SEMANTICALLY ENHANCING MULTIMEDIA DATA WAREHOUSES

Using Ontologies as Part of the Metadata

Andrei Vanea and Rodica Potolea
Technical University of Cluj-Napoca, Computer Science Department, 28 Gh. Baritiu Street, Cluj-Napoca, Romania

Keywords: Data warehouse, Semantic, Metadata, Ontology, Business intelligence.

Abstract: Data warehouses are versatile systems capable of storing and processing large quantities of data. They are most suited for aggregating and reporting. The data managed by these systems vary from simple, numeric data, to more complex, multimedia data. One of the domains in which multimedia data is intensively produced is medicine. We present a method for semantically enhancing the metadata stored in a medical multimedia data warehouse. This semantically rich environment will gain in autonomy, reducing the dependence on human intervention to resolve new, unforeseen queries. Furthermore, the use of the semantic relations defined in the ontology allows the system to speed up the execution of a query, by computing the results of new, unforeseen queries, from the fact data already stored in the data warehouse.

1 INTRODUCTION

Large quantities of data are being produced daily in every domain, creating storage problems for business and organizations. A related issue is the time needed to process the data stored.

Data warehouses are systems mostly used to structure, store and process historical data. The data is usually numerical (Vassiliadis, 1998), symbolic (textual) (Diday, 2003) and recently multimedia (Mbarki, 2004). A particularity of the last decade is the significant increase of quantity and quality of multimedia data. As a result, specific issues related to storing, processing and knowledge extraction from such data have arisen. Data warehouses provide preprocessed and aggregated data. They also store a large amount of metadata (Object Management Group, 2003), which is used to offer information on the system and data.

In this article we present a method for semantically enhancing the metadata repository, by using an ontology based technique. We consider that adding semantics to the data warehouse provides a higher level of independence to the system. Thus it will be able to automatically solve tasks that are usually the responsibility of the data warehouse administrator. Such tasks mainly refer to the user queries, which are new to the system (i.e. they have not been considered at the design time).

The rest of the article is organized as follows. In Section 2 we briefly present similar intelligent systems, which deal with data management and information extraction. Section 3 describes the general system structure, as well as our particular implementation and the data handled by it. Section 4 presents the way in which we used the semantic metadata. In Section 5 we present experimental results and, finally, Section 6 presents a summary of our work and some conclusions.

2 RELATED WORK

Many researchers are studying data warehouses. Although the technology of data warehouses which store and process numerical data is considered to be mature, there is always room for improvement.

In recent years, the interest for data warehouses which focus on data types other then the classical numeric type has increased. In (Mahboubi, 2009) a data warehouse model for complex data (text files images, data bases, sounds) is presented. The semi-structured format of these objects is captured via XML files, which are parsed and validated against a requirements pattern. A medical multimedia data

Vanea A. and Potolea R..
SEMANTICALLY ENHANCING MULTIMEDIA DATA WAREHOUSES - Using Ontologies as Part of the Metadata.
DOI: 10.5220/0003434701630168
In *Proceedings of the 13th International Conference on Enterprise Information Systems* (ICEIS-2011), pages 163-168
ISBN: 978-989-8425-53-9

warehouse containing image data and symbolic data and focused on ECG signal recordings is detailed in (Arigon, 2007).

Semantically enhanced data warehouses have been also taken into consideration. In (Pardillo, 2008), the authors propose a method to semantically translate conceptual models into their platform specific counter parts, by using an OLAP algebra. In (Xie, 2007) a data warehouse which has two ontologies has been built; one for the specific business terms and one for the technical terms, specific to the aggregation and knowledge extraction tools. This requires a one-time collaboration between the business experts and data warehouse designers, to produce a mapping between the two ontologies. As a result, whenever a new query is requested by the business analysts, the warehouse administrator can quickly create the appropriate data mart, without the need of long and repetitive meetings between the two expert teams. In (Nebot, 2010) the authors propose an ETL tool for extracting semantically annotated data into fact tables.

Medical ontologies have been intensively researched and developed in the later years. This resulted in a large number of ontologies dealing with different sub domains (Freitas, 2009) and/or serving specific purposes (Rubin, 2007). A data warehousing system which uses semantic data to aid in the diagnosis process of mitral valve prolapse is described in (Podgorelec, 2009).

