APPLYING ONTOLOGIES IN THE KNOWLEDGE DISCOVERY IN GEOGRAPHIC DATABASES

Guillermo Nudelman Hess, Cirano Iochpe
Universidade Federal do Rio Grande do Sul, Caixa Postal 15064, Porto Alegre, RS, Brasil
Instituto de Informática

Keywords: Geographic Database, Ontologies, Semantic integration, Conceptual modeling

Abstract: This article proposes a software architecture to integrate geographic databases conceptual models. The goal is the preprocessing phase on the knowledge discovery in database, using geographic databases conceptual schemas as input data, in order to obtain analysis patterns candidates. The semantic unification is very important in this process, since the data mining tools are not capable to recognize synonyms neither to distinguish between homonymous. In this way a methodology to refer the knowledge basis was developed.

1 INTRODUCTION

Because of the increasing use of the Geographic Information Systems (GIS) in the last past years, the conceptual modeling of the Geographic Database (GDB) has become a very important task. However, each one of the GIS software has its own data model, which has it focus in the logical phase of the database project (Silva, 2003).

Plenty of conceptual models to the GDB project have been proposed, attending to make the GDB modeling independent from the implementation platform. Among them, some are the UML-GeoFrame (Rocha, 2001) and MADS (Parent, 1999). The core of most of them is equivalent, and a complete comparative study is presented in (Bassalo, 2002).

The use of the conceptual modeling allows also the project documentation and the reuse of the model, or part of it, several times. This reuse is specially interesting in GDB, since its modeling is quite complex, and part of the geographic concepts of the real world being modeled is repeated for distinct applications. In this way the use of analysis patterns (Gamma, 1995) is usefeull. Analysis patterns are the essence of the conceptual modeling for the solution of a recurrent problem in a specific context.

To support the acknowledgment of analysis patterns automatically, the Knowledge Discovery in databases (KDD) (Fayyad, 1996) can be applied. This process has several steps, as shown in Figure 1.

The data mining (DM) and post-processing of conceptual schemas was performed in (Silva, 2003), by the use of DM tools that produce associative rules. However, a few conceptual schemas could be

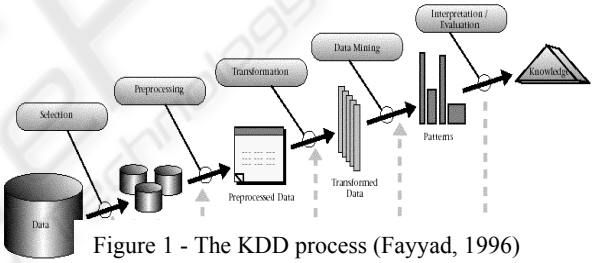

Figure 1 - The KDD process (Fayyad, 1996)

mined, because the pre-processing phase was not addressed. The present work is focused in this phase.

To mine a number of conceptual schemas in possibly different data models it is necessary put them all in the same format, because they have to be stored to be reused after. Unfortunally, this does not occurs with the GDB conceptual models, because there is not a modeling pattern addopted as a standard.

To reach a correct data preparation, this schemas integration accomplishes two levels, the syntatic and the semantic. The first one relates to the equivalence of the constructors of each model. A research over the unification of the refered models was initiated in (Bassalo, 2002), where is presented the constructors union set.

The semantic level of the integration comprises the problem of unification between names used to describe the real world phenomena being modeled and the associations among them. In this sense it is necessary to build a Knowledge Organization System (KOS) (Hodge, 2000), such as a controlled vocabulary, a taxonomy, a thesaurus (Qin, 2001), a

Nudelman Hess G. and Iochpe C. (2004).
APPLYING ONTOLOGIES IN THE KNOWLEDGE DISCOVERY IN GEOGRAPHIC DATABASES.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 509-512
DOI: 10.5220/0002634505090512

semantic network or an ontology (Guarino, 1998) to store the concepts concerning of the geographic applications domain.

Section 2 of this article presents the context of the GDB semantic integration problem. A software architecture for the integration of GDB conceptual schemas is presented in Section 3. Section 4 details the proposed metodology to query ontology in the designed software architecture. At last, conclusions and future works are presented in Section 5.

2 WHY SEMANTIC UNIFICATION OF GDB SCHEMAS

The semantic integration, even in databases or conceptual models, is a very complex and costly task, once it has to address variety kinds of heterogeneity.

Bergamaschi et. al (Bergamaschi, 1998) classifies the heterogeneity in terms of nomenclature and structure. The first case englobes the naming conflicts, such as synonyms and homonymous. Structural heterogeneity concernes the differences existing in the conceptual model, in terms of attributes and associations of the modeled concepts.

Geographic databases try to model the real world phenomena. Thus the set of elements to be modeled are concrete and quite restricted. The attributes and associations between the geographic elements are always the same. The only thing that varies is the approach, which depends on the application and the designer's knowledge, and also the names used to represent the same things. In this sense, the development of a set of definitions about names, attributes and associations of the geographic phenomena is usefull, in at least two aspects, described in the next subsections.

