Modelo para la Búsqueda de Metadatos con Características Semi-Inteligentes

Investigación

Dr. Enrique Luna Ramírez[1], Lic. Jorge H. Dzul Bermejo[2], Aarón Rangel Villalobos[3]
(1) Miembro del Cuerpo Académico de Sistemas Distribuidos del Instituto Tecnológico de Aguascalientes
(2) Tesista del Programa de Maestría en Ciencias Computacionales del Instituto Tecnológico de Toluca
(3) Tesista del Programa de Licenciatura en Informática del Instituto Tecnológico de Aguascalientes
Departamento de Sistemas y Computación del Instituto Tecnológico de Aguascalientes, Av. A. López Mateos 1801, Fracc. Bonagens, Aguascalientes, Ags., C.P. 20256 Tel. (449) 9105002 ext. 152, Fax (449) 9700423,
eluna@ita.mx, jdzul09@yahoo.com.mx, arangel23@hotmail.com

Resumen

En este proyecto de investigación se propone el diseño de un modelo para la búsqueda de metadatos con características semi-inteligentes, orientado a recuperar los metadatos asociados a un data warehouse de una manera rápida, flexible y confiable. El modelo incluye un conjunto de funcionalidades distintivas consistentes en el almacenamiento temporal de los metadatos de uso frecuente en un almacén distinto al almacén global de los metadatos de un data warehouse y en el uso de procesos de control para recuperar información de ambos almacenes mediante alias de conceptos.

Palabras Clave

Sistema de Soporte a la Decisión, Data warehouse, Almacenamiento y Recuperación de Metadatos, Alias de Conceptos.

Introducción

En las organizaciones actuales se genera una gran cantidad de información procedente de diversas fuentes, la cual debe ser integrada adecuadamente para que los responsables de la toma de decisiones puedan obtener un óptimo beneficio de ella. Con este propósito, durante los últimos años, varias compañías han comenzado a implantar diversas tecnologías de soporte a la decisión, entre las que se destaca la tecnología data warehousing como una estrategia para presentar una vista integrada y consistente de toda la información de una organización. Es decir, la información de las diferentes áreas de una organización es concentrada en un único almacén de datos (data warehouse) para una mejor explotación de la misma.

Es fundamental hacer énfasis en la importancia que tienen los metadatos no sólo en los sistemas data warehousing, sino en general en los sistemas de soporte a la decisión. Cuando se habla de los metadatos de un sistema de este tipo, se está haciendo referencia concretamente al *conocimiento* que se tiene sobre el sistema, tanto en su parte técnica como en su parte semántica. Por esta razón, en la medida que se tenga a la mano este conocimiento, se podrá obtener un mayor beneficio del sistema. Es importante señalar que a pesar de que la tecnología data warehousing no es un tema nuevo (inicio formalmente a principios de los años 90), la mayor parte de los productos e investigaciones actuales relacionadas con el tema no consideran de manera explícita la gestión de los metadatos [1,2], centrándose más en otros aspectos relacionados con la funcionalidad de los sistemas tales como la integración y explotación de datos, entre otros.

En conclusión, para obtener un mayor beneficio de la funcionalidad de un sistema data warehousing es necesario sacar más provecho de la potencialidad que tienen sus metadatos. Es decir, las operaciones sobre un data warehouse, tanto en el back-end como en el front-end, pueden ser realizadas más eficientemente con la disponibilidad oportuna y consistente de sus metadatos. Es por esta razón que hemos considerado conveniente y pertinente realizar una propuesta de solución para este problema tan poco abordado; cabe señalar que no sólo el estudio del estado del arte hace evidente esta problemática, sino también las propias experiencias que se han tenido en la región, donde no existe un sólo caso que presente una estrategia para almacenar el conocimiento de sus sistemas de manera ordenada, menos aun una estrategia para explotarlo convenientemente.

Teniendo en cuenta lo anterior, en este proyecto de investigación se propone el diseño de un modelo para la búsqueda de metadatos con características semi-inteligentes, con base en el cual se construirá en el futuro un prototipo que, como hipótesis de trabajo, permitirá recuperar los metadatos asociados a un data warehouse de una manera rápida, flexible y confiable. Además, nuestra propuesta será sometida a un análisis comparativo con las propuestas consideradas en el estado del arte.

