II WORKSHOP DE INGENIERÍA DE SOFTWARE Y BASES DE DATOS (WISBD)

Arquitectura conceptual para combinar los procesos de Data Warehousing y Data Mining basada en Objetos Simbólicos

Sandra González Císaro, Héctor Oscar Nigro, Daniel Xodo
INCA/INTIA. Departamento de Computación y Sistemas, Facultad de Ciencias Exactas. Universidad Nacional del Centro de la Pcia de Bs. As.
Campus Universitario, Paraje Arroyo Seco S/n Tandil, Buenos Aires, Argentina
TE +54-2293-432466 – FAX: +54-2293-440363
E-mail: { sagonci, onigro, dxodo}@exa.unicen.edu.ar

Resumen

Este trabajo presenta una arquitectura conceptual para la combinación de los procesos de Data Warehousing con el Data Mining por medio de objetos simbólicos.

En los últimos años, las empresas han recopilado una cantidad muy importante de datos, es deseable organizarlos para coordinar las tareas de análisis con la intención de mejorar los Procesos de Toma de decisiones. La organización de datos es realizada con la implementación de un Data Warehouse. En el cual, la información es seleccionada, limpiada y enriquecida; debido a ello es posible integrar varias fuentes e incluir el conocimiento propio del negocio, también llamado conocimiento contextual.

De este punto de vista, extraer el conocimiento potencialmente valioso de los volúmenes masivos de datos coleccionados por sistemas operacionales es un desafío siendo modelado por objetos simbólicos. Los cuales, representan los principales conceptos que definen el negocio u organización.

De esta manera, mejoramos la Gestión del Conocimiento, ya que el conocimiento implícito en las mentes de los miembros de la organización es transformado en explicito bajo el formalismo de objetos simbólicos.

Palabras clave: Bases de Datos, Data Warehouse, Data Mining, Objeto Simbólico, Análisis de Datos Simbólicos, Gestión del Conocimiento, Sistemas de Apoyo a la Toma de Decisiones.

Introducción

La importancia de la información en las empresas pensada como un capital que puede dar utilidades (Davenport et al., 1998), está en la base de la toma de decisiones en la organización. Razón por la cual el proceso de almacenamiento orientado al análisis y el posterior análisis de esa información adquiere relevancia en la actualidad.

Las bases de datos existentes contienen gran cantidad de información, la cual no siempre es presentada al nivel de agregación necesario para los procesos de toma de decisiones. Dependiendo de la evolución tecnológica que se presente en la empresa, el tipo de requerimientos de información, características propias del negocio; algunas organizaciones tienen implementado el uso de una base de datos central llamada Data Warehouse, sobre la cual se desarrollan las tareas de análisis de la información a través de sistemas front end: Olap, Data Mining o paquetes estadísticos profesionales.

Los sistemas Olap resuelven el problema de presentar distintos niveles de agregación y visualización de los datos a través del paradigma del cubo. Las técnicas clásicas de análisis de datos

(análisis factorial, regresión, dispersión, etc.) son aplicadas a individuos, tuplas o individuos en bases de datos transaccionales u operativas. Los objetos clásicos de análisis no resultan lo suficientemente expresivos para contener distribuciones, reglas lógicas, atributos multivaluados e intervalos. Simultáneamente es necesario respetar la variación interna y taxonomía de la información, manteniendo el dualismo entre individuo y clase.

Por consiguiente, necesitamos un nuevo tipo de dato que sea capaz de contener estas características. Este es justamente el modelo matemático de concepto introducido por Diday, llamado objeto simbólico. Formalmente un objeto simbólico s es una tripla s = (a, R, d) donde R es una relación entre descripciones, d es una descripción y a es una correspondencia definida desde Ω (universo del discurso) en L dependiendo de R y d (Diday, 2003).

Los objetos simbólicos nos permiten modelar entidades físicas o conceptos del mundo real. Los primeros son las tuplas almacenadas en las bases de datos, los segundos son entidades de más alto nivel obtenidas por agregación, clasificación automática, análisis de distintos expertos o alguna agregación en particular tomada de la unidad de análisis (Bock & Diday, 2000; Diday & Billard, 2002; Diday, 2004).

La integración de datos es un problema central en el diseño de Data Warehouses y Sistemas de Apoyo a la Toma de Decisiones DSS (Calvanese, 2003), con esta idea preparamos una arquitectura basada en objetos simbólicos, la cual nos permite la integración mediante la combinación de dos procesos claves en el área de toma de decisiones: Data Warehousing y Data Mining. El beneficio principal de esta arquitectura es la posibilidad de realizar análisis exploratorio en un ambiente integrado de trabajo, bajo una misma interfase.

Las principales ventajas que posibilita el uso de objetos simbólicos son: -preserva la confidenciabilidad de la información, -soporta el lenguaje inicial en el que ellos fueron creados y permite compartir conceptos entre bases de datos. Independizándose de la tabla inicial de la que fueron creados, son capaces de identificar algunas coincidencias en individuos en otras tablas (Bock & Diday, 2000). El resultado de trabajar con entidades de más alto nivel llamadas conceptos necesariamente descriptos por datos más complejos extiende el Data Mining a Knowledge Mining (Diday 2004).

