledge and Data Ingineering, Vol 8, No. 6, December 1996.

[17] Laboratorio de Sistemas de Información y Bases de Datos, "Manual de Usuario, ANASIN Minería de Datos, Versión 2.1", CIC-1998

[18] Vnekatesh Ganti, Johannes Gehrke, Raghu Ramakrishnan, "*Minig Data Streams under Block Evolution*", Micrsoft Research, Cornell University, UV-Madison.

[19] Martínez G, Albores Y, Castillo C, "Automatización del Proceso de Mineria de Datos", 2001, Memoria del 2do. Foro "Computación de la Teoría a la Práctica", Ago-2001, Canacintra CIC-IPN.

[20] Martínez G., "Instalador Automático de Sistemas, Agentes de Software con una Arquitectura de Pizarra", Sección de Computación, Departamento de Ingeniería Electrica, CINVESTAV-1998, Tesis de Maestría.

[21] Martínez G., Núñez G., "Instalación de Agentes Móviles, utilizando una Arquitectura de Pizarra", Laboratorio de Sistemas de Información y Bases de Datos, Centro de Investigación en Computación CIC-IPN, 1999

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

[23] Bigus J.,. "*Data Mining With Neural Networks* ", McGraw-Hill, 1996.

[24] Martínez L. G. L., Guzmán A. A., Mikhail Alexandrov, "Modelo de Minería Datos con Ajuste de Curvas", en revisión por comité de evento CIICC-2003, http://www.sd-cenidet.com.mx/~ciicc03/.

[25] Noguerón G. C., Guzmán A. A., "Buscador automático de Situaciones Interesantes en la minería de Datos", "3er. Taller de Minería de Datos", Universidad Panamericana CIC-IPN, ITZ, 2001.