Integración del Data Warehouse y Herramientas de Análisis de Datos por medio de Objetos Simbólicos

Héctor Oscar Nigro, Sandra González Císaro
INCA/INTIA - Departamento de Computación y Sistemas
Facultad de Ciencias Exactas - UNICEN – Tandil
Campus Universitario - Paraje Arroyo Seco s/n
TE: +54-2293-432466 – FAX: +54-2293-444431
e-mail: onigro@exa.unicen.edu.ar, sagonci@exa.unicen.edu.ar

Resumen

En este proyecto se darán las bases para la investigación y el desarrollo de un sistema ejecutivo que integre el Data Warehouse con las herramientas de Análisis de Datos. Basado en el modelo matemático de conceptos introducidos por el Prof. Diday, conocidos como Objetos Simbólicos. [1, 2, 3]. Con este desarrollo se evitarán problemas en el formato de los datos y se optimizarán los beneficios en el proceso de análisis. Ya que estaremos trabajando con entidades más abstractas que representarán conocimiento dentro de la organización.

Introducción

Las actuales bases de datos contienen cantidades gigantescas de información, no siempre presentada al nivel de agregación necesario para la toma de decisiones, ya sea en las organizaciones públicas o privadas. ”El almacenamiento de datos (Data Warehousing) y el procesamiento analítico en línea (on - line analytical processing) son elementos esenciales en el soporte de decisiones, que se están convirtiendo de forma creciente en un foco de la industria de las bases de datos”[10].

Los sistemas OLAP resuelven el problema de presentar diferentes niveles agregación y visualización para datos multidimensionales a través del paradigma del cubo; las técnicas clásicas de análisis de datos (análisis de factorial, regresión, dispersión) son aplicadas a individuos(tuplas o registros para nosotros en una base de datos). Los objetos clásicos de análisis no resultan lo suficientemente expresivos como para representar tuplas con celdas conteniendo una distribución, reglas lógicas, atributos multivaluados, intervalos, etc.

Surge así, la necesidad de tener un nuevo tipo de dato que sustente estas características y mantenga el dualismo ente individuo y clase.[11] Por ejemplo: con este simple Objeto Simbólico s = (cant_reclamos [2:10] $\wedge$ consumo_en_$ (0,3 Bueno, 0,7 Regular)) podemos describir dos hechos: 1) un cliente que posea entre 2 y 10 reclamos en un período determinado, con consumos: el 30 % de las veces regular y el 70 % bueno. 2) Un conjunto de clientes que posean entre 2 y 10 reclamos y con consumos: el 30 % de las veces regular y el 70 % bueno.

Este nuevo tipo de dato es la noción de Objeto Simbólico introducida por el Prof. Diday, quién lo define como un modelo matemático, a través del cual se puede modelar una entidad física o un concepto del mundo real. Las entidades físicas están representadas por las tuplas almacenadas en la base de datos. Los conceptos son representados por entidades de nivel superior obtenidos del análisis de uno o varios expertos, por clasificación automática o de alguna agregación en particular surgida de la unidad de estudio.[1,2,3,11].

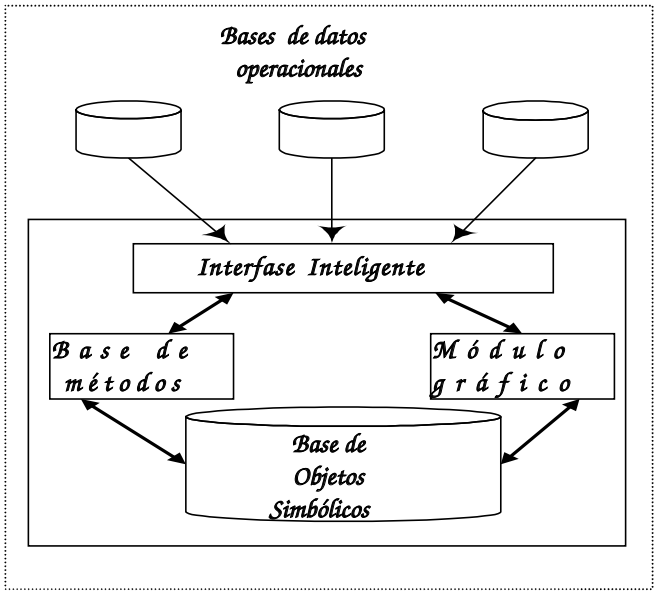
Más formalmente, un Objeto Simbólico s es una tripla s=(a, R, d) donde R es una relación entre descripciones, d es una descripción y “a” es una función definida desde Ω(universo de estudio) en L dependiendo de R y d.[1, 3, 4].

