DESIGN OF A WEB SEMI-INTELLIGENT METADATA SEARCH MODEL APPLIED IN DATA WAREHOUSING SYSTEMS

DISEÑO DE UN MODELO SEMI-INTELIGENTE DE BÚSQUEDA DE METADATOS EN LA WEB, APLICADO A SISTEMAS DATA WAREHOUSING

Enrique Luna Ramírez[1] Humberto Ambriz Delgadillo[2] J. Antonio Nungaray Ornelas[3]
Francisco Javier Álvarez Rodríguez [4] Jorge N. Mondragón Reyes[5]

Recibido 18 de abril de 2007, aceptado 18 de noviembre de 2008
Received: April 18, 2007 Accepted: November 18, 2008

RESUMEN

En este artículo se propone el diseño de un modelo para la búsqueda Web de metadatos con características semi-inteligentes. El modelo ha sido concebido para recuperar de manera rápida, flexible y fiable los metadatos asociados a un data warehouse corporativo. Nuestra propuesta incluye un conjunto de funcionalidades distintivas consistentes en el almacenamiento temporal de los metadatos de uso frecuente en un almacén exclusivo, diferente al almacén global de metadatos, y al uso de procesos de control para recuperar información de ambos almacenes a través de alias de conceptos.

Palabras clave: Búsqueda de metadatos, data warehousing.

ABSTRACT

In this paper, the design of a Web metadata search model with semi-intelligent features is proposed. The search model is oriented to retrieve the metadata associated to a data warehouse in a fast, flexible and reliable way. Our proposal includes a set of distinctive functionalities, which consist of the temporary storage of the frequently used metadata in an exclusive store, different to the global data warehouse metadata store, and of the use of control processes to retrieve information from both stores through aliases of concepts.

Keywords: Metadata search, data warehousing.

INTRODUCTION

In current organizations, a large amount of information originating in diverse sources is generated. This information should be integrated adequately, so that decision makers can obtain an optimum benefit of it. With this purpose, during the last years, several companies have begun to implement diverse decision support technologies. Particularly, the data warehousing technology is emphasized as a strategy to present an integrated and consistent view of all the information of an organization. That is, the information of the different areas of an organization is concentrated in a unique data warehouse for a better exploitation of it.

It is fundamental to emphasize the metadata importance, not only in a data warehousing system, but in any decision support system. Metadata means the *data on data* of a system, on its technical part as well as its semantics. Thus, as we use the metadata of a system, we will be able to obtain a greater benefit of that system. It is important to mention that despite the fact that the data warehousing technology is not a new topic (it started formally at the beginnings of the 90´s [1]), most of the commercial products and current research woks related to this topic do not consider explicitly the metadata management [2-4], being focused more in other aspects such as the quality, security and refreshment of a data warehouse, than in metadata itself.

[1] Instituto Tecnológico El Llano Aguascalientes, México. E-mail: eluna@itllano.edu.mx
[2] Instituto Tecnológico El Llano Aguascalientes, México. E-mail: hambriz@itllano.edu.mx
[3] Instituto Tecnológico El Llano Aguascalientes, México. E-mail: janungaray@itllano.edu.mx
[4] Universidad Autónoma de Aguascalientes, México. E-mail: fjalvar@correo.uaa.mx
[5] Secretaría de Gestión e Innovación del Estado de Aguascalientes, México. E-mail: jmondragon@aguascalientes.gob.mx

In few words, to obtain a greater benefit of the functionality of a data warehousing system, it is necessary to take advantage of the potentiality that its metadata possess. That is, the operations on a data warehouse, at the back-end as well as at the front-end, can be carried out more efficiently with the consistent and opportune availability of its metadata. In this sense, we have considered convenient and pertinent to carry out a proposal of solution for this problem. It is worth to mention that not only the study of the state of the art showed us this problem, but also the common lack of clear strategies to store and retrieve the metadata of corporative decision support systems.

Based on this, in this research project, the design of a metadata search model with semi-intelligent features is proposed, whose hypothesis is the fast, flexible and reliable retrieval of the metadata associated to a data warehouse. It is important to mention that this hypothesis will be evaluated in another phase through the construction of a whole prototype and its application to real cases, so that in this paper only the conceptual design of the metadata search model and an initial prototype will be discussed.

STATE OF ART

There is a variety of research works and technological developments in the data warehousing technology field, but very few of them pay attention to an adequate metadata use for improving a data warehouse functionality. Next, the main approaches that make an explicit use of the metadata in this type of systems are described briefly. These approaches discuss specific data warehouse aspects, but, as mentioned above, they do not focus on metadata itself.