2.1 Integration of Geographic Applications

To make the integration of geographic application possible, three requisites must be satisfied (Bergamaschi, 1998):

1. The conceptual schemas of each source must be available;
2. There must be semantic information in the schema;
3. A canonical data model must exists. This standard model must have enough expressivity power to describe all the models to be integrated;

Once the target of the integration proposed in this paper is of conceptual schemas, the first requisite is autmatically satisfied. The other requisites are satified by the use of the work developed in (Bassalo, 2002) and by the use of GML (OpenGIS, 2001).

Through the use of a KOS to eliminate semantic heterogeneity not only the data mining is possible, but also at least other three capabilities can be reached (Sheth, 2000):

4. Terminological transparency: Ambiguities created by homonymous and synonyms are eliminated;
5. Context sensitive processing: Depending on the context (attributes and associations) in which a concept is in, it is possible to infer its meaning;
6. Semantic correlation: Integration between conceptual schemas, combining both aspects above.

2.2 New applications modeling aid

The database conceptual modeling process is a complex task, but really important to guarantee the correct working and the manutenability of the database. In order to automate this process and aid the designer, a number of CASE tools are disposable. However, this is not true for the GDB conceptual modeling. There are some academic proposals, but specific for one data model, such as RoseGIS (Hess, 2003) and MADS editor (Parent, 1999).

None of these CASE tools has information about the real world and how is its behavior. The consequence is that the designer is who has to give all sort of information about the application's domain. Thus, the build of a KOS containing the elements (phenomena, in case of GDB) and the associations of the domain may contribute to the database project (Sugumaran, 2002), and thus to the GDB project. The designer can face his modeling against the existing KOS, so he can detect possible inconsistencies and incompleteness, such as missing entities, attributes and associations.

3 THE ARCHITECTURE

To reach correct data preparation of conceptual schemas based on different data models, it is necessary to develop a mechanism to unify those models. This integration aims to eliminate possibles ambiguities of understanding and data redundancy.

As the ontology is the KOS chosen technique, Figure 2 presents a generic architecture to translate conceptual schemas, independent of the data model.

A conceptual schema is primarily converted into a syntactic canonical file format (SCFF), that is, only in the syntactic level. According to the data

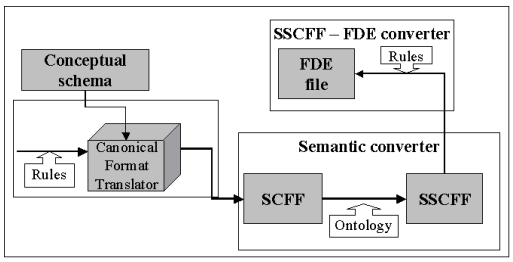

Figure 2 - Conceptual schema integrator architecture

model in which the schema is based a specific set of rules is applied.

This syntactic integration turns conceptual schemas into a canonical data model, totally independent of platform. The Geographic Markup Language (GML) (OpenGIS, 2001) encoding is the chosen format to be used as the canonical data model.

Even knowing the GML is not capable to represent all of the constructors from all the data models it was adopted for having a significative set of elements used in the GDB modeling and because it is standard data format proposed by the OpenGIS. Moreover, in the future GML can be extended to handle the missing constructors.

The second step of the process consists in pass the SCFF through an ontology, to guarantee the semantic level of the data preprocessing. The result is a semantic and syntactic canonical file format (SSCFF). The last step of the data preprocessing consists in trasform the SSCFF file to the FDE file (Silva, 2003) which can be handled by the data mining tools.

4 THE ONTOLOGY'S ROLE

Ontologies are used, in this work, to conceptualize. The use of an ontology by itself does not provide a complete solution to the semantic integration problem. It is impossible to the ontology to contemplate all the ways to express a real world phenomenon. Depending on the geographical location of the designer the names used to the same concept may vary. Moreover, the spelling of the same concept may vary too, specially in case of abreviation.

To solve the situations cited above the ontology process may use some similarity matching (Cohen, 1998) techniques. This matching has to occurs in the level of names and in the level of the structure of a term, considering hierarchies, associations and attributes of the candidate concepts stored in the ontology (Bergamaschi, 1998).

4.1 The algorithm to search and update the ontolgy

Figure 3 illustrates the algorithm to search the ontology.

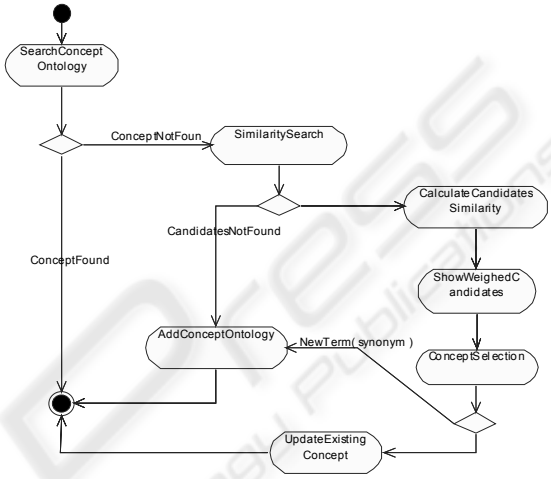

Figure 3 - The ontology searching process

Step 1 – Search a concept in the ontology: Each concept in the conceptual schema is searched in the ontology. If the term is found with the same name and exactly the same attributes and associations, the query ends and starts for the next concepts. If the name, attributes or associations are different from the ones stored in the ontology, go to step 2.

