A METADATA MODEL FOR KNOWLEDGE DISCOVERY IN DATABASE

José Rafael Carvalho and Maria Madalena Dias
Departamento de Informática, Universidade Estadual de Maringá, Av. Colombo 5790, 870200-900 Maringá-PR, Brazil

Keywords: Knowledge Discovery in Database, Data Warehouse, Metadata, XML.

Abstract: Metadata are deeply necessary in an environment of Knowledge Discovery in Database (KDD), once they are the responsible for the whole documentation of information on data that integrate a data warehouse (DW), being the latter used to store the data about the organization business. Such data usually come from several data sources, thus the metadata format should be independent on the platform. The system interoperability can be solved, by using XML (Extensible Markup Language). Therefore, in the present paper a metadata model for KDD, in XML format, and a manager of metadata are presented. The manager was implemented in Java, which provides support to the model presently proposed.

1 INTRODUCTION

The growing demand for systems to support decision-making has impelled a search for both, new technologies and methods to solve problems relative to the treatment of the great volume of existing data, in bases maintained by transactional systems, and the search for information and knowledge.

The knowledge discovery in database appeared as a process that establishes how to organize the data, as well as to extract relevant knowledge to make decisions.

In KDD systems, the knowledge for decision-making are obtained by applying data mining techniques that can be accomplished on data comprised in databases of transactional systems, after being integrated and transformed. However, the ideal is to build a DW, where data can be prepared to serve as entrance for the application of those techniques, thus the DW can also be used by OLAP tools (On-line Analytical Processing) (Chaudhuri and Dayal, 1997).

Bearing in mind the importance of DW, it is necessary to know the information on the data that comprise it, that is, how they are integrated and transformed to be recorded in the DW, and how they can be accessed. Such information should be registered in a metadata. Therefore, it is necessary to define a metadata model to represent the organization of its content.

The use of XML format, in the definition of the metadata model, enables to be more independent from the computational platform. Thus, the application receives a fundamental characteristic for any software, that is, the so searched independence from any product manufacturing company.

To meet this goal and others on the quest for knowledge in database, a research project is under development and partial results are presented in this paper, which are: a metadata in the XML format to meet the needs of record and access to data about the data involved in KDD systems, and a software component to manage this metadata.

A software component was developed to validate the defined model and to make possible the registration and the access of information in the metadata. That component was implemented in Java language, because that language is free and supplies portability.

In the next section of the present paper some related works are presented. In Sections 3 and 4 the metadata model proposed, and the manager of metadata are described, respectively. Finally, in the last section, conclusions regarding the present investigation and some suggestions for future studies are presented.

Rafael Carvalho J. and Madalena Dias M. (2008).
A METADATA MODEL FOR KNOWLEDGE DISCOVERY IN DATABASE.
In *Proceedings of the Tenth International Conference on Enterprise Information Systems - DISI*, pages 469-472
DOI: 10.5220/0001702404690472

2 RELATED WORKS

Tannenbaum (2002) presented a generic metadata model. In that model the initial step is to identify the beneficiaries, by using the metadata suggested, as being: developers, TI (Technology of Information) project managers, final users and DBMS (Database Management System) catalogs. The next step is to relate the sources of metadata, that is, to know where the origin of a metadata requirement comes from. The last step is to classify the metadata in three possible categories, as follows: specific one, only one and common category.

Castoldi (1999) proposed a relational model of specific metadata for KDD systems that offers support to the data preparation phase, including definition of tables, sources of data, definition of transformation rules and load in the DW, definition of the frequency with which each rule should be executed, and the metadata documentation.

Huynh et al. (2000) proposed a metadata that shows a mapping between object environment and relational environment in metadata layer of an O-R data warehouse.

Melchert et al. (2005) defined a CWM-based model of DW metadata that allows for the integration of metadata from different software tools within the DW system. The purpose of the paper was reporting the experience made with the application of the CWM pattern within the bank's metadata management project.

3 THE MODEL PROPOSED

The models proposed for Tannembaum (2002) and Castoldi (1999), together with the CWM (CWM, 2001) pattern, supplied subsidies for the definition of a new metadata model in XML to KDD system. Besides that, that model is applicable to the DW architecture proposed by Menolli and Dias (2006).

In the model proposed, all the necessary information for the general understanding of a KDD system is available to the users, such as: the sources of data that feed DW, the data comprised in the DW, the relationship between the precedence or origin data and DW data, etc. Thus, the metadata serve as a guide in the creation, load and access to DW. Items of metadata identified as being fundamental to the proposed model are listed below:

1) Pre-processing:
   a) Search:
      - Addresses of data sources;
      - Type of DBMS;
      - Version of DBMS;
      - User name/password to access the DBMS;
      - Database name;
      - Names of tables, attributes, and keys;
      - Names of aliases;
      - Names of the drivers and APIs;
      - Types of source files ;
      - Names of legacy systems.
   b) Transformation:
      - Items of data from source database;
      - Rules of transformation;
      - Transformations made.
   c) Load:
      - Stage database:
         - Items of data processed;
         - Information on stage database;
         - Frequency of load;
         - Volume of data;
         - Date of the last load;
         - User login.
   d) DW:
      - Items of data from stage database;
      - Information on DW database;
      - Rules of transformation;
      - Transformations made.
      - Frequency of load;
      - Volume of data;
      - Date of the last load;
      - User login.