El concepto de objeto simbólico es muy importante para la construcción del Data Warehouse y es un importante desarrollo para Data Mining, especialmente para la manipulación y análisis de información agregada (Nigro & González Císaro, 2005).

Conceptos Fundamentales sobre Objetos Simbólicos

Teniendo en cuenta la definición de Gowda: “los objetos simbólicos son extensiones de los tipos de datos clásicos y son definidos por una conjunción de eventos que unen valores y variables en las cuales las variables pueden tomar uno o varios valores, y todos los objetos simbólicos no tienen que ser definidos sobre las mismas variables” (Gowda, 2004); consideramos que los objetos simbólicos constituyen un nuevo tipo de dato para datos complejos. Dado que definen un álgebra en el Análisis de Datos Simbólico (ADS), la cual determina un conjunto de operadores básicos y un espacio de trabajo propio sobre ellos.

Como fue mencionado en la sección introductoria, un objeto simbólico puede modelar a un individuo o una clase manteniendo su taxonomía y variación interna. En realidad, representamos un concepto por su descripción intencional, es decir los atributos necesarios para caracterizar al fenómeno estudiado o unidad de análisis; su descripción nos permite la distinción entre ellos.

Las características claves enumeradas por Gowda que hacen del objeto simbólico datos complejos son (Gowda, 2004):

- Todos los objetos de un conjunto de datos simbólico puede no ser definido sobre las mismas variables.
- Cada variable puede tomar más de un valor o hasta un intervalo de valores.
- La descripción del objeto simbólico puede depender de las relaciones que existen entre otros objetos.
- Los valores alcanzados por las variables pueden tener valores típicos, los cuales indican la frecuencia de acontecimiento, probabilidad relativa, el nivel de la importancia de los valores, etcétera.

Esencialmente, hay dos tipos de objetos simbólicos (Diday, 2002):

- *Objetos Simbólicos Booleanos*: el caso de una relación binaria entre el descriptor del objeto y el dominio de definición, que es definida para tener valores verdaderos o falsos. Si [y (w) R d] ∈ {verdadero, falso} es un Objeto Simbólico Booleano (Diday & Billard, 2002). Ejemplo: s = (modo-de-pago ∈ {bueno; regular}), aquí describimos a un individuo / clase de cliente cuyo modo de pago es bueno o regular.
- *Objetos Modales Simbólicos*: En algunas situaciones, no podemos decir verdadero o falso, tenemos un grado de pertenencia, o alguna imprecisión lingüística como siempre verdadero, a menudo verdadero, a mitad, a menudo falso, siempre falso; ahora decimos que la relación es difusa. Si [y (w) R d] ∈ L = [0,1] es un Objeto Simbólico Modal (Diday & Billard, 2002; Diday, 2003). Ejemplo: s = (modo-de-pago ∈ [(0,25) bueno; (0,75) regular]), en esta instancia describimos a un individuo / clase de cliente cuyo modo de pago es: 25 % bueno; 75 % regular. En este caso, los pesos corresponden a una probabilidad (frecuencia relativa), pero ellos podrían representar también posibilidades, capacidades, creencias (detalles en Diday & Billard, 2002).

Podemos trabajar con objetos simbólicos de dos formas: - conociendo los valores de sus atributos entonces queremos saber a que clase ellos pertenecen, - formamos una clase a partir de la generalización / especialización de los valores de los atributos de unos individuos. A la primera práctica se la denomina *inducción* y la segunda *generalización.*

Diday llama "*aserción*" a un caso especial de un Objeto Simbólico definido por s = (a, R, d) donde R es definido por $\wedge_{i=1,\,p}$ [d'$_i$ R_i d_i], donde "∧ " tiene el sentido lógico estándar y a es definido por: a(w) = $\wedge_{i=1,\,p}$ [y_i (w) R_i d_i] en el caso Booleano. Note que considerando la expresión a(w) $\wedge_{i=1,\,p}$ [y_i (w) R_i d_i] somos capaces de definir el Objeto Simbólico s = (a, R, d) (Diday & Billard, 2002).

La *extensión* de un Objeto Simbólico es una función que nos permiten reconocer cuando un individuo alcanza la descripción de la clase o una clase cabe en otra más genérica. En el caso Booleano, la extensión de un Objeto Simbólico es denotada Ext (s) y definida por la extensión de a, que es: Ext (s) = {w ∈ Ω / a (w) = verdadero}.

En el caso Modal, se considera un umbral α, con el cual la extensión es definida por Ext_α (s) = Ext (a) = {w ∈ Ω / a (w) ≥ α} (Diday & Billard, 2002).

Diday (2002) presenta las características más significativas para los algoritmos:

1. Comienzan con una tabla de datos simbólica como entrada y dan como salida un conjunto de Objetos Simbólicos. Estos Objetos Simbólicos dan la explicación de los resultados en un lenguaje cercano al del usuario.
2. Se recurre a procesos eficientes de generalización en los algoritmos, con el fin de seleccionar las variables y los individuos más representativos.