Estado del Arte

Existen una infinidad de trabajos de investigación y desarrollo en el área de la tecnología data warehousing, pero muy pocos prestan atención al uso eficiente de los metadatos con propósitos de contribuir en el mejoramiento de la funcionalidad de un data warehouse. La inmensa mayoría de los trabajos existentes simplemente utilizan un repositorio para almacenar los metadatos y recuperarlos de manera ocasional. A continuación se describen brevemente los principales trabajos que hacen uso de una u otra forma de los metadatos con propósitos principalmente de calidad, seguridad o refrescamiento de un data warehouse.

Jarke et al. [3] describen un enfoque para gestionar la calidad de un data warehouse mediante su repositorio de metadatos. Los autores proponen agregar a este repositorio funciones de soporte a la calidad con base en el enfoque genérico GQM ("Goal-Question-Metric"), desarrollado para gestionar la calidad del software. De esta manera, para gestionar la calidad de un data warehouse, se realizan preguntas (consultas) al repositorio sobre la calidad de un determinado componente o un determinado proceso, y éste responde con métricas de calidad obtenidas a través de agentes que se comunican con los diferentes componentes y procesos del data warehouse.

Katic et al. [4] describen un enfoque para incrementar la seguridad de un data warehouse basado en metadatos. En su propuesta, los autores utilizan aquellos metadatos que describen los mecanismos de seguridad de un data warehouse. De esta manera, el enfoque consiste en presentar vistas reducidas del data warehouse de acuerdo al perfil de usuario.

Martin et al. [5] describen un enfoque para consultar fuentes pasivas de datos (fuentes que no cuentan con un motor de búsqueda) basado en la extracción de los metadatos de este tipo de fuentes. Su propuesta se centra en el diseño de un esquema para capturar las estructuras de los datos y las relaciones entre éstos, el cual es usado para construir herramientas que permitan extraer los metadatos y almacenarlos en un repositorio. De esta manera, un usuario puede localizar datos de interés en las fuentes pasivas mediante consultas directas al repositorio de metadatos.

Nelson [6] propone un modelo para alimentar un repositorio de metadatos con información proveniente de las aplicaciones que interactúan con él, y que éstas a su vez puedan recuperar información del mismo. El modelo se compone de clases, asociadas entre sí, cada una de las cuales representa a un componente del repositorio. Cabe mencionar que su diseño soporta solamente consultas de datos estadísticos, aunque, de acuerdo al autor, el modelo puede ser adaptado para recuperar datos multidimensionales (propios de un data warehouse).

Vavouras et al. [7] describen un enfoque para modelar y ejecutar el proceso de refrescamiento de un data warehouse basado en especificaciones sobre este proceso, las cuales son almacenadas en el repositorio de metadatos. Como parte de su propuesta, los autores definen un componente llamado "Data Warehouse Refresh Manager", usado para gestionar las tareas que deben realizarse durante el proceso de refrescamiento. Este componente está compuesto por el repositorio de metadatos mismo y por subcomponentes de extracción, transformación y carga de datos.

Modelo Propuesto

Tomando como base la forma en que el cerebro humano procesa la información [8], en la Figura 1 se muestra una abstracción para modelar la estructura de un repositorio de metadatos.

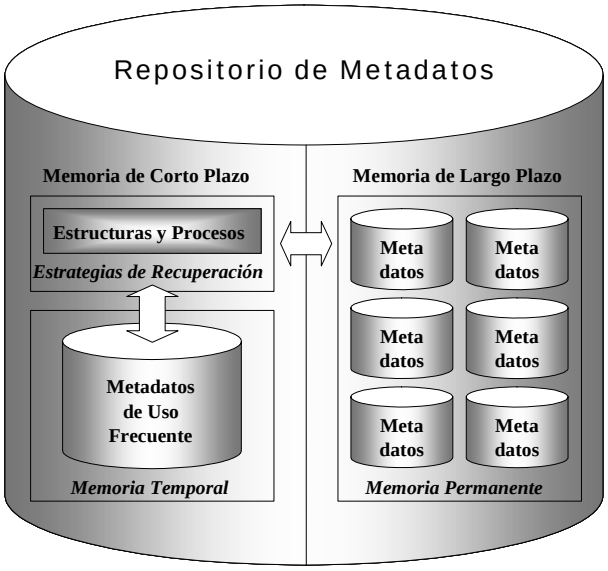

Figura 1. Enfoque para modelar la estructura de un repositorio de metadatos.