3 THE STRUCTURE OF THE MULTIMEDIA WAREHOUSE

In today's medicine the physicians benefit from heterogeneous data sources, which provide different types of data. Images and sounds are intensively used in order to better asses the health state of a patient. The data is produced in large quantities, and at very fast rates. The abundance of data increases the probability that quality information can be produced. This is possible only if new, specialized data management methods are designed and developed. Powerful tools are required mainly because humans are not able to efficiently process large amounts of data. Therefore, intelligent systems that manage such large quantities of data are helpful in assisting physicians in the decision process.

Based on this premise, we designed and built a medical multimedia data warehouse. The purpose of this system is to manage both classical and multimedia (medical) data in such a manner that meaningful medical information can be efficiently produced (fast and easy).

In order to aid them in the diagnosis decision process, the medical physicians need a particular piece of information at the moment they are examining the patient. Therefore, an issue which we must address, due to the medical nature of the system, is the speed of query processing.

3.1 General Description

We propose a five section structure (Figure 1): ETL tools section, Warehouse, Semantic metadata, Processing and Metadata Maintenance and Query processor.

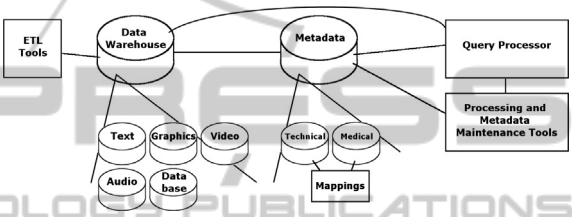

Figure 1: The structure of a multimedia data warehouse.

The ETL Tools Section handles the acquisition, cleaning and transforming of raw data. The Warehouse Section stores the data handled by the system: custom format data, raw data and aggregated data. The Semantic Metadata Section is a special repository which stores both classical metadata and semantically enhanced metadata. The later metadata type is used to provide machine understandable information on the data being stored and the domain in which the system is being deployed. The Processing and Metadata Maintenance Section provides tools for data processing and manipulation and tools for CRUD operations on the metadata. The Query Processor Section provides tools for understanding the users query and tools for computing the query result.

3.2 Implementation

A multimedia data warehouse needs to offer support for heterogeneous types of data: numeric and symbolic data, text files, image and sounds. Multimedia data is often semi-structured or non-structured. Therefore, it is difficult to extract structured data from it. To overcome this problem we decided to use XML to store the unstructured or semi-structured data extracted from the multimedia files. XML is useful in this situation because of its ability to model semi-structured data.

Using the proposed model, we built a Medical Multimedia Data Warehouse, which we deployed in the field of pneumology. The system deals at this time with three data types: symbolic (textual), numeric and graphic (images). Symbolic and numeric data are used to store information on the patient, such as identification data and physical and/or medical features. Images are used to store information on spirometry tests performed by the patients. At this time, our system deals only with the exhaling part of the tests.

XML files containing information on the patient and on the spirometry tests, as well as images containing spirometry test results are loaded by the ETL tools. These data are transformed and loaded into custom internal XML files. These internal files store information on each patient, their tests and results, such as physical features at the moment of each test. The internal files are stored in a staging area, in the Data Warehouse Section.

The images present graph data which model the flow air being exhaled by the patient. Different medical relevant features are extracted from the images: Peak Expiratory Flow (PEF), Normal Peak Expiratory Flow (NPEF), Forced Expiratory Flow at 25%, 50%, 75% and 25-75% (FEF25, FEF50, FEF75 and FEF25-75 respectively), Normal Forced Expiratory Flow at 25%, 50%, 75% and 25-75% (NFEF25, NFEF50, NFEF75 and NFEF25-75 respectively), Forced Vital Capacity (FVC), Normal Forced Vital Capacity (NFVC), Flow-Volume Line (FV line), Forced Expiratory Volume in one second (FEV1), Normal Forced Expiratory Volume in one second (NFEV1). The *normal* indicators represent values obtained by healthy patients, with the same physical features.

The internal files are processed and the resulted information is stored in fact tables. Facts are computed in two ways. First, the classical numeric form, in which particular data is aggregated. The second manner is by creating images similar to the ones stored in the warehouse, but with aggregated multimedia features.

4 THE SEMANTIC METADATA MODEL

In particular situations, dimensional hierarchies cannot be used in order to provide the level of granularity desired by the medical specialist. Such a situation occurs when sets of low level features which determine high level features are disjoint.

As an example, consider the *diagnosis* dimension. Each diagnosis depends on a specific subset of medical test set. Unfortunately, the subsets differ from diagnosis to diagnosis. In our case, the *restriction* diagnosis depends on the following tests: fvc, fev1. The *small air ways obstruction* diagnosis however, is set based on the following tests: fvc, fef25, fef50, fef75, fef2575. These two subsets are not disjoint, as they have some tests in common. Therefore, drill-down, from diagnosis to tests, is not easily achievable using the classical hierarchies.