2) Implementation of Technical Analysis:
   a) Items of data from DW;
   b) Information on data mining algorithms;
   c) Information to access data mining algorithms;
   d) Information on OLAP tools;
   e) Information to access OLAP tools.

3) Post-processing:
   a) Items of data from DW used to implementation of technical analysis;
   b) Technical analysis applied in the generation of results;
   c) File name of results;
   d) File address of results;
   e) Business name.

The XML of the metadata was structured by naming the three main nodes, as being the main stages of KDD process (Feldens et al., 1998), as follows: 'Pre-processing', 'Application of Analysis Techniques' and 'Post-processing'. However, it was necessary to adapt the model proposed to a metadata pattern for KDD systems. CWM was the pattern chosen, due to the fact of being one of the most suitable approaches in the DW domain. It has been constructed as a metadata pattern for this very application domain and it is already supported by a large number of DW software vendors (Melchert et al., 2005).

Only the main packages (Foundation, Resource, Analysis and Management) of CWM were used. However, there is the possibility of other packages are supported by this model. This is a proposal for future work.

4 THE MANAGER OF METADATA

A manager of metadata was developed, as being a software component, to give support to the registration and access to all information contained in the metadata here proposed. Such a component is based on the reference architecture proposed by Valentin (2006). That architecture is divided into layers, and it is based on architecture oriented to service, named SOA (Service Oriented Architecture). Thus, a layer is defined, in order to join the business processes, whereas another layer is used for services regarding application. Besides that, other layers offer support for the accomplishment of some services, among them, the services related to the metadata management.

Some UML diagrams were built in the design of the manager of metadata. Figure 1 shows the Diagram of Classes, where classes represent the model tags in XML.

A case study was realized using the manager of metadata and a XML was generated to register data about the data involved in the KDD process. After, the XML was used to realize operations of simple queries about the organization business. More details can be found in (Carvalho, 2007).

5 CONCLUSIONS

The metadata model described in the present paper is based on CWM pattern; on metadata models proposed by Tannenbaum (2002) and Castoldi (1999); on the main stages of the process of development of a KDD System defined by Feldens et al. (1998); on the DW architecture defined by Menolli and Dias (2006), and in addition, on the reference architecture by Valentin (2006). In the metadata model the fundamental information are represented for a KDD system, that is, the techniques (e.g.: the ones for finding the tables) used as sources of data for DW, as well the management techniques (e.g. for getting information on the date of the last load accomplished in DW).

The XML language was used in order to provide the interoperability of the metadata model proposed, beyond the matter concerning performance, because the metadata persistence are in XML, and due to that, it reduces the overhead, and justifies its use.

Therefore, the main contribution of the present paper is the definition of a metadata model in XML, that it represents the items of fundamental metadata to a KDD system. The main goal of the work described in this paper was defining a physical and portable model of metadata to support all the activities of the KDD process, to make easy the registration and the access of data of a DW, that it is used in KDD systems.

Considering the possibility of further investigations, some aspects can be suggested, such as: a more detailed study/investigation, and the use of a CWM pattern in a thorough way, and in addition, the development of a KDD system that includes the component of metadata management here proposed.

REFERENCES

Carvalho, J. R., 2007. Modelo de metadados para sistemas de descoberta de conhecimento em banco de dados utilizando XML. Programa de Pós-Graduação em Ciência da Computação. *Dissertação de Mestrado*. Departamento de Informática. Universidade Estadual de Maringá.

Castoldi, A. V., 1999. Um modelo de metadados para suporte a sistemas de descoberta de conhecimento em banco de dados. *Monografia de Graduação* – Ciência da Computação. Departamento de Informática. Universidade Estadual de Maringá.

Chaudhuri, S., Dayal, U., 1997. An overview of data warehousing and OLAP Technology**.** *ACM SIGMOD.* New York. Vol. 26. N° 1.

CWM, 2001. *Common warehouse metamodel specification*. OMG (Object Management Group). Version 1.0.

Feldens, M. A., Moraes, R. L., Pavan, A.; Castilho, J. M. V., 1998. Towards a methodology for the discovery of useful knowledge combining data mining, data warehousing and visualization. In *CLEI'98 (XXIV Conferencia Latino-Americana de Informática).* Quito. Equador.