3. Proveen descripciones gráficas teniendo en cuenta la variación interna de los objetos simbólicos.

Los algoritmos desarrollan las tareas de análisis sobre los objetos simbólicos están organizados en el libro de Bock y Diday (2000) de la siguiente manera: Estadísticas Descriptivas (Capítulo 6), Similaridad y Disimilaridad (Capítulo 8), Análisis Factorial Simbólico (Análisis de Componentes Principales, Análisis Factorial; Capítulo 9), Discriminación (Análisis Simbólico Discriminante de núcleo, Árboles de Decisión, Árboles de Segmentación; Capítulo 10), Agrupamiento (Agrupamiento Divisivo, Agrupamiento Piramidal; Capítulo 11), Visualización y Edición (Zoom Star y Editor; Capítulo 7).

Estos algoritmos fueron implementados en Software Sodas 1.200. Los proyectos SODAS y ASSO, ambos terminados y desarrollados por un numeroso equipo de investigadores europeos, Universidades europeas, Instituciones Estadísticas Oficiales y empresas. El primer proyecto implementa el software capaz de extraer SOs de bases de datos relacionales, analizarlos y visualizarlos. El segundo, Sodas 2.5, mejora la funcionalidad y añade nuevos métodos para recuperación, análisis y visualización, tales como: Agrupamiento Dinámico, Agrupamiento por Distancias, mapas de Kohonen, Árboles no supervisados de Agrupamiento, Árboles Bayesianos, Regresión y Red Neuronal Multi-Capa Perceptrón (Diday, 2004; Asso Home Page; Diday E & Billard L, 2002; El Golli & Lechevallier, 2004; El Golli et al., 2004; Rossi et al., 2002; Noirhomme, 2004).

El objetivo principal de ambos softwares es analizar datos oficiales de Instituciones Estadísticas Oficiales (para más detalle, Sitios ASSO o SODAS).

Construyendo Objetos Simbólicos

Fundamentalmente, la mayor parte de la información almacenada en un Data Warehouse es agregada. Por ejemplo, supongamos que nuestra unidad de análisis esta constituida por nuestros clientes agrupados de acuerdo a la actividad primaria que ellos realizan. Los atributos que necesitamos para caracterizarlos son: los valores mínimos y máximos de los créditos que tomaron en el año pasado, - los continentes a los cuales los clientes pertenecen y su performace en el pago. Este nos define un cierto nivel de granularidad, que estamos interesados en almacenar y analizar de acuerdo a las peticiones de información establecida en la compañía. ¿Cómo modelamos a este tipo de situaciones con OSs?

El descriptor del objeto simbólico debe tener los siguientes atributos:

a. Performance de pago: representa el comportamiento del cliente con respecto al pago de los préstamos.
b. País: esto significa el país de la empresa.
c. Monto en Euros especifica el rango de valores del importe de los prestamos tomados por los clientes el año pasado.

Supongamos que en nuestras bases de datos operacionales tenemos almacenadas las siguientes relaciones:

Tabla 1 Clientes

#Cliente	…	Transacción Inicial	Performance de pago	País	Actividad Principal
041	…	23-May-03	bueno	España	Manufacturas
033	…	25-Jul-03	regular	China	Manufacturas
168	…	30-Jul-03	bueno	Australia	Agricultura
457	…	2-Ene-04	malo	Sudan	Servicios

#Cliente	…	Transacción Inicial	Performance de pago	País	Actividad Principal
542	…	12-Feb-04	regular	Argentina	Agricultura
698	…	13-Abril-04	bueno	India	Servicios
721	…	22-Ago-04	regular	Francia	Servicios
844	…	15-Sep-04	NC	Canadá	Servicios
987	…	25-Oct-04	NC	Italia	Agricultura
1002	…	10-Nov-04	bueno	Alemania	Manufacturas
1299	…	28-Dic-04	regular	México	Agricultura

Tabla 2 Créditos

#Crédito	Monto en Euros	Plazo en meses	Fecha	#Cliente
1234	35000	12	12-Ene-04	041
1343	44000	6	2-Feb-04	033
1498	62000	18	15-Mar-04	844
1455	50000	24	29-Abr-04	987
1567	75000	12	20-May-04	457
1580	37800	6	11-Jun-04	542
1625	24500	12	12-Jul-04	721
1654	65000	18	20-Ago-04	698
1679	80230	36	21-Sep-04	168
1740	29000	6	7-Oct-04	1002
…	…	…		…
2920	45000	12	24-Dic-04	1299

Tabla 3 Taxonomía

País	Continente
España	Europa
China	Asia
Australia	Oceanía
Sudan	África
Argentina	América
India	Asia
Francia	Europa
Canadá	América
Italy	Europa
Alemania	Europa
México	América

Como podemos observar, tenemos tantos objetos simbólicos como valores tiene la variable actividad principal, variable por la cual generalizamos para armar la unidad de análisis. Los descriptores de los objetos simbólicos fueron expresados siguiendo la misma notación utilizada en el libro de Bock y Diday, ellos son:

OS-Agricultura (4) = [Performance-Pago={ "bueno"(0.50), "regular"(0.25), "NC"(0.25)}] ∧ [Continente = {"América"(0.5), "Europa"(0.25), "Oceanía"(0.25)}] ∧ [Monto-en-Euros = [37800:80230]].

OS-Manufacturas (3) = [Performance-Pago = {"bueno"(0.66), "regular"(0.33)}] ∧ [Continente = { "Asia"(0.33), "Europa"(0.66)}] ∧ [Monto-en-Euros = [45000:65000]].