Step 2 – Similarity search: Applying techniques of similarity matching cadidates to synonyms of the input concept are identified. The similarity coefficient is calculated, based on criteria of name and structure similarity. If candidates are found, go to step 3. If there are no candidates, go to step 4.

Step 3 – Terms selection: Each one of the terms identified as possible synonyms to the input concept is presented to the domain expert, who identifies the most appropriate. If it is an insertion of a new synonym of a concept already stored, without the need to update its structure (attributes and associations), go to step 4. If it is necessary to update the concept structure, go to step 5.

Step 4 – Insertion of a concept in the ontology: The term is added to the ontology. If it comes from step 3, it is associated to its equivalent in the ontology. If it comes from step 2, it is added to the ontology with all attributes and associations. The algorithm searches for the next term.

Step 5 – Update of an existing concept: The structure (attributes and associations) of an existing concept is updated in the ontology.

5 CONCLUSIONS

The use of analysis patterns can contribute to the improvement of GDB conceptual models because they are tested and aproved solutions. This can reduce the time needed to the conceptual project and also reduce the possibility of making mistakes. The obtainment of these patterns can be done by the KDD process application. One of the important phases of this process is the data preprocessing.

Specifically in GDB conceptual schemas the data preprocessing consists in the integration of the conceptual schemas designed based on different data models and with naming variations to the same real world concepts. Thus the integration must be performed in two levels, syntacticaly and semanticaly which was the focus of this paper. The semantic integration among distinct conceptual schemas must be aided by an ontology, which allows searching by names and also searching by structure as attributes and associations.

Another benefit of using ontologies, is the fact the knowledge is stored and can be updated and interchanged. Not only analysis patterns can be deduced but also the ontologies existing concepts can help the designer in modeling a new conceptual schema. However to explore all the ontologies potentialities and in an efficient way it is necessary to combine it with another technique very used in heterogeneous databases, known as similarity matching.

The next steps of this research are the study of the similarity matching techniques and more important the definition of a set of criteria to be considered in the similarity coefficient calculus and also the weight of each one. The implementation of the algorithm proposed in section 4 is also a future work to test the efficiency of this solution to the semantic unification.

REFERENCES

Bergamaschi, S. et al., 1998. An Intelligent Approach to Information Integration. *In Internation Conference on Formal Ontology in Information Systems (FOIS'98)*. Italy.

Bassalo, G.H.M.; Iochpe, C.; Bigolin, N., 2002. Representando esquemas no Formato Atributo-Valor para a Inferência de Padrões de Análise. *In: IV Brazilian Symposium on GeoInformatics - GeoInfo 2002*. Caxambu, Brazil.

Cohen, W.W., 1998. Integration of Heterogeneous Databases Without Common Domains Using Queries Based on Text Similarity, *In Proceedings of the 1998 ACM SIGMOD international conference on Management of data*. USA.

Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P, 1996. From Data Mining to Knowledge Discovery in Databases. AI Magazine, v.17, n.3, p.37-54.

Gamma, H.E.; Johnson, R.; Vlissides J., 1995. Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley.

Guarino, N., 1998. Formal Ontology and Information Systems. In Proc. of In Internation Conference on Formal Ontology in Information Systems (FOIS'98). Italy.

Hess, G.N.; Iochpe, C.; Silva, C.M.S., 2003. RoseGIS: Uma ferramenta CASE para projeto de banco de dados geográficos. *In GISBrasil 2003*. Brazil.

Hodge, G., 2000. Knowledge Organization Systems: An Overview. In System of knowledge Organization for Digital Libraries: Beyond Traditional authority files.

OpenGIS Consortium, 2001. Geography markup Language (GML) 2.0. Open GIS Implementation Specification. Available in http://www.opengis.net.

Parent, C. et al., 1999. Spatio-temporal conceptual models: data structures + space + time. In *Proc.7th ACM GIS*, Kansas City, USA.

Qin, J.; Paling, S., 2001. Converting a controlled vocabulary into an ontology: the case of GEM, *Information Research 6,* 2001.

Rocha, L. V.; Edelweiss, N.; Iochpe, C., 2001 GeoFrame-T: A Temporal Conceptual Framework for Data Modeling. *In: ACM Symposium on Advances in GIS*. Atlanta, USA.

Sheth, A.P., 2000. Changing focus on interoperability in information systems: From systems, syntax, structure to semantics. *In Interoperating Geographic Information Systems*".

Silva, C.M.S.; Iochpe, C.; Engel, P.M., 2003. Using Knowledge Discovery in Database to Identify Analysis Patterns, *5th International Conference on Enterprise Information System*, Angers, France.

Sugumaran, V.; Storey, V., 2002. Ontologies for Conceptual Modeling: their creation, use and management. *In Data & Knowledge Engineering.* Elsevier.