Chapter 7

CONCLUSIONS

The continuous interaction among people and their natural home, the Planet Earth, generates everyday new requirements associated to spatial data. For example, urban analysis, natural risks prevention, space exploration, contamination in oceans, and reforestation of lands just to mention some of them. Spatial data mining involves the integration of methods from different scientific fields which help us, by means of data analysis and discovery algorithms, to produce a particular enumeration of patterns from spatial data.

In Chapter 2 we presented several approaches developed for mining spatial data (i.e., generalization-based methods, clustering, spatial associations, approximation and aggregation, mining in image databases, spatial classification, and spatial trend detection). However, our argumentation about those approaches was that they do not consider all the elements found in a spatial database (spatial data, non-spatial data and spatial relations among the spatial objects) in an extended way. We proposed in this dissertation a new approach based on graphs. Our feeling is that if we are able to represent those elements as a unique dataset and if we are also able to mine them as a whole, then, we might be able to

get patterns that might contain both types of data and spatial relationships enhancing the quality of the results since the generated pattern(s) would describe (a) spatial object(s) meeting (a) spatial relation(s) with other spatial object(s) and which is(are) that (those) relationship(s). We proposed to use a graph-based representation since it provides the desirable flexibility to describe these elements and their relations together.

As mentioned, in our model spatial relations among spatial objects are included since a significant characteristic of spatial data is the influence of the neighbors of an object may have on the object itself. In the model we included the representation of three types of spatial relations.

Derived from the general graph-based schema we have proposed five operative models. Three aspects define the characteristics of a graph created from those models: first, the representation of equivalent spatial relations (the relation *A* touch *B* can be represented by two directed edges, $A \rightarrow B$ and $B \rightarrow A$, or by one undirected edge, $A \text{—} B$; we used the second approach). Second, the representation of symmetric spatial relations (the relation *A* North_of *B* implies a relation *B* South_of *A*, some models represent only the first relation and other both relations). The third aspect is the way to represent the objects and their relationships. Our intention is to represent the spatial objects and their relation as much as possible but also considering a balance among the representative of the data and the complexity of the created graph. This last issue has a major importance since the tie among the complexity of the created graph and the mining phase. For example, huge graphs may require more computational resources than small graphs, but by creating small graphs we may loss data representativeness and perhaps we may not to discover significant patterns.

As component of our methodology for mining spatial data using a graph-based approach, we used the Subdue system as our mining tool. The overlap feature plays a relevant role in the Subdue's substructure discovering system. But, as we described, it is implemented in an orthodox way: allows overlap among any instances of a substructure or allows overlap among all instances of a substructure. The first option is better regarding to the processing time, but the second one may discover more instances of a substructure (patterns). Both cases do not give to the user the capability to specify the set of vertices allowed for overlapping.

Therefore, we proposed a new overlap approach named limited overlap. The new approach gives to the user the means to specify over which vertices the overlap will be allowed. These vertices may represent significant elements in the context we work with. For example, in the use-case presented in Section 6.2 (implementation of evacuation plans in case of volcanic contingences in the Popocatépetl volcano zone) vertices representing roads were allowed for overlapping since these spatial objects represent relevant elements in the evacuation plans.

Moreover, we visualized three motivation issues to propose the implementation of the new approach. First, we demonstrated that by using limited overlap we obtain a search space reduction in the substructure discovering process since we allow overlap but it is restricted to the set of elements chosen by the user. Second, as result of a search space reduction we also get a processing time reduction. Remember that as part of the substructure discovery process there exist a validation and discarding phase over the instances of a substructure.

Instances that are not discarded may become candidates to further expansion in order to discover new substructures. Third, giving the user the capability to choose the set of vertices allowed for overlapping, we are orienting the search over particular overlapping instances and, at the same time, discarding irrelevant overlapping instances.

In order to show the feasibility, capacity to mine and to discover patterns by using a graph-based approach as proposed, we developed a prototype system implementing our model to create graph-based datasets, to mine those datasets (by calling the Subdue system), and to visualize the discovered patterns.

In Chapter 6 we described three illustrative use-cases showing the applicability of our proposal. We used two real world domains: a population census from the year of 1777 in Puebla downtown, and a Popocatépetl volcano database. The generated results from those test domains give us a panorama about how and what we can achieve using our approach. It is important to remark the fact that we can use this approach in any domain that can be represented as a graph.

In this context, we visualize perspectives related to enhance our work in issues such as the graph-based model for creating the graphs, the data mining algorithm, and the prototype system. They include some of the following:

Visualization/presentation of discovered knowledge. For example, visualization of discovered knowledge over the spatial layers (patterns over the maps); to use charts for

showing comparison among patterns; give the user the capability to navigate through the discovered patterns hierarchy using a hypergraph approach.

Enhancing the algorithms used to create the graph-based datasets according to the proposed models. The validation of the spatial relations among the spatial object is phase that in most of the cases requires a lot of computational resources. (i.e., creation and manage of spatial indices, creation of bounding box for the spatial objects, manage of spatial reference systems, etc.) Therefore, the algorithms must as most as possible to execute efficiently this task.

Mining the graph. We proposed and used the Subdue system as our data mining tool, moreover we implemented a new algorithm name limited overlap. Subgraph isomorphism is an NP-complete problem, so we must be able to face this problem in order that our processing times for discovering knowledge meet acceptable parameters of efficiency.

Relationships among non-spatial data describing spatial objects. Implicit relations among non-spatial data (i.e., attributes) describing the spatial objects may be included in the model in order to enhance the representativeness of the data.

Spatial data mining is a promising research field. Several approaches have been developed, and without any doubt, new approaches will be proposed; it is a field in continuous improvement. Our contribution to the spatial data mining goes in that direction.