OS-Servicios (4) = [Performance-Pago = {"bueno"(0.25), "regular"(0.25), "malo"(0.25), "NC"(0.25)}] ∧ [Continente = {"África"(0.25), "Asia"(0.25), "Europa"(0.5)}] ∧ [Monto-en-Euros = [24500:75000]].

Ahora tenemos unidades de segundo orden representando el concepto de actividad principal en nuestra compañía. Los números entre paréntesis representan la cantidad de individuos en esa clase, las variables están notando los valores para la clase. Por ejemplo el objeto simbólico Servicios está constituido por 4 individuos, cuya Performance-Pago tiene una distribución uniforme en los valores del dominio de definición; los clientes se hallan distribuidos un 25 % en África, 25 % en Asia y un 50% en Europa; los montos en Euros de sus préstamos oscilan entre 24500 y 75000.

Podríamos realizar, con las mismas tablas relacionales otra agregación, por ejemplo para conocer el comportamiento de los clientes agrupados ahora por continente, con lo cual los objetos simbólicos que obtendríamos serian los siguientes:

OS-África (1) = [Performance-Pago = { "mala" (1)}] ∧ [País = {Sudán}] ∧ [Actividad ={"Servicios" (1)}] ∧ [Monto-en-Euros = [75000:75000]].

OS-América (3) = [Performance-Pago = {"regular" (0.66), "NC" (0.33)}] ∧ [País = {Argentina, Canadá, México}] ∧ [Actividad = {"Agricultura" (0.66), "Servicios" (0.33)}] ∧ [Monto-en-Euros = [37800:62000]].

OS-Asia (2) = [Performance-Pago = { "buena" (0.50), "regular" (0,50)} ∧ [País = {China, India}] ∧ [Actividad = {"Manufacturas" (0.50), "Servicios" (0.50)}] ∧ [Monto-en-Euros = [44000:65000]].

OS-Europa (4)= [Performance-Pago ={"buena" (0.50), "regular" (0.25), "NC" (0.25)}] ∧ [País ={Alemania, España, Francia, Italia}] ∧ [Actividad = {"Agricultura" (0.25), "Manufacturas" (0.50), "Servicios" (0.25)}] ∧ [Monto-en-Euros = [24500: 50000]].

OS-Oceanía (1)= [Performance-Pago ={ "buena" (1)}] ∧ [País = {Australia}] ∧ [Actividad ={"Agricultura" (1)}] ∧ [Monto-en-Euros = [80230:80230]].

Luego de ejercitarnos con estos casos sencillos, podemos ver que planear una análisis con objetos simbólicos no es simple. Es necesario por ejemplo: - conocimiento del dominio, - reglas del negocio, - tipo de información almacenada en los sistemas operacionales, - estructuras organizativas. Algunos de estos elementos que mencionamos antes, constituyen el Conocimiento Contextual.

Proceso Integrado

La integración entre bases datos relacionales y algoritmos de mining fue direccionada por Netz A, Chaudhuri S., Bernhardt J. & Fayyad U. (2000). En los últimos años las principales empresas proveedoras de bases de datos, tales como IBM, Microsoft y Oracle han agregado algunas funcionalidades de Data Mining a sus Sistemas Gestores de Bases de Datos (DBMS). IBM incorporó como componente principal a su Information Warehouse Framework funciones de Data Mining (IBM, 2005). SQL Server 2005 de Microsoft mejora las capacidades de análisis de la versión 2000 (Mac Lennan J., 2004). Oracle, por su parte, ha integrado Olap y Data Mining directamente al servidor de bases de datos (Oracle Corporation, 2002).

El proceso combinado, Data Warehousing - Data Mining con objetos simbólicos (Figura 1) facilita las tareas de gestión de la información a ser analizada y provee un marco integrador que

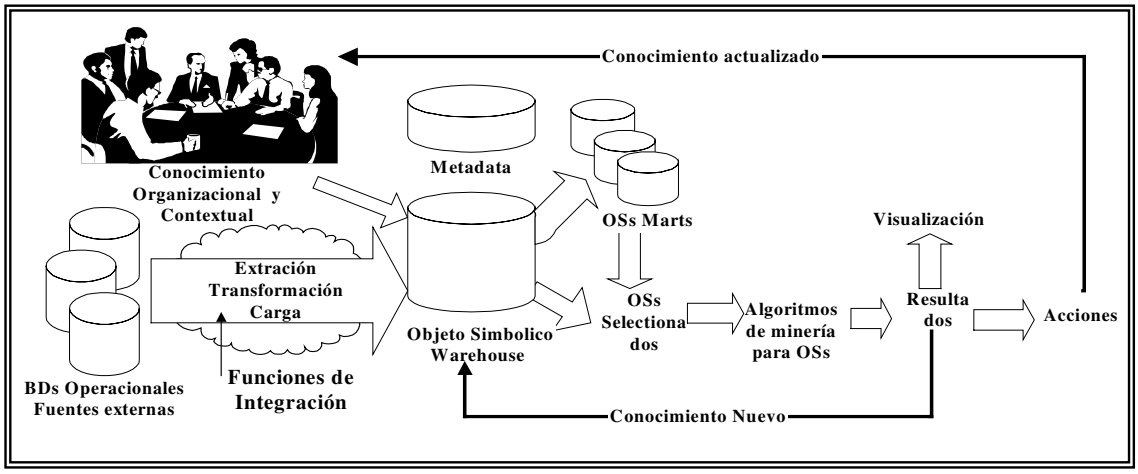

Figura 1 Combinación de DW & DM con objetos simbólicos.

posibilita el análisis exploratorio y permite compartir los conocimientos nuevos. De esta manera los usuarios tienen disponible la información para analizar.

