

Proyecto “Técnicas y herramientas para diseño lógico y mantenimiento de data warehouses relacionales”

Informe de Avance de la Investigación

Agosto 2001

1 Principales actividades realizadas en el proyecto

Un data warehouse (DW) puede ser definido como una base de datos que almacena información en un modelo de datos orientado a satisfacer requerimientos de toma de decisiones. Este modelo debe tener en cuenta a su vez que la información proviene de sistemas operacionales ya existentes y debe ser mantenida a medida que estos sistemas cambian.

El proyecto aborda problemas de diseño y mantenimiento de data warehouses usando un modelo de datos basado en el modelo relacional. El objetivo es definir técnicas e incorporarlas en una herramienta de software para resolver principalmente los problemas de definición del esquema del DW y la gestión de su evolución. El enfoque de la solución se basa en el uso de primitivas de transformación de esquemas. Estas operaciones permiten por un lado, representar conocimiento de criterios de diseño para llegar al esquema final a partir de los esquemas fuentes y por otro lado, proveer la traza del diseño realizado utilizada en la gestión de la evolución.

Las principales actividades desarrolladas hasta el 30/6/2001 las podemos clasificar en 2: específicas y de difusión.

Dentro de las actividades específicas se encuentran:

- Revisión bibliográfica de las técnicas y estrategias de diseño de data warehouses.
- Definición de las primitivas de transformación de esquemas para diseño de data warehouses relacionales.
- Estudio de características de las primitivas y clasificación de las mismas. Definición de la traza de primitivas.
- Definición de guías de uso de las primitivas definidas.
- Desarrollo de un primer prototipo de una herramienta de ayuda al diseño de DW basada en las primitivas incluyendo la traza de aplicación de las mismas.

Dentro de las actividades de difusión destacamos:

- Dictado del curso “Sistema de data warehouses” de 60 hs. para el diploma de actualización profesional ofrecido por el Instituto de Computación de la Facultad de Ingeniería. Participaron en esta actividad los docentes Alejandro Gutiérrez, Adriana Marotta, Verónica Peralta y Raúl Ruggia.
- Dictado del curso “Sistemas de Data Warehousing y OLAP” de 14 hs. en la 8va Escuela de Verano de Cs. Informáticas (RIO 2001) organizada por la Universidad Nacional de Río Cuarto. Área de Computación. Córdoba, Argentina. en Río IV, Argentina. Participaron en esta actividad los docentes Adriana Marotta y Verónica Peralta.

- Trabajo de tres días con la profesora Ana Moura del Instituto Militar de Ingeniería (IME), Río de Janeiro, Brasil. Ana Moura fue invitada por el grupo Concepción de Sistemas de Información del Instituto de Computación para establecer lazos de cooperación. En dicha oportunidad pudimos intercambiar en particular el trabajo en torno al diseño de data warehouses.
- Elaboración y participación en un seminario interno dentro del Instituto de Computación de la Facultad de Ingeniería sobre temas relacionados a la representación de los procesos de carga y refresco de data warehouses usando conceptos del área conocida como *workflow*.

2 Avances obtenidos hasta la fecha

- Un primer relevamiento de trabajos existentes sobre técnicas de diseño de data warehouses [1].
- La definición de un conjunto de primitivas de transformación de esquemas que permiten representar estrategias para construir data warehouses relacionales. Se estudiaron características de las primitivas definidas y se realizó una clasificación de las primitivas según diferentes criterios [2].
- El desarrollo de un primer prototipo consistente en un ambiente gráfico de ayuda al diseño de DW relacionales que incluye la implementación de las primitivas y provee mecanismos para aplicar reglas de consistencia que aseguran invariantes sobre los esquemas de DW [3].
- La proposición de una arquitectura de sistema de información que permita incorporar 2 categorías de técnicas de propagación de cambios a un data warehouse con información extraída de la web: (1) propagación de cambios en esquemas fuentes y (2) propagación de agregado o eliminación de fuentes [4].

3 Análisis de las etapas realizadas con respecto a las previstas

Las actividades realizadas se corresponden con las previstas inicialmente. Existieron modificaciones a nivel del cronograma de ejecución que surgieron en el curso del trabajo. La modificación más destacada es la fusión de las actividades de definición de las primitivas de diseño, la identificación de invariantes del esquema de DW y la definición de guías de uso de las primitivas.

Nos resulta importante informar una decisión que debimos tomar cuando trabajamos en la construcción de una especificación formal de las primitivas de diseño. El objetivo principal de esta actividad en el contexto del proyecto era contar con un tiempo para encontrar una forma de expresar con precisión la funcionalidad de las primitivas. Estudiamos dos posibilidades en ese sentido. Por un lado, el uso de un lenguaje de especificación como B y por otro lado el uso de un prototipo. La primera posibilidad la evaluamos demasiado larga para este proyecto previendo retrasos en el resto de las actividades. Por esta razón hemos seguido el camino de precisar nuestras proposiciones mediante el desarrollo de prototipos.

Anexos

Publicaciones

- [1] A. Gutiérrez, A. Marotta, *An Overview of Data Warehouse Design Approaches and Techniques*. Reporte Técnico INCO-01-09. InCo - Pedeciba, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay. Octubre 2000. ISSN 0797-6410.

- [2] A. Marotta, *Designing Relational Data Warehouses through Schema-Transformation Primitives*. Reporte Técnico INCO-01-10. InCo - Pedeciba, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay. Diciembre 2000. ISSN 0797-6410.
- [3] A. Gutiérrez, A. Marotta, *Designing Relational Data Warehouses through Schema-Transformation Primitives - Prototype*. Reporte Técnico INCO-01-11. InCo - Pedeciba, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay. Junio 2001. ISSN 0797-6410.
- [4] A. Marotta, R. Motz, R. Ruggia, *Managing Source Schema Evolution in Web Warehouses*. International Workshop on Information Integration on the Web, WIIW '2001. Brazil. Abril 2001.

Ponencias

- Presentación del trabajo “Diseño y Mantenimiento de Data Warehouses a través de Transformación de Esquemas” por Adriana Marotta en las VI Jornadas de Informática e Investigación Operativa organizada por el Instituto de Computación de la Facultad de Ingeniería y el Pedeciba Informática en agosto de 2000. Se adjunta la presentación.
- Presentación del artículo “Managing Source Schema Evolution in Web Warehouses” por Adriana Marotta en el workshop WIIW’ 2001 en abril de 2001. Se adjunta la presentación.
- Presentación de avance del trabajo de Ignacio Larrañaga “Especificación de Primitivas de Diseño de Data Warehouses Relacionales en SQL” en mayo de 2001 en el seminario interno sobre temas relacionados a la representación de los procesos de carga y refresco de data warehouses usando conceptos del área conocida como *workflow*.
- Presentación de avance del trabajo de Verónica Peralta “Modelización del Pasaje del Esquema Conceptual al Esquema Lógico de Data Warehouses” en junio de 2001 en el seminario interno sobre temas relacionados a la representación de los procesos de carga y refresco de data warehouses usando conceptos del área conocida como *workflow*.

Formas de difusión empleadas

Además de las anteriores, se utilizaron los siguientes medios para difundir el trabajo en curso.

- Página web accesible mediante la dirección <http://www.fing.edu.uy/~csi/Proyectos/csic2000>
- Curso “Sistemas de data warehousing”. Se adjunta el programa del curso.
- Curso “Sistemas de Data Warehousing y OLAP”. Se adjunta el programa del curso.