Utilizaremos los Objetos Simbólicos para representar los conceptos que deseamos almacenar en el Data Warehouse, y realizaremos las tareas de análisis sobre ellos. Como punto de partida tenemos las técnicas desarrolladas para Objeto Simbólico[1,2,3], pero también resulta interesante ampliarlas con otras provenientes de Data Mining, como por ejemplo: redes neuronales, reglas de asociación.

Primero analizaremos la arquitectura propuesta para la aplicación, luego expondremos las área involucradas en la investigación y para finalizar las conclusiones.

2)Arquitectura

En el siguiente gráfico vemos los principales componentes que constituyen la aplicación, en los apartados posteriores describiremos su funcionalidad y veremos que es necesario investigar para su desarrollo.



Dentro de esta arquitectura necesitamos el desarrollo de un lenguaje que nos permita manejar los Objetos Simbólicos. Si bien ya existe un lenguaje (SOL)[1], consideramos que es necesario el desarrollo de uno que no tenga las limitaciones para la inclusión de reglas de dependencia y trabaje con una semántica más lógica. Sin llegar a ser un lenguaje puramente lógico o deductivo.

2.1) Interfase Inteligente

Esta componente será la encargada de brindar la interfase con las bases de datos operacionales y la funcionalidad del sistema con el objetivo de ser amigable a los usuarios.

También será importante que tenga un comportamiento de aprendizaje en cuanto a cada uno de los perfiles de los usuarios gerenciales, con el propósito de ayudarlos en las tareas o brindarle sugerencias. Por lo que sería interesante la inclusión de agentes.

2.2) Base de métodos

Este es el centro de la funcionalidad en cuanto al análisis. El módulo es una biblioteca con todas las técnicas que se aplican al Objeto Simbólico [1,2,3,4]. Por lo tanto, debe conocer los parámetros de cada una de las mismas, como así también la configuración en particular que deben poseer los Objetos Simbólicos para cada una de ellas.

También debe construirse un parser para que realice un chequeo sintáctico del lenguaje en el que escribamos los Objetos Simbólicos para que verifique su correctitud sintáctica.

La creación de los Objetos también será incluida en este módulo, a través de un editor. Para la extracción de los Objetos a partir de las bases operacionales podríamos recurrir a consultas SQL sobre las mismas. Con dos posibilidades:1) un conjunto básico de Objetos Simbólicos modelando los requerimientos de información que la empresa tenga(basado en la heurística, por el conocimiento de los procesos de análisis que se desarrollen en la empresa), junto con un mecanismo de actualización automática. 2) Objetos creados sobre demanda para determinados tipos de análisis.

Con la primera opción creamos en forma estática un conjunto de O.S., con la segunda tenemos la posibilidad de flexibilizar, para permitir la evolución del Data Warehouse en términos de necesidad de nueva información.

Un tema que requeriría un esfuerzo en la investigación sería: como adaptar técnicas de Data Mining, tales como reglas de asociación y redes neuronales a los Objetos Simbólicos.

2.3) Módulo Gráfico

Una de las tareas de mayor importancia en los procesos de Análisis de Datos es la visualización. Sobre todo , porque estamos trabajando con datos hipervariados, que pueden tener dependencias lógicas y jerárquicas. La mejor forma de representar Objetos Simbólicos es a través Zoom Star, desarrollada por Monique Noirhomme. Se trata de un gráfico radial en el que se muestran los valores de cada variable incluida en el descriptor del O.S.[12]

Le concierne también a éste módulo, la inclusión de otros tipos de gráficos como barras, histogramas, box-plot, árboles, etc.