Arquitectura

La arquitectura conceptual es ilustrada en la figura 2, en la cual podemos identificar los módulos más importantes del sistema. Un gestor es asociado a cada uno de los componentes, con el objetivo de proveer flexibilidad (es simple agregar nuevos servicios) y encapsulación de la funcionalidad (ayuda en la organización del diseño y modularización).

En consecuencia, con una simple observación al esquema podemos determinar: cuales son los componentes, las funcionalidades del sistema en general y los flujos de información / conocimiento.

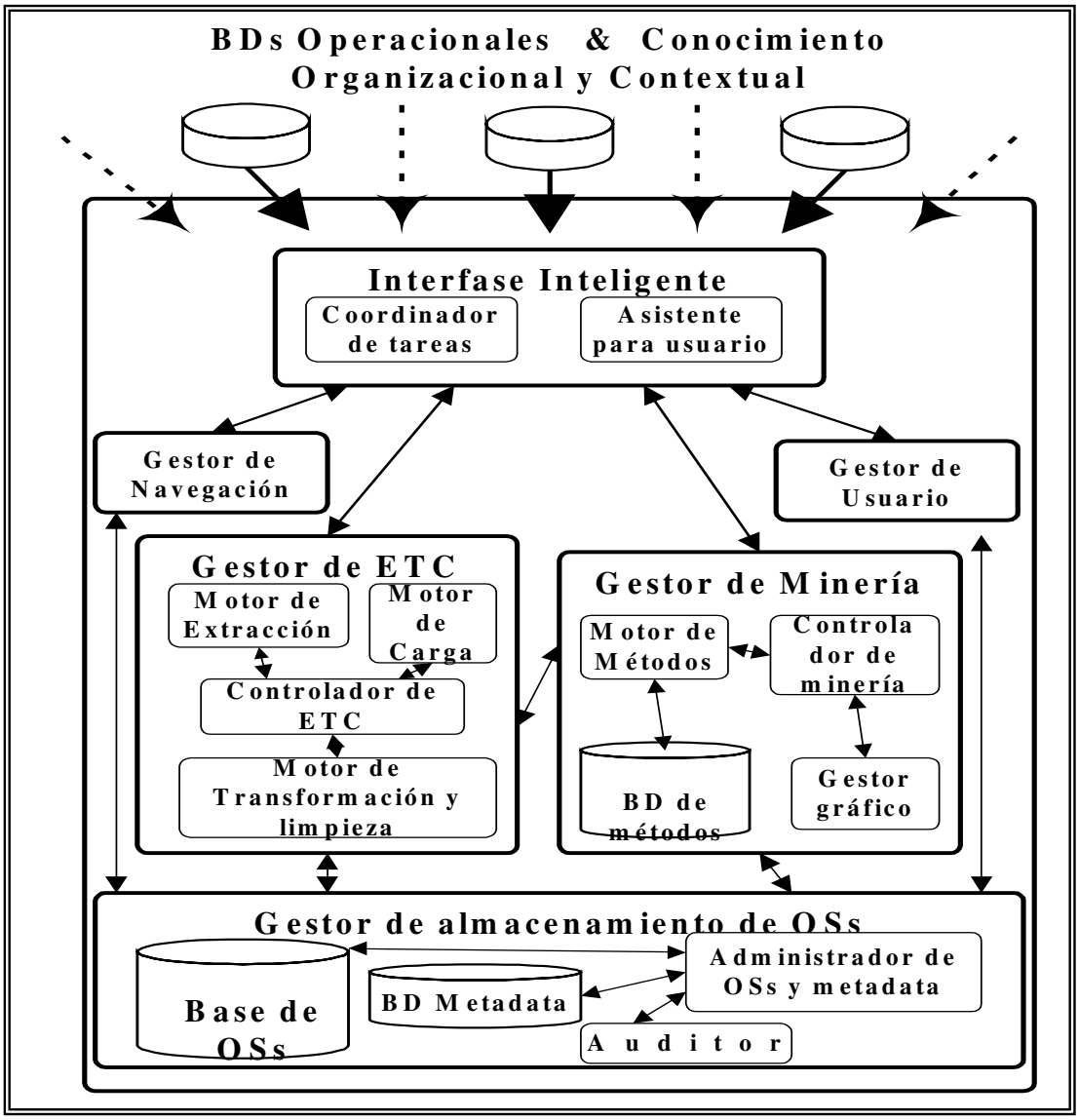

Figura 2. Arquitectura

Un aspecto muy importante a tener en cuenta para el Data Warehouse es: *la meta data*. Según Staudt constituye un factor clave para el éxito en el desarrollo de proyectos de Data Warehousing (Staudt et al. , 1999).

Los meta datos para objetos simbólicos, como afirma Vardaki, deben describir las variables simbólicas, su naturaleza, componentes y dominio. Todos los meta datos necesarios para creación de datos simbólicos y su procesamiento pueden ser presentados como una plantilla de metadata o modelados en un esquema separado. El progreso del modelado indicará no sólo los items meta datos considerados sino también, en un formato estructurado, especificarán su relación y los operadores / transformaciones que pueden ser aplicados para manipulaciones adicionales (Vardaki M., 2004).