Jarke, Jeusfeld, Quix and Vassiliadis [5] describe an approach to manage the quality of a data warehouse by means of its metadata repository. The authors propose adding quality support functions to this repository based on the generic approach GQM ("Goal-Question-Metric"), developed to manage software quality. In this way, to manage the quality of a data warehouse, questions (queries) on the quality of a certain component or process are asked to the metadata repository, and it responds with metrics of quality obtained through agents that communicate with the different components of the data warehouse.

Katic, Quirchmayr, Schiefer, Stolba and Tjoa [6] describe an approach to increase the security of a data warehouse based on metadata. In their proposal, the authors use the metadata that describes the security mechanisms of a data warehouse. Thus, the approach consists of presenting to users reduced views of the data warehouse depending on their profile.

Martin, Powley, Weston and Zyon [7] describe an approach to consult passive data sources (sources that do not include a search engine) based on the extraction of their metadata. This proposal focuses in the design of a meta-schema to capture the data structures, as well as the relations among data. The meta-schema is used to build tools that permit to extract the passive sources metadata and store them in a repository. In this way, a user can locate data of interest in this type of sources by making queries directly against the metadata repository.

Nelson [8] proposes a model to feed a metadata repository with information coming from the applications that interact with it and, at the same time, these applications can retrieve information from the repository. The model is composed of classes associated among themselves, each of which represents a repository component. It is worth to mention that this model is oriented only to the statistical data retrieval, although, according to the author, it can be adapted to retrieve multidimensional data.

Vavouras, Gatziu and Dittrich [9] describe an approach for modeling and executing the data warehouse refreshment process based on specifications on this process, which are stored in the metadata repository. As the main part of their proposal, the authors define a component called "Data Warehouse Refresh Manager", used to manage the tasks that should be carried out during the refreshment process. This component is composed of the metadata repository itself and subcomponents for extracting, transforming and loading data into the data warehouse.

It is important to mention that besides the approaches focused on data warehousing aspects; there exist others that consider concept schemes to retrieve information in the Semantic Web context. As an example, Miles et al. [10] provide a model for expressing the basic structure and content of concept schemes such as thesauri, subject heading lists, taxonomies, classification schemes, and other similar types of controlled vocabulary. SKOS (Simple Knowledge Organization System) [11], an implementation of this model, allows concepts to be composed and published on the World Wide Web, linked with data on the Web and integrated into other concept schemes. As it will be discussed, our proposal is also based on concepts; however, its purpose is different. That is, it is not oriented to the Semantic Web.

PROPOSED MODEL

Based on the human information processing [12], an approach for modeling the structure of a metadata repository is shown in figure 1. This figure shows how the repository has been structured in two blocks corresponding to the short-term and long-term memories of the human brain, the first one composed of a *temporary memory* and the so-called *retrieval strategies*, while the second one basically consists of a *permanent memory* that accumulates all the knowledge acquired through a lifetime. This structure will be the base for designing our metadata search model.

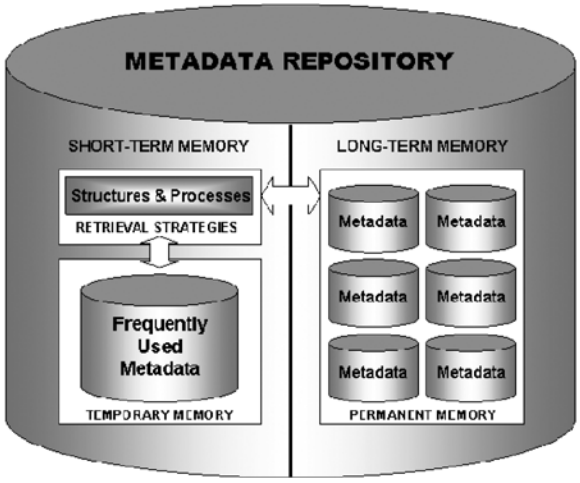

Figure 1. Approach for modeling the structure of a metadata repository.

All the metadata that belong to a data warehouse are stored in the permanent memory. These metadata can be stored in different databases (metadata bases), each of which may be structured differently. Some of these metadata will surely be used more frequently, either because of their importance or for any other reason. So, these recurrent metadata are extracted from the permanent memory and stored in the temporary memory through the retrieval strategies. The idea behind this approach is to enable recurrent metadata to be retrieved faster. That is, when a query is made against the metadata repository, the retrieval strategies turn first to the warehouse corresponding to the temporary memory, which is often significantly smaller than the set of warehouses corresponding to the permanent memory.

According to Pfeifer and Scheier [12], the retrieval strategies are structures and processes used by the human brain to retrieve information from the permanent and temporary memories with the purpose of giving response to external stimuli. Based on this idea and on the metadata repository structure proposed previously, our metadata search model was designed (see figure 2).

