Analysis Patterns for the Geographic Database Conceptual Schema: An Ontology Aided Approach

Guillermo Nudelman Hess, Cirano Iochpe, José Palazzo Moreira de Oliveira

Universidade Federal do Rio Grande do Sul, Instituto de Informática
Porto Alegre, Brazil
{hess,ciochpe,palazzo}@inf.ufrgs.br

Extended Abstract. Conceptual modeling of Geographic Databases (GDB) has become a very important task due to both the increasing exchange and reuse of geographic information. Most of the GDB support a proprietary data model, specific for a particular system architecture. This design scenario leads to a non-reusable as well as implementation-dependent geographic database project. The reuse is especially interesting in GDB, since the precise modeling process can be of significant complexity. Furthermore, a wide set of real world entities must be represented in the database of most geographic applications. To support this kind of reuse, the analysis patterns methodology [2] is a very useful technique. Analysis patterns are the core of the conceptual modeling for the solution of a recurrent problem in a specific context. Having the patterns stored in a knowledge base that can be easily searched and updated, a designer is able to search, retrieve and reuse the most adequate concept representation that applies to a specific modeling process. To support this approach, a ontology can be used to play the role of the knowledge base mentioned above.

To support the recognition of analysis patterns automatically, the Knowledge Discovery in Databases (KDD) [1] is a candidate technology. This process can be briefly outlined as a sequence of steps or phases that can be revisited when necessary, an iterative process. In the preprocessing step, one of the main tasks is integration of all selected data. To reach a correct data preparation for mining, this schemata integration process must deal with syntactic and semantic heterogeneities. It must solve the unification problem of different representations used to describe the same real world phenomenon, and the relationships between them. The present work deals with this step. The architecture consists in a syntactic part, which intends to convert the different data models to the Geographic Markup Language (GML), which is used in our methodology as the canonical data model, since it is a standard for storing and exchanging geographical data.

To support the semantic phase of the pre-processing, we propose the use of an ontology to be used as the knowledge repository. In this sense, we developed semi-automate algorithm to search and update the ontology, illustrated in figure 1. In [3] there is a complete description of the algorithm.

To minimize the need of the expert intervention two parameters have to be set at the beginning of the algorithm execution: the minimum and the maximum accepted probabilities. Only the concepts having similarity probability higher than the minimum specified are shown to the expert. If one ore more of the ontology candidates have similarity probability higher than the maximum specified by the user, the one with the higher value is considered as equivalent (a synonym) of the input concept.

To perform the algorithm it is necessary to measure the similarity between two concepts, one from the input conceptual schema and the other from the ontology. Here we adopt a hybrid approach, combining syntactic matching between strings and semantic matching. In [3] the mathematical part of the similarity measuremente is explained. It takes into account the name, attributes and relationships features of the concept
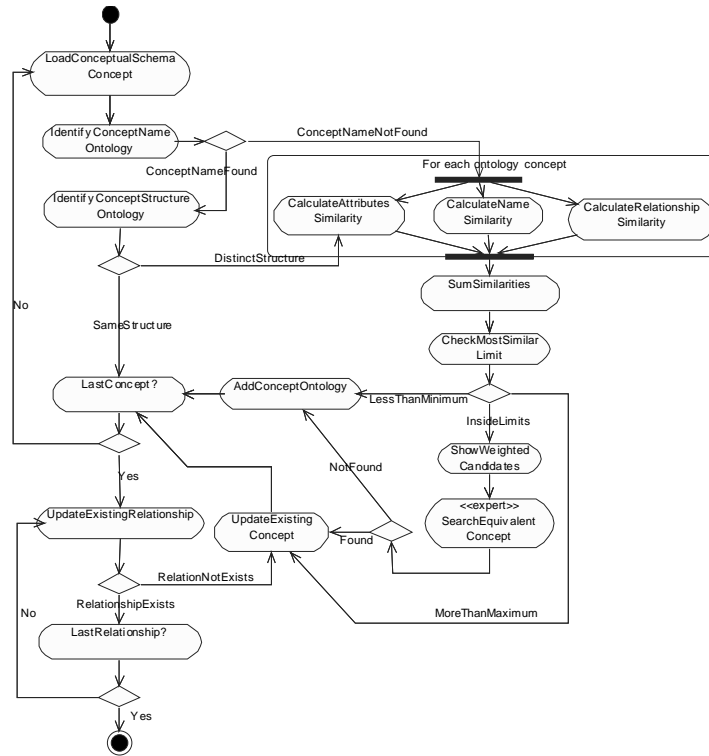
Fig. 1: The ontology's search algorithm

During this process, the original conceptual schema is kept unaltered and a new one is generated, in the canonical semantic format, and this is the one to be used later in the data mining. The ontology can also be dinamically updated depending on the concept's matches. Attributes and relationships can be added, and even new concepts may be inserted.

References

1. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery in Databases. In AI Magazine, v.17, n.3,p.37-54, 1996.
2. Fowler, M. Analysis patterns : reusable object models. Menlo Park: Addison-Wesley, 1997. 357 p. : il.
3. Hess, G.; Iochpe, C. Applying Ontologies in the KDD Pre-Processing Phase. In Proc. of 16th International Conference on Software Engineering and Knowledge Engineering (SEKE'04). Banff, Canada. June 2004 (to be published).