En los párrafos siguientes desarrollaremos una explicación breve de la funcionalidad planificada para cada componente.

Interfase Inteligente: Esta componente es la responsable del vinculo entre el usuario y el sistema. Además debe realizar las tareas de coordinación entre los módulos, es como un mediador y una fachada en términos arquitectónicos.

Está compuesta por dos subcomponentes fundamentales:

- **Coordinador de tareas** debe garantizar el uso adecuado del sistema por medio de la coordinación de tareas.
- **Asistente para usuario:** facilita la utilización del sistema ayudando al usuario en sus tareas.

Gestor de ETC: Este módulo debe brindar las interfases necesarias para los procesos de extracción, transformación y carga de la información, asegurando el traspaso seguro y confiable de los datos en objetos simbólicos. Dos cargas de tipos diferentes son asumidas: *inicial y ad hoc*. La primera modela los conceptos principales del negocio, establecidos en el análisis de requerimientos informacionales y la segunda responde a nuevos conceptos o necesidades de información en la organización. Un nuevo concepto práctico, que asumimos como la guía para mantener la carga de los objetos simbólicos inicial es ETC activo (Adzic y Fiore, 2003).

Los subcomponentes esenciales del Gestor de ETC son:

- ***Controlador ETC:*** Coordina todo el proceso de extracción, transformación y carga. También, el módulo debe verificar la edad de los objetos simbólicos y cuando sea necesario actualizarlos. Es decir, cuando el ciclo de vida de un objeto simbólico ha finalizado, éste debe ser reemplazado en forma automática; el objeto anterior debe permanecer almacenado en la base para mantener la historia y los meta datos actualizados.
- ***Motor de Extracción:*** Comenzando por las consultas de los usuarios o por las predefinidas, esta componente debe almacenar los datos tomados de las bases de datos operacionales en una memoria transitoria.
- ***Motor de Transformación y Limpieza:*** transforma datos estándar almacenados en la memoria temporal por el motor de extracción, en objetos simbólicos. Cuando el usuario trabaja con consultas ad-hoc, un editor permite la interacción para que el usuario pueda decidir acerca de los valores nulos, valores perdidos, renombre de variables, etc. Para las consultas predefinidas, la transformación y limpieza es realizada en forma automática por medio de reglas establecidas.

- ***Motor de Carga*** Ahora los conjuntos de objetos simbólicos están listos en la memoria temporal, este módulo los almacena en la base de datos; lo mismo realiza con los Metadatos.

Gestor para Usuarios: el sistema debe registrar, controlar, adjudicar o cambiar roles a los usuarios.

Gestor de Navegación: Una funcionalidad significativa del Data Warehouse es la capacidad para mostrar y navegar en los datos. La extensión del Álgebra relacional para objetos simbólicos sugerida por Wan & Zeitouni (2004) puede ser aplicada a nuestro módulo.

Este componente de software debe hacerlo como un catálogo, del modo iterativo. También esto permite hacen el reporte de objetos simbólicos y su metadata.

Gestor de Minería: Este componente básico es el corazón de la funcionalidad analítica del sistema. Está compuesto por los siguientes subcomponentes principales:

- ***Controlador de Minería*** es el responsable de la coordinación entre los otros subcomponentes del gestor de Minería.
- ***Motor de Métodos***: Ejecuta los algoritmos de análisis sobre objetos simbólicos, está pensado como un interprete de métodos.
- ***Base de Métodos***: Base conteniendo el código de cada uno de los algoritmos de análisis para objeto simbólico, mencionados en la sección correspondiente a la presentación de objetos simbólicos.
- ***Gestor Gráfico***: Realiza todo tipo de gráficos sobre objetos simbólicos; particularmente debe implementar el gráfico Zoom Star (Noirhomme, 2000, 2002, 2004), el cual constituye la mejor forma de representación para este tipo de datos.

Gestor de Almacenamiento de OSs: Aquí tenemos el Objeto Simbólico Warehouse, propiamente dicho; debe contener los objetos simbólicos, sus meta datos, realizar el control de concurrencia, auditoria y preservar la seguridad de los objetos y de los meta datos.

El gestor de almacenamiento de objetos simbólicos tiene cuatro subcomponentes claves:

- ***Administrador de OS y Metadata*** administra el acceso o control de concurrencia a los objetos simbólicos y metadata. Mientras no haya actividad en la organización este módulo debe ser capaz de efectuar backup del OS Warehouse, de manera tal que los errores y potenciales demoras sean minimizados.
- ***Base de Datos Simbólica:*** esta es la memoria no volátil para la información, donde físicamente almacenaremos los objetos simbólicos.
- ***Base de Datos para Metadata***: aquí almacenaremos la información necesaria para administrar y controlar a los objetos simbólicos.
- ***Auditor***: debe realizar las funciones de auditoria en el Data Warehouse, tales como: Control de accesos, historial de cambios (para los objetos y meta datos). Con la información proporcionada por este modulo se podrá monitorear la utilidad de los objetos, de manera tal que cuando se detecten objetos que no se utilizan en los análisis, estos sean evaluados en su importancia.