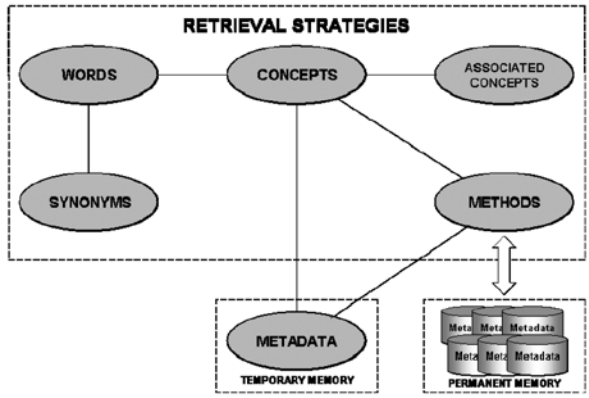

Figure 2. Metadata search model.

The model proposed is composed of structures and processes that have a functional dependence among them, since the structures are used to store the information that the processes need to carry out the search of metadata. This is described in detail next, starting with the model static part, composed of the structures shown in figure 2: CONCEPTS, ASSOCIATED CONCEPTS, WORDS, SYNONYMS, METHODS and METADATA.

The central point of the model is the CONCEPTS structure, which stores all the concepts that conform the metadata belonging to a corporation data warehouse. That is, this structure stores sentences (concepts) related to metadata sets in the permanent memory. These metadata can be extracted and stored in the temporary memory through adequate methods (components, dynamic hypertext pages and ad hoc programs). Thus, a concept can be related to certain types of metadata such as, to mention but a few examples, its own definition, data models, executed processes during data load and existing business reports in the data warehouse.

The WORDS structure stores the meaningful words that at any given time can form part of a concept (viewed as a sentence), while the SYNONYMS structure stores the synonyms of the words contained in the former structure. It is important to mention that articles and connectives are not meaningful words. This idea, as will be discussed shortly, will serve to locate concepts through alias. With respect to the ASSOCIATED CONCEPTS structure, this stores those concepts that provide additional information on some concepts in the CONCEPTS structure, according to a relation of similarity. To conclude with the static part of the model, the METHODS and METADATA structures are discussed next. The METHODS structure stores the

necessary methods to extract, from the warehouses that compose the permanent memory, the metadata associated to a given concept and place them in the METADATA structure (the temporary memory). Such methods consist basically in components, dynamic hypertext pages or ad hoc programs, depending on the type of information that should be extracted.

The dynamic part corresponds to the processes that operate on the structures above-described. So, to find a concept and retrieve its associated metadata, the concept sought is decomposed into the words of which it is composed, removing any that are not meaningful (articles and connectives). Each one of these meaningful words is located in the WORDS structure, and then the corresponding synonyms are detected in order to derive aliases of the concept sought. In this way, a concept can be located either through an alias, or directly, in the CONCEPTS structure. Once the concept is located, its associated concepts are detected (in case they exist), so that the metadata associated to any of these concepts can be retrieved from the METADATA structure. If the value corresponding to the concept sought is not found in this structure, it is extracted from the metadata bases by using the corresponding method(s) in the METHODS structure. Thus, with this model, it is not necessary to search a concept in the exact form (not even similar) how it is stored, what provides a great flexibility during the metadata search process.

INITIAL PROTOTYPE

Currently a prototype based on the model proposed in the previous section has been built. To follow the trends in the information systems field, and particularly in the data warehousing systems, the prototype is being built as a Web application supported by the Apache server [13], the MySQL database management system [14] and the PHP scripting language [15]. In the following, some representative examples of metadata search are completed using our prototype.

The first example refers to the search of a typical concept associated to every data warehousing system: its data model. This concept, like any other, can be sought in any way a user wants to, on the condition of searching something related to the concept. Thus, this concept can be sought as "data model", "data warehouse model", or even, "model" or "data". As shown in figure 3, the concept was sought simply as "datos" (Spanish version), whose results were those concepts that contain this word. Of course, like with any search engine, this search can be carried out more specifically in order to obtain a more reduced list of results.

It is important to observe that for the "data model" concept, the standard concept stored in our metadata repository is "Data Warehouse Model" ("Modelo del Data Warehouse" in Spanish), which was located through an alias, considering that "datos" is synonymous with "data".

Figure 3. Location of the concept "Data Warehouse Model" through an alias.

Proceso ETL	Tabla de Destino	Tipo de Tabla	Tiempo de Procesamiento	Fecha de Carga	Hora de Carga
etl007	Artículo	Dimensión	00:37:10	15-May-06	11:37 p.m.
etl123	Marca	Dimensión	00:24:25	15-May-06	11:24 p.m.
etl182	Segmento	Dimensión	00:18:15	15-May-06	11:18 p.m.
etl188	Sucursal	Dimensión	00:19:27	15-May-06	11:19 p.m.
etl205	Ventas_Diarias	Hecho	02:23:07	15-May-06	01:23 a.m.
etl244	Zona	Dimensión	00:32:18	15-May-06	11:32 p.m.