De acuerdo a esta estructura, los metadatos de un sistema pueden ser almacenados en una memoria permanente (compuesta comúnmente por almacenes separados), mientras que los metadatos de uso frecuente pueden ser almacenados en una memoria temporal para que su recuperación sea más rápida.

El hecho de que los metadatos de uso frecuente puedan ser recuperados más rápidamente cuando son almacenados en una memoria temporal se debe principalmente a la diferencia significativa en el tamaño de ambas memorias (permanente y temporal). De esta manera, cuando se realiza una consulta sobre el repositorio, las estrategias de recuperación acuden primeramente a la memoria temporal, dejando siempre como segunda opción a la memoria permanente. Es decir, esta última memoria es utilizada sólo en caso de que la información buscada no sea encontrada en la memoria temporal.

Por su parte, las estrategias de recuperación son estructuras y procesos usados por el cerebro humano para recuperar información de las memorias temporal y permanente con la finalidad de dar respuesta a los estímulos externos. Con base en esta idea y en la estructura de repositorio propuesta, se diseñó nuestro modelo para la búsqueda de metadatos, cuyo esquema conceptual se muestra en la Figura 2.

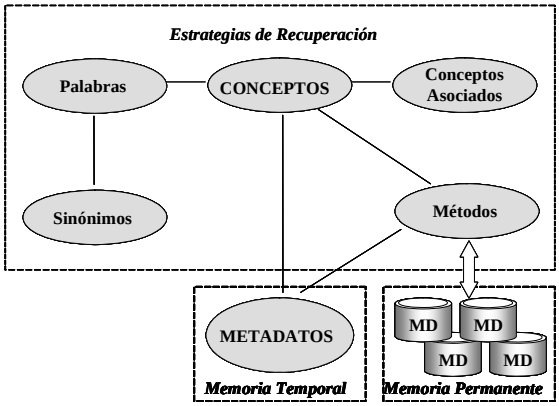

Figura 2. Modelo propuesto para la búsqueda de metadatos.

El modelo propuesto se compone de estructuras y procesos que guardan una dependencia funcional, ya que las estructuras almacenan la información necesaria para que los procesos puedan llevar a cabo su tarea de recuperar metadatos. Como se puede observar en la figura anterior, el punto central del modelo es la estructura denominada *CONCEPTOS*, la cual almacena todos los conceptos que conforman los metadatos de un data warehouse. Es decir, esta estructura almacena frases (cada una de las cuales identifica a un concepto) relacionadas con conjuntos de metadatos en la memoria permanente, mismos que pueden ser extraídos y depositados en la memoria temporal mediante los métodos correspondientes.

Por su parte, la estructura *Palabras* almacena las palabras significativas que son parte de un concepto (visto como una frase), entendiéndose por palabra significativa aquella palabra que no es artículo ni conectivo; cada concepto está asociado a las palabras significativas que éste incluye. La estructura *Sinónimos*, como su nombre lo indica, almacena los sinónimos de las palabras contenidas en la estructura Palabras, por lo que cada palabra significativa está asociada a todos sus sinónimos. Esta estrategia servirá para localizar conceptos mediante alias. Respecto a la estructura *Conceptos Asociados*, ésta almacena aquellos conceptos que proveen información adicional sobre los conceptos en la estructura CONCEPTOS. Dicho de otra manera, esta estructura almacena los conceptos asociados a otros conceptos con base en una relación de afinidad.