Huynh, T. N., Mangisengi, O., Tjoa, A M., 2000. Metadata for object-relational data warehouse. In

International Workshop on Design and Management of Data Warehouse (DMDW12000). Stockholm, Sweden.

Melchert, F., Schwinn, A., Herrmann, C., Winter, R., 2005. Using reference models for data warehouse metadata management. In *Eleventh Americas Conference on Information Systems.* Omaha, NE, USA.

Menolli, A. L. A., Dias, M. M., 2006. A data warehouse architecture in layers for science and technology. In *SEKE'06, Eighteenth International Conference on Software Engineering & Knowledge Engineering.* California, USA.

Tannenbaum, A., 2002. *Metadata solutions: using metamodels, repositories, XML, and enterprise portals to generate information on demand*, Addison Wesley, New York, USA.

Valentin, L. G., 2006. Uma arquitetura para um sistema de descoberta de conhecimento**.** Programa de Pós-Graduação em Ciência da Computação. *Dissertação de Mestrado*. Departamento de Informática. Universidade Estadual de Maringá.

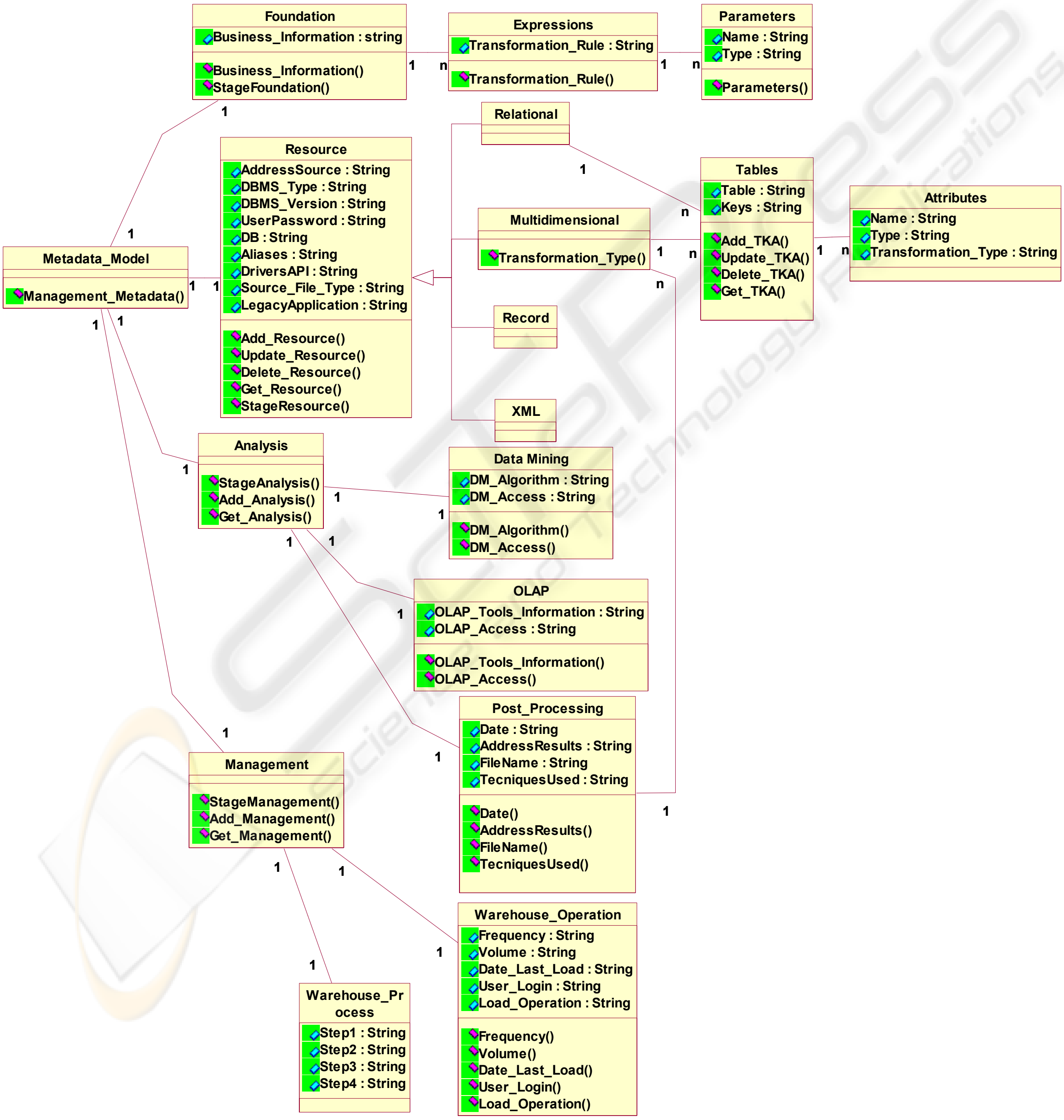

Figure 1: Diagram of Classes of the Manager of Metadata.