Figure 4. Location of the concept "Load of Data".

The first metadata associated to a concept is its definition and then, through a link, other metadata are provided (diagrams, tables, etc.). Notice that the "Data Warehouse Model" concept has two associated concepts ("2 Conceptos Asociados" in Spanish) that provide additional information on the data model.

Another typical concept associated to the back-end of a data warehouse is the "load of data" concept, whose location is shown in figure 4. This time the search was performed in a more specific way, being obtained the standard concept "Data Load Statistics" ("Estadísticas de Carga de Datos" in Spanish). This concept was also obtained in the previous example, since it contains the word "data", which shows the flexibility of our model.

The "load of data" concept is naturally associated to all that information (metadata) on the process of extraction, transformation and load of data from the data sources to the data warehouse.

This process, commonly known as ETL process (Extraction, Transformation and Load), provides statistics on the sub-processes and tables involved in the whole process, the type of each table (dimension or fact), and the date and time of data load.

CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a metadata search model with semi-intelligent features, oriented at efficiently retrieving the metadata of a corporative data warehouse. Our proposal consists of searching concepts through aliases and retrieving the metadata associated to them. This allows to carry out semi-intelligent searches in the sense that it is not necessary to search a concept in the exact form (not even similar) how it is stored, providing in this way a great flexibility during the metadata search process. With a view to achieving this objective, the model, composed of an approach for modeling the repository structure and by a metamodel for storing and retrieving metadata, is based on the human information processing paradigm. So, our model considers a set of distinctive functionalities that can be built into a metadata repository system to assure that it works efficiently. These functionalities refer to the use of two memories for storing the repository metadata and a set of structures and processes for retrieving the information passing from one memory to another. One of the memories in particular is used to store the most recurrent metadata in a corporate environment, which can be rapidly retrieved with the help of the above-mentioned structures and processes. It is important to mention that these structures and processes can also serve to contextualize the information of the different business areas of a data warehouse. In this sense, template and scenario querying is the next stage we are going to carry out, for retrieving sets of metadata that belong to a same business area or project.

Currently, we have developed a prototype based on the proposed model, built as a Web application supported by the Apache HTTP server, the MySQL database management system and the PHP scripting language. So

far, we have managed to implement concept query and associated metadata retrieval, as discussed in this paper, showing the potentiality of the prototype by means of the examples presented.

ACKNOWLEDGEMENTS

Thanks to the reviewers for their pertinent comments and to *Sistema Nacional de Investigadores (SNI)* of Mexico for its support to the development of this research work.

REFERENCES

[1] W.H. Inmon. "Building the Data Warehouse". Wiley. 1992.

[2] D. Marco. "Building and Managing the Meta Data Repository". Wiley. 2000.

[3] J. Mundy and W. Thornthwaite. "The Microsoft Data Warehouse Toolkit with SQL Server 2005 and the Microsoft Business Intelligence Toolset". Wiley. 2006.

[4] J.C. Hancock and R. Toren. "Practical Business Intelligence with SQL Server 2005". Addison-Wesley. 2007.

[5] M. Jarke, M.A. Jeusfeld, C. Quix and P. Vassiliadis. "Architecture and Quality in Data Warehouses: An Extended Repository Approach". Information Systems. Vol. 24 Nº 3, pp. 229-253. 2003.

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

[7] P. Martin, W. Powley, A. Weston and P. Zyon. "Using Metadata to Query Passive Data Sources". International Journal of Cooperative Information Systems. Vol. 9 Nº 1-2, pp. 147-169. 2004.

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

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

[10] A. Miles, B. Matthews, M. Wilson and D. Brickley. "SKOS Core: Simple Knowledge Organisation for the Web". Proceedings of the International Conference on Dublin Core and Metadata Applications: Vocabularies in Practice, pp. 313-321. 2005.

[11] A. Isaac and E. Summers. "S. Simple Knowledge Organization System Primer. W3C Working Draft". August 29, 2008. Date of visit: October 31, 2008. URLs: http://www.w3.org/TR/skos-primer/

[12] R. Pfeifer and C. Scheier. "Understanding Intelligence". The MIT Press. 1999.

[13] Apache HTTP Server Project, version 2.2.10. Date of visit: October 2008. URLs: http://httpd.apache.org/

[14] MySQL version 4.1. Date of visit: October, 2008. URLs: http://www.mysql.com/

[15] PHP version 4.4.9. Date of visit: October 31, 2008. URLs: http://www.php.net/