We propose an ontology driven metadata in order to overcome this difficulty. By providing the system with semantically enhanced metadata, it will gain in autonomy.

4.1 Query Analysis

We started by formalizing medical queries to identify concepts which are of interest for the domain expert and describe them in technical terms in the metadata (i.e. the ontology). The relevance of the step is provided by the semantic metadata which aids the querying process via matching domain expert and technical terms. By analyzing use cases, we identified the relevant domain cases, and extracted their general template, such as:

```
select the [minimum | mean | maximum
| most frequent] value of feature F
for the [males | females | patients]
with [feature_1 = value_1 {and feature_n
= value_n}]
```

A special case of query deals with the “age” feature. This is because the age affects both the diagnostic and statistical assessment. Therefore, the age appears twice in the query.

4.2 The Ontology

The ontology provides structured information about the real world model. Moreover, it ensures a mapping between the specific domain (medical in our case) and technical elements used by the system. Its main roles are: to model the specific domain in which the system is deployed, to provide mappings from that domain to the technical data and to create a more flexible environment for query submission.

The first role is to define relations and provide support for all the entities identified in the previous section. The first entity (i.e. class) we need to address is the “patient”. All the other entities are in some specific way linked to the patient therefore, it is arguably the most important entity in the entire

ontology. The full list of entities, together with their instances is presented bellow:

- medical_test: spirometry_test;
- physical_feature: date_of_birth, gender, race, height, weight.
- medical_result: pef, fvc, fev1, fef25, fef50, fef75, fef2575;
- graph: flow_volume, volume_time
- diagnosis: normal, restriction, obstruction, small_air_ways_obstruction, mixed;

The current version of our ontology contains two types of inter-class relations. These relations are: "has_a", which depicts the fact that an instance of a class is the owner or beneficiary of an instance of some other class; "influenced_by", which depicts the fact that an instance of a class is defined by a complete set of instances of other classes.

Using these two relations, we define the following structure:

```
patient has_a physical_feature
patient has_a medical_test
medical_test has_a medical_result
medical_test has_a graph
medical_test has_a diagnosis
diagnosis influenced_by medical_result
```

Another feature which is present in the ontology is the data type of the concept (numerical, textual or Boolean). We have also added a property which defines a numerical concept as a set. This allows the system to treat numerical concepts as sets, using set theory to compute query results. A concept can be viewed as a "continuous_set", "discrete_set" or "not_a_set". For example, the age is defined as "discrete_set". We present in Section 5 a case in which this property is proven to be useful.

The second role of the ontology is to provide a link to the technical terms, reducing the dependency on the Data Warehouse Administrator, when new, unforeseen queries are formulated. To do so, we designed a special relation in the ontology. The relation is called "mapping" and maps one domain specific entity, in this case medicine, to a technical domain entity. This technical entity is relevant to the data warehouse designer or developer administrator, rather than to the warehouse user. By providing this technical mapping, the system can automatically resolve some unforeseen task, which could by time consuming if the data warehouse administrator would have had to solve. A more detailed description on how the mapping works is given in Section 4.3.

The third role is to add greater flexibility to the method in which the query is formulated by the medical specialists. Very often, in medicine, physicians encounter multiple names or specific multiple terms, which refer to the same medical entity or concept. In order to make the medical physician feel more comfortable when working with the medical data warehouse, we decided to use synonymy. By allowing synonyms into the ontology, the system could collaborate with other systems which work on similar ontologies. This enables information exchange between two heterogeneous (medical) systems. More precisely, our medical system can exchange data or even information, with other medical systems, deployed in other hospitals, in order to retrieve meaningful information to aid the physician in the decision making process.

Concepts and terms which are defined as synonyms can be collected from both the physicians and other medical ontologies, such as Foundational Model of Anatomy or Disease Ontology. Medical ontologies are often developed by different organizations, model distinct medical sub-domains and do not follow a standardized structure or a name convention. Therefore, defining synonyms does not limit the expressiveness of physicians to a particular implementation of an ontology.

4.3 Term Mapping

By term mapping we refer to the process which links two terms, or concepts, from different domains. In the most general form, this process translates a business concept into a technical concept, which could be in some way handled and processed by the system. In the case of our specific implementation of the system, the mapping is done for the medical specific terms and the technical terms, which are of relevance from the informational system's point of view. Term mapping provides a direct correlation between the real world model and the technical one.

