MDBE: Una Herramienta Automática para el Modelado Multidimensional

Oscar Romero
Dept.de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
oromero@lsi.upc.edu

Alberto Abelló
Dept.de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
aabello@lsi.upc.edu

1. Introducción

Los sistemas *Data Warehousing* tienen como objetivo ayudar en la toma de decisiones de las organizaciones. Para ello, estos sistemas integran y homogenizan los datos de la organización en un *Data Warehouse* (DW), con el objetivo de obtener una representación única y detallada de nuestro negocio.

Por definición, un DW es una gran base de datos que, por sí misma, es incapaz de aportarnos información relevante sobre nuestro negocio. Al igual que en otros sistemas de bases de datos, estos sistemas requieren herramientas que exploten el DW. En nuestro trabajo nos centramos en las herramientas OLAP (*On-line Analytical Processing*), que tienen como objetivo facilitar la navegación y el análisis de la información contenida en el DW, siguiendo el paradigma de la *multidimensionalidad*.

La multidimensionalidad se caracteriza por mostrar los datos como si éstos estuvieran situados en un espacio n-multidimensional, dónde los *hechos* (o sujetos de estudio) se analizan desde las diferentes perspectivas de análisis (o puntos de vista) de interés de nuestro dominio (las *dimensiones*). Los hechos contienen *medidas* de nuestro negocio (ventas, ingresos, etc.) mientras que las *dimensiones* se descomponen en diferentes *niveles* de detalle que nos permitirán modificar la granularidad de los datos. Un ejemplo clásico sería el de una venta (*hecho*) analizada desde el punto de vista del cliente, el vendedor, el producto vendido y la ubicación de la venta (*dimensiones*).

Como hemos comentado anteriormente, un DW no es más que el resultado de la integración y homogenización de los datos de nuestra empresa. Así pues, el esquema multidimensional de un DW debe derivarse de los esquemas de las fuentes de datos. Tradicionalmente, este proceso se ha desarrollado manualmente por un experto en sistemas DW, por lo que el esquema resultante dependía, en gran parte, de la habilidad y experiencia del experto. Este proceso, principalmente, puede llevarse a cabo de dos formas: o bien empezamos analizando los datos de los que dispone la organización para hacer emerger todo el conocimiento multidimensional posible disponible (a filtrar por los requisitos del usuario), o, justamente al revés, empezamos analizando los requisitos del usuario para posteriormente, intentar mapearlos sobre nuestras fuentes de datos.

Para facilitar el proceso de modelado multidimensional de un DW, en este trabajo presentamos MDBE (*Multidimensional Design By Examples*): nuestra propuesta de herramienta para validar requisitos multidimensionales proporcionados por el usuario final y expresados como consultas SQL sobre las fuentes de datos operacionales. MDBE descompone la consulta SQL de entrada para extraer el conocimiento multidimensional relevante que contiene y acorde con dicha información, deriva un conjunto de esquemas multidimensionales que satisfacen los requisitos (consultas) del usuario. Es decir, nos propone posibles esquemas multidimensionales de forma automática.

MDBE identifica conceptos multidimensionales tales como *hechos*, *medidas*, *dimensiones* y *niveles* de dimensiones a partir de conceptos

relacionales. De esta forma, nos da soporte a la hora de diseñar nuestros DW multidimensionales. Por último, cabe destacar que MDBE es el resultado de implementar las propuestas presentadas en [1] y [2].

2. Nuestra Propuesta: MDBE

Como hemos comentado en la sección anterior, MDBE descompone las consultas SQL entrantes (que representan requisitos multidimensionales de usuario mapeados sobre las fuentes de datos relacionales de la organización) para extraer el conocimiento multidimensional que éstas contienen. Dicho conocimiento, se almacena apropiadamente en un grafo que nosotros denominamos *grafo multidimensional*, tal y como se especifica a continuación (el lector puede encontrar las demostraciones pertinentes de cada uno de estos pasos en [2]):

- Primer paso: Cada una de las tablas del FROM de la consulta SQL representa un nodo del grafo multidimensional. A lo largo del proceso, cada uno de los nodos se etiquetará de acuerdo con el rol multidimensional que debe jugar de acuerdo con la consulta entrante. Si se encuentra alguna incoherencia en el etiquetaje, o no es posible hacerlo, podremos asegurar que dicha consulta no tiene significado multidimensional.
- Segundo paso: Si la consulta dispone de cláusula GROUP BY, ésta identifica *niveles* de *dimensión*.
- Tercer paso: Los campos agregados en la cláusula SELECT son identificados como *medidas*, y las tablas a las que pertenecen se identifican como *hechos*.
- Cuarto paso: Las selecciones multidimensionales (es decir, cualquier comparación en la clásula WHERE), identifican *niveles* de dimensión.
- Quinto paso: Así como los pasos anteriores se utilizan para etiquetar nodos del grafo, este paso crea y etiqueta aristas. Por cada *join* en la cláusula WHERE creamos una arista relacionando las tablas implicadas. Además, inferiremos la multiplicidad de dicha relación para poder determinar que roles multidimensionales pueden estar jugando los extremos (nodos) de la arista (que se etiqueta con los etiquetajes válidos de sus nodos). Si sólo existe una combinación válida, los nodos se etiquetan como corresponde y se propaga dicho conocimiento.

Nótese que estos pasos extraen el conocimiento multidimensional contenido en la consulta entrante. No obstante, este etiquetaje no nos garantiza un esquema multidimensional correspondiente válido; nos falta validar todos los etiquetajes como conjunto. Con ese objetivo, aplicaremos los siguientes pasos:

- El grafo debe ser conexo.
- Cada subgrafo de *niveles* debe representar una *dimensión* de análisis válida (cada nodo del subgrafo representará un *nivel*).
- Los *hechos* identificados en el grafo deben estar relacionados entre sí mediante relaciones multidimensionales válidas que eviten problemas de agregación en los datos multidimensionales (ésto es, *medidas*).
- Por último, se validan las conexiones entre subgrafos de *niveles* y los *hechos* para garantizar, en conjunto, que no habrá problemas de agregación de datos.

Si nuestro grafo ha superado con éxito todos estos pasos, el etiquetaje de los nodos denota los posibles esquemas multidimensionales que satisfacen los requisitos del usuario.

Referencias

[1] O. Romero and A. Abelló. Improving Automatic SQL Translation for ROLAP Tools. *Proc. of JISBD 2005*, 284(5):123–130, 2005.

[2] O. Romero and A. Abelló. Multidimensional Design by Examples. In *Proc. of DaWaK 2006*, volume 4081 of *LNCS*, pages 85–94. Springer, 2006.