Una de las cuestiones más importantes en Data Warehousing y Data Mining es la preservación de la privacidad. Oliveira y Zaïane (2004) enfatizan el conflicto de equilibrar la preservación de privacidad y el descubrimiento de conocimiento. Como podemos observar, los objetos simbólicos conservan las exigencias de privacidad, ya que ellos son unidades de alto nivel,

agrupando individuos o clases de individuos. El usuario no puede inferir individuos específicos por medio de consultas.

Futuras Investigaciones y Desarrollo

El paso siguiente en este trabajo, será la especificación formal de esta arquitectura en términos de diseño, para luego analizar las alternativas más eficientes en programación. Algunos de los problemas más importantes a resolver están constituidos por: -El lenguaje interno para manipular a los objetos simbólicos, dado que eficiencia espacial y temporal son necesarias; - Almacenamiento de los datos.

Considerando la modularidad funcional, una implementación orientada a objeto sería la más conveniente; otra implementación que sería muy atractiva es a través de un sistema multi-agentes, también orientado a objetos.

Los algoritmos nuevos que trabajarán sobre objetos simbólicos, serán dirigidos por técnicas que aún deben ser exploradas y desarrolladas. Las más importantes y útiles en Data Mining son: Reglas de Asociación, Regresiones, Interpretación de Clases y otros tipos de Redes Neuronales.

Otra área, potencialmente útil y muy atractiva para investigación y desarrollo es el desarrollo de sistemas expertos, especializados en ayuda a la toma de decisiones mediante el formalismo de objetos simbólicos.

Conclusiones

Como fue explicado en la sección introductoria, los objetos simbólicos permiten representar entidades física o conceptos del mundo real en forma dual, respetando su variación interna y estructura. La integración de los procesos de Data Warehousing y Data Mining a través de objetos simbólicos proyecta una visión integral de la dirección al nivel de toma de decisiones, permitiendo la descripción intencional / extensional de los conceptos más importantes en un lenguaje cercano al utilizado por los usuarios.

El control de calidad, la seguridad y la veracidad de la información son obtenidos en los objetos simbólicos en el proceso de creación, ya que en este proceso son establecidos los significados de valores nulos y son incluidos los meta datos (los últimos son importantes sobre todo en Data Warehousing y en Procesos de Toma de Decisiones).

Una de la ventaja más valorada proporcionada por el uso de objetos simbólicos es la facultad para realizar varios niveles de análisis. Lo que conlleva a que la salida de un método sea la entrada de otro algoritmo. Principalmente, esto puede observarse en clustering o métodos de clasificación, debido a que en la mayoría de los casos la salida del algoritmo es un conjunto de objetos simbólicos.

Las desventajas principales incorporadas por el uso de objetos simbólicos son: - la complejidad para determinar cuales serán los mejores objetos que representarán las tareas de análisis en la organización y - cuando los objetos deben ser cambiados o actualizados.

Nuestra arquitectura permite un ambiente integrado del trabajo, con posibilidades de mejora y crecimiento; esto se debe a la flexibilidad y a la modularidad de su diseño. Desde el punto de vista del análisis de datos es muy importante la integración Data Warehouse y Data Mining porque la posibilidad de añadir el conocimiento descubierto en el Data Warehouse es muy práctica, ahorramos pasos, ya que los objetos simbólicos representan conceptos, los cuales no tienen que ser integrados, ni enriquecidos. Al mismo tiempo facilitan el análisis exploratorio. Por ejemplo si descubrimos nuevas características de clientes potenciales o relaciones entonces los descriptores de

los objetos simbólicos almacenados en el Data Warehouse pueden ser actualizados, creando nuevos objetos simbólicos.

Por lo tanto, al estar trabajando con entidades de mayor nivel de abstracción, se mejora notablemente la Gestión del Conocimiento en la organización y se logra optimizar los procesos de toma de decisiones. Debido a que los objetos simbólicos suceden como consecuencia de la transformación del conocimiento implícito en las mentes de los analistas en conocimiento explícito.

Referencias

Adzic, J. and Fiore, V. (2003). Data Warehouse Population Platform. In Proceedings of 5th International Workshop on the Design and Management of Data Warehouses (DMDW'03), Berlin, Germany

Appice A., D'Amato C., Esposito F. & Malerba D. (2004). k-Nearest Neighbors Classification of Symbolic Objects. In Proceeding of Workshop on Symbolic and Spatial Data Analysis: Mining Complex Data Structures, pg 19-30. ECML/PKDD. Pisa, Italy.

ASSO, Project Home Page. http://www.info.fundp.ac.be/asso/.

Bock H.H. & Diday E. (2000). Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data. Studies in Classification, Data Analysis and Knowledge Organization. Springer Verlag.

Cali A, Calvanese D., De Giacomo G. & Lenzerini M. (2004). Data integration under integrity constraints. In Information Systems 29 pg.147 –163. Elsevier Science B.V.

Calvanese D., De Giacomo G. Lenzerini M., Nardi D. & Rossati R. (2004). Data integration in Data Warehousing. In International Journal of Cooperative Information Systems 10(3), pg. 237-271. World Scientific Publishing Company.