The ontology provides semantic annotations about a particular domain. These annotations are computer readable. The purpose of the term mapping is to extend these semantic annotations to the technical terms. By doing this, the system can automatically determine which technical concepts handled by it should be computed or accessed in order to provide a meaningful and correct answer to the query inputted by the user. In our particular case, a medical physician will input a query using domain specific terms and the system will infer the class of the terms, existing relations among them and also the correspondence to the concepts used by the system in order to model the real world scenarios.

We implemented the mapping for our system by using the internal format in which we store the data. Each XML file corresponds to a patient. Therefore, the concept of "patient" described in the ontology is mapped to the root element of the XML file. When a new type of query containing a reference to one or several patients will be submitted, the system will infer that it has to access the root elements of the XML files that are in the staging area of the data warehouse. All the mappings are expressed in a similar way, using XML tags from the staging area files. Ideally, every medical (domain) concept present in the ontology should have a corresponding XML tag, i.e. a mapping with a technical term. Although our particular implementation is done using XML tags (i.e. the internal data format) the mappings can also be formulated between the concepts in the ontology and attributes from any type of database.

Multimedia concepts are dealt with in a similar way. Atomic multimedia concepts, such as lines, curves or any pixel sets bound to each other by a common feature, can be mapped either to a field in the internal format, or to a tool which directly extracts these concepts from the multimedia file.

When a query is submitted, the system will try to identify the concepts contained by the query and present in the ontology. If this step is successfully done, it will identify the technical concepts mapped to the domain concepts found in the query. Next it will search for a fact table which contains the information needed to answer the query; that is, a table which stores aggregated data derived of the technical terms. If a corresponding fact table exists, the system will return the aggregated data which answers the query; if not, the system will check the synonyms of the concepts present in the query. Based on every possible combination of concepts and their synonyms, corresponding fact tables are searched for and the aggregated data stored in the tables are checked to see if the answer the query. If the system still does not find a suitable fact table from which it can produce the result, it tries to infer the domain relations. This is done by determining based on the "influenced_by" relation, the concepts which are influencing the ones specified in the query. After determining these concepts, the whole process of finding a corresponding fact table is repeated until a valid fact table is found, or until there are no more fact tables to check or influencing (lower level) concepts.

5 EXPERIMENTAL RESULTS

All the concepts and techniques presented in the previous sections are implemented in an experimental system. We present some of our results, to better illustrate the functionality of the proposed method. The first example illustrates the method in which the system parses a query and then computes the result based on the semantic metadata available and data already stored in fact tables.

The query in the first example is the following:

```
select the number of patients in the
age group of 10 to 25 and with the
diagnosis restriction
```

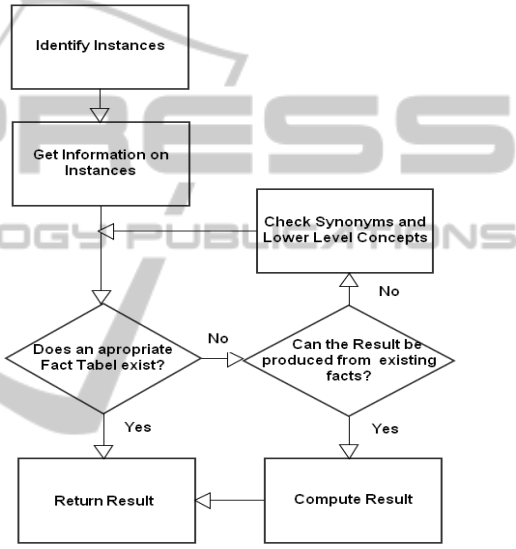

Figure 2: Query resolving steps.

Figure 2 presents the flow for resolving the query. First, the system identifies the instances in the query, which are specified in the ontology, i.e. "age" and "restriction". Second, the classes and the corresponding data types are identified: "age" is a numerical "physical_feature" and "restriction" is a string "diagnosis". Next, the system checks if an existing fact table stores aggregated data referencing the two classes. In our particular case the system found such a fact table, though a valid fact to answer the query was inexistent. By examining the stored facts which reference the "restriction" diagnosis, the system identified that there where two facts which could be used in order to respond to the query. The two facts stored the number of patients for the age group of 10 to 19 and 20 to 25 respectively. "Age" is a "discrete_set", therefore the union of 10-19 and 20-25 results in the age group of 10-25. The response was computed by adding the two measures.