También resultaría práctico el desarrollo de algún editor para reportes gráficos, en los que se podría seleccionar que partes del descriptor del O. S. se desea visualizar.

2.4) Base de Objetos Simbólicos

Este modulo es el Data Warehouse propiamente dicho, el repositorio donde tendremos la información necesaria para el análisis y la toma de decisiones. Pero ahora no trabajamos con los modelos clásicos en el tema como: son el diagrama estrella o copo de nieve[9]. Sino que tenemos una red de Objetos Simbólicos, razón por la cual se hace necesario un análisis diferente para el almacenamiento de los mismos, teniendo en cuenta las complejidades espaciales y temporales.

Los aspectos más relevantes a tener en cuenta para este módulo serán: la administración de los metadatos, ya sea en términos de los datos sobre sí mismos o sobre los usuarios que accedan a los mismos[13], nivel de granularidad, control de concurrencia, seguridad e integridad de la información.

3)Temas involucrados en el trabajo

Las áreas básicas incluidas en el proyecto son: 1)Data Warehouse, 2)Bases de Datos, 3)Estadística, 4)Análisis de Datos, 5)Análisis de Datos Simbólicos, 6)Data Mining, 7)Inteligencia Artificial, 8)Sistemas Inteligentes, 9)Aprendizaje Automático, 10)Sistemas expertos gerenciales, 11)Agentes Inteligentes, 12) Visualización de datos.

4)Conclusiones

Este trabajo pretende, a partir de la integración del Data Warehouse con las herramientas de análisis de la información, una visión integral de la gestión a nivel gerencial.

Optimización en los procesos de toma de decisiones, con la mejora en el conocimiento que la organización tiene de sí misma. Control en la calidad , la seguridad y la veracidad de la información.

Referencias

1- Bock H.H. and Diday E. "*Analysis of Symbolic Data*". Studies in Classification, Data Analysis and Knowledge Organization. Springer Verlag. (2000).

2- Diday E, Moscoloni N. *Análisis de Datos Simbólicos*. Conferencias dadas por el Prof. Diday en el IRICE. Conicet. UNR. 1993

3- Diday Edwin, *An introduction to symbolic Data Analysis and the Sodas software*. University Paris 9 Dauphine, Ceremade. May 2002. Paper enviado por el autor.

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

5- Han J. And Kamber M. *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2000.

6- Hahsler Michael *Knowledge Management Data Warehouses and Data Mining*. Dept. of Information Processing. Vienna Univ. of Economics and BA. 2001.http://wwwai.wuwien.ac.at/~hahsler/research/datawarehouse_webster2001/talk/download/talk_2.pdf.

7- Fayyad U. Piatetsky-Shiapiro G., Smyth P., and Uthurusamy R. *Advances in Knowledge Discovery and Data Mining*. Merlo Park, California: AAAI Press 1996.

8- Teste Olivier *Towards Conceptual Multidimensional Design in Decision Support Systems*. http://www.science.mii.lt/adbis/local1/Olivier_Testen.pdf.

9- Theodoratos D., Sellis T. *Designing Data warehouse*. Data and Knowledge Engineering (DKE), 31, 3, , pp. 279 - 301. Oct. 1999. http://www.dbnet.ece.ntua.gr/~dwq/p42.pdf

10- Diday Edwin *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

11- Dayal U. Chandhuri S., "*An overview of Data Warehousing and OLAP Technology*" ACM Sigmod Record, 26 (1) March, 1997.

12- Noirhomme Monique, *Multimedia Support for Complex Multidimensional Data Mining*. Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000. http://www.cs.ualberta.ca/~zaiane/mdm_kdd2000/papers/mdm00-08.pdf.

13- Staundt M., Vaduva A. and Vetterli T. *The Role of Metadata for Data Warehousing*. http://ftp.ifi.unizh.ch/pub/techreports/TR-99/ifi-99.06.ps.gz