La parte estática del modelo concluye con las estructuras Métodos y METADATOS. La estructura denominada *Métodos*, almacena los métodos necesarios para extraer los metadatos en la memoria permanente asociados a un determinado concepto y depositarlos en la memoria temporal. Tales métodos consisten básicamente en componentes, páginas dinámicas de hipertexto o programas ad hoc, dependiendo del tipo de información que deba ser extraída. Por su parte, la estructura *METADATOS* almacena los metadatos extraídos de la memoria permanente, por lo que esta estructura corresponde propiamente a la memoria temporal del modelo.

La parte dinámica del modelo está conformada por los procesos que operan sobre las estructuras antes descritas. Para encontrar un concepto y recuperar sus metadatos asociados, éste se descompone en las palabras que lo forman, eliminándose aquellas que no son significativas y localizándose, en la estructura Palabras, aquellas que si lo son. Como se mencionó previamente, las palabras significativas están asociadas a sus sinónimos, por lo que el concepto buscado puede ser localizado ya sea directamente, o ya sea a través de un alias, en la estructura CONCEPTOS. Así, con este modelo no es necesario escribir un concepto en la forma exacta (ni siquiera parecida) a como está almacenado para poder encontrarlo, lográndose así una gran flexibilidad durante el proceso recuperación de información.

Una vez que el concepto buscado es localizado, se procede a identificar los conceptos asociados a éste (en caso de que existan) y posteriormente se procede a recuperar sus metadatos de la memoria temporal. No obstante, como se mencionó anteriormente, tales metadatos deben ser extraídos de la memoria permanente a través de los métodos correspondientes en caso de que no se encuentren en la memoria temporal.

Conclusiones y Trabajo Futuro

El avance de investigación presentado en este artículo consistió en el diseño conceptual de un motor de búsqueda de metadatos con características semi-inteligentes, orientado a sistemas de soporte a la decisión, particularmente a data warehouses. Para lograr esto, en nuestra propuesta se plantea el uso de un almacén exclusivo para mantener los metadatos que son usados con mayor frecuencia en el entorno de un sistema y se definen estructuras y procesos de control para recuperarlos mediante alias. Esta es una de las principales características de nuestro modelo, la cual implica de antemano un problema de combinatoria relativamente complejo para su implementación.

Nuestro modelo será evaluado en principio mediante una comparación analítica del mismo con los diferentes modelos que conforman el estado del arte, con base en características definitorias tales como el dominio de aplicación y los tipos de metadatos contemplados.

Agradecimientos. Queremos agradecer al *Cuerpo Académico de Sistemas Distribuidos* del Instituto Tecnológico de Aguascalientes por brindar todos los recursos necesarios para llevar a cabo el desarrollo de este trabajo.

Referencias

[1] Marco, D. (2000) *Building and Managing the Meta Data Repository*, Wiley.

[2] Mundy, J. and Thornthwaite, W. (2006) *The Microsoft Data Warehouse Toolkit with SQL Server 2005 and the Microsoft Business Intelligence Toolset*, Wiley.

[3] Jarke, M., Jeusfeld, M. A., Quix, C., and Vassiliadis, P. (2003) "Architecture and Quality in Data Warehouses: An Extended Repository Approach", *Information Systems*, 24: 3, pp. 229-253.

[4] Katic, N., Quirchmayr, G., Schiefer, J., Stolba, M., and Tjoa, A. M. (2001) "A Prototype Model for Data Warehouse Security Based on Metadata", *Proceedings of the 9th International Conference on Database and Expert Systems*, pp. 300-308.

[5] Martin, P., Powley, W., Weston, A., and Zyon, P. (2004) "Using Metadata to Query Passive Data Sources", *International Journal of Cooperative Information Systems*, 9: 1-2, pp. 147-169.

[6] Nelson, C. (2005) "Use of Metadata Registries for Searching for Statistical Data", Proceedings of the 14th International Conference on Scientific and Statistical Database Management.

[7] Vavouras, A., Gatziu, S., and Dittrich, K. R. (2002) "Modelling and Executing the Data Warehouse Refreshment Process", *Proceedings of the International Symposium on Database Applications in Non-Traditional Environments*, pp. 66 -73.

[8] Pfeifer, R. and Scheier, C. (1999) *Understanding Intelligence*, The MIT Press.