The second query example shows how a query

on graphical data is resolved:

```
select the mean flow-volume curve
for patients with the age of  23 and
the diagnosis obstruction
```

As with the first query, the system identified the instances, classes and types present in the second query. The "flow-volume curve" was identified as a multimedia feature ("graph"). Based on the "graph" mapping available in the ontology, the system retrieved all the graph data for patients of age 23 who have been diagnosed with obstruction. The graphs where scaled to a common dimension and the mean graph was computed.

6 CONCLUSIONS

Data warehouse technology can be of great value in many domains. One such domain is medicine. This particular field also benefits from semantic web technologies, as numerous academic and industry researchers have developed different ontologies.

In this article we presented a general multimedia data warehouse model, with a semantically enhanced metadata repository. The enhancement was achieved by developing an ontology which models part of a sub-domain in medicine (pneumology). Based on semantic annotations which map the specific terms to the technical terms, the system becomes more dynamic and gains autonomy, as it no longer needs administrator's intervention.

The advantages of our method are twofold. First, it is adapted to work with multimedia files and data, by breaking larger multimedia "objects" into sets of smaller ones. Second, the majority of unforeseen queries can be resolved by the system, if the ontology is properly built. This is achieved by using proper semantic annotations like the type of a concept or instance, synonyms and "influenced_by" relations. The use of synonyms allows the system to communicate with heterogeneous medical systems, making information sharing a relatively simple task.

For future work we plan to extend the mappings to the multimedia data extraction tools, defining a set of semantic annotations that allow the system to work with a larger number of multimedia "objects". We also plan to improve the methods with which the system computes the results of a query from existing fact data.

ACKNOWLEDGEMENTS

The work is supported by the project "Doctoral studies in engineering sciences for the development of knowledge based society - SIDOC" contract no. POSDRU/88/1.5/S/60078, project co-funded by the European Social Fund through the Regional Operational Human Resources Program 2007-2013.

REFERENCES

Vassiliadis, P., 1998. Modeling Multidimensional Databases, Cubes and Cube Operations. In *Proceedings of the 10th SSDBM Conference* IEEE Computer Society.

Diday, E., Esposito, F., 2003. An introduction to symbolic data analysis and the SODAS software. In *Journal Intelligent Data Analysis*, Volume 7, December 2003, IOS Press Amsterdam, The Netherlands.

Mbarki, M., Dupuy, C. S., 2004. A Conceptual Modeling of Multimedia Documents. In *Proceedings of IADIS International Conference WWW/Internet 2004* IEEE Computer Society Press.

Object Management Group, 2003. Common Warehouse Metamodel (CWM) Specification.

Mahboubi, H., Ralaivao, J. C., Loudcher, S., Boussaid, O., Bentayeb, F., Darmont, J., 2009. X-WACoDa: An XML-based approach for Warehousing and Analyzing Complex Data. In *Advances in Data Warehousing and Mining*, IGI Publishing.

Arigon, A. M., Miquel, M., Tchounikine, A., 2007. Multimedia data warehouses: a multiversion model and a medical application. In *Multimedia Tools and Applications*, Springer.

Pardillo, J., Mazón, J. N., Trujillo, J., 2008. Bridging the Semantic Gap in OLAP Model: Platform-independent Queries. In *Proceedings ACM 11th International Workshop on Data Warehousing and OLAP*, ACM.

Xie, G., Yang, Y., Liu, S., Qiu, Z., Pan, Y., Zhou, X., 2007. EIAW: Towards a Business-friendly Data Warehouse Using Semantic Web Technologies. In *Proceedings of The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference*, Springer-Verlag.

Freitas, F., Schulz, S., Moraes, E., 2009. Survey of current Terminologies and Ontologies in Biology and Medicine. In *Electronic Journal of Communication, Information & Innovation in Health*, Institute of Communication and Scientific and Technological Information in Health.

Rubin, D. L., Shah, N. H., Noy, N. F., 2007. Biomedical ontologies: a functional perspective. In *Briefings in Bioinformatics. Volume 9, 75-90*, Oxford University Press.

Nebot, V., Berlanga, R., 2010. Building Data Warehouses with Semantic Data. *In Proceedings of the 2010 EDBT/ICDT Workshops*, ACM.

Podgorelec, V., Grasic, B., Pavlic, L., 2009. Medical diagnostic process optimization through the semantic integration of data resources. In *Computer Methods and Programs in Biomedicine Volume 95, Issue 2, Supplement 1, August 2009, Pages S55-S67*, Elsevier.