Chavent Marie (1997). "Analyse des Données symboliques. Une méthode divisive de classification". Thèse de doctorat in Sciences. I'Université Paris IX-Dauphine. Downloaded on August 2002. http://www.math.u-bordeaux.fr/~chavent/thesemc.ps

Clear John et al. (1999). NonStop SQL/MX Primitives for Knowledge Discovery. In Proceedings *KDD pg* 425-429.

Davenport, T. and Prusak, L. (1998). Working knowledge. Harvard Business School Press.

Diday E & Billard L. (2002). Symbolic Data Analysis: Definitions and examples. http://www.stat.uga.edu/faculty/LYNNE/tr_symbolic.pdf

Diday E. (2003). Concepts and Galois Lattices in Symbolic Data Analysis. JIM 2003 Journées de l'Informatique Messine. JIM'2003. Knowledge Discovery and Discrete Mathematics Metz, France. September 3-6, 2003.http://www.inist.fr/uri/jim03/diday.pdf

Diday Edwin (2004). From Data Mining to Knowledge Mining: Symbolic Data Analysis and the Sodas Software. Workshop on Applications of Symbolic Data Analysis. Lisboa Portugal. January 2004. http://www.info.fundp.ac.be/asso/dissem/W-ASSO-Lisbon-Intro.pdf

El Golli A., Connan-Guez B. & Rossi F. (2004). Self-organizing maps and symbolic data. JSDA Electronic Journal of Symbolic Data Analysis- 2(1). ISSN 1723-5081. Downloaded on November 2004. http://www.jsda.unina2.it/newjsda/volumes/Vol2/No1/V2n1Aicha.pdf.

El Golli A & Lechevallier Y (2004). Extraction de classes homogènes et création d'objets symboliques. 4èmes journées d'Extraction et de Gestion des Connaissances, Clermont Ferrand, Université Blaise Pascal. Fouille de données complexes dans un processus d'extraction des connaissances. FDC- EGC04.

Gowda K. (2004). Symbolic Objects and Symbolic Classification. Invited paper in Proceeding of Workshop on Symbolic and Spatial Data Analysis: Mining Complex Data Structures. ECML/PKDD. Pisa, Italy.

Han J. & Kamber M. (2001). Data Mining: Concepts and Techniques, Morgan Kaufmann.

IBM (2005): IBM Software - DB2 Intelligent Miner - Family Overview. http://www-306.ibm.com/software/data/iminer/.

Inmon, W. (1996). Building the Data Warehouse. 2nd edition. John Wiley & Sons.

Mac Lennan J. (2004) SQL SERVER 2005 Unearth the New Data Mining Features of Analysis Services 2005. http://msdn.microsoft.com/msdnmag/issues/04/09/AnalysisServices2005/default.aspx

Netz A, Chaudhuri S., Bernhardt J. & Fayyad U. (2000). Integration of Data Mining and Relational Databases. Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt.

Nigro H. O. & González Císaro, S. (2005), Symbolic Object and Symbolic Data Analysis. In Encyclopedia of Database Technologies and Applications. Laura C. Rivero, Jorge H. Doorn, Viviana E. Ferraggine, Editors. Idea Group Inc Publishers. June 2005.

Noirhomme Monique (2004). Visualization of Symbolic Data. Workshop on Applications of Symbolic Data Analysis. Lisboa Portugal. January 2004 http://www.info.fundp.ac.be/asso/dissem/W-ASSO-Lisbon-Visu.pdf

Oliveira S. & Zaïane O.(2004). Toward Standardization in Privacy-Preserving Data Mining. In Proceedings of the Second Annual Workshop on Data Mining Standards, Servicios and Platforms. KDD 2004. Seattle WA.

Oracle Corporation (2002): Oracle9i Data Mining, www.oracle.com/technology/products/oracle9i/pdf/o9idm_bwp.pdf.

Proceeding of Workshop on Symbolic and Spatial Data Analysis: Mining Complex Data Structures. ECML/PKDD. Pisa, Italy. http://ecmlpkdd.isti.cnr.it/workshops/W2.pdf.

Rossi F. & Conan-Guez B (2002). Multi-layer Perceptron on Interval Data. In Proceedings of IFCS'2002.

Staudt M., Vaduva A. & Vetterli T. (1999). The Role of Metadata for Data Warehousing. http://ftp.ifi.unizh.ch/pub/techreports/TR-99/ifi-99.06.ps.gz

Teste Olivier (2004). Towards Conceptual Multidimensional Design in Decision Support Systems. Fifth East-European Conference on Advances in Databases and Information Systems. Vilnius, Lithuania.

Theodoratos D. & Sellis T. (1999). Designing Data Warehouse. Data and Knowledge Engineering (DKE), 31, 3, pg. 279 - 304. Oct. 1999.

Touati M. & Diday E. Sodas Home Page. http://www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm.

Vardaki Maria (2004). Metadata for Symbolic Objects. JSDA Electronic Journal of Symbolic Data Analysis- 2(1). ISSN 1723-5081. http://www.jsda.unina2.it/newjsda/volumes/Vol2/No1/V2n1Vardaki.pdf

Wan T. & Zeitouni K (2004). Extension de l'algèbre relationnelle aux données symboliques. 4èmes journées d'Extraction et de Gestion des Connaissances, Clermont Ferrand, Université Blaise Pascal. Fouille de données complexes dans un processus d'extraction des connaissances. FDC-EGC04.