ABSTRACT

Motivation

Several approaches have been developed for mining spatial data (i.e., generalization-based methods, clustering, spatial associations, approximation and aggregation, mining in image and raster databases, spatial classification and spatial trend detection). However, we argue that these approaches 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. Some of them focus first on spatial data and then on non-spatial data or vice versa, and others consider restricted combinations of these elements. We think that it is possible to enhance the generated results of the data mining task by mining them as a whole and not as separate elements (they are related elements). A graph representation provides the flexibility to describe these elements together and this is the motivation to explore the area of graph-based spatial knowledge discovery.

Proposal

Our idea is to create a unique graph-based model to represent spatial data, non-spatial data and the spatial relations among spatial objects. We will generate datasets composed of graphs with a set of these three elements. We consider that by mining a dataset with these characteristics a graph-based mining tool can search patterns involving all these elements at the same time improving the results of the spatial analysis task. A significant characteristic of spatial data is that the attributes of the neighbors of an object may have an influence on the object itself. So, we propose to include in the model three relationship types (topological, orientation, and distance relations).

In the model the spatial data (i.e., spatial objects), non-spatial data (i.e., non-spatial attributes), and spatial relations are represented as a collection of one or more directed graphs. A directed graph contains a collection of vertices and edges representing all these elements. Vertices represent either spatial objects, spatial relation types between two spatial objects (binary relation), or non-spatial attributes describing the spatial objects. Edges represent a link between two vertices of any type. According to the type of vertices that an edge joins, it can represent either an attribute name or a spatial relation name. The attribute name can refer to a spatial object or a non-spatial entity. We use directed edges to represent directional information of relations among elements (i.e., object *x* covers object *y*) and to describe attributes about objects (i.e., object *x* has attribute *z*).

We propose to adopt the Subdue system, a general graph-based data mining system developed at the University of Texas at Arlington, as our mining tool. Subdue discovers

substructures using a graph-based representation of structural databases. The substructures (a connected subgraph within the graphical representation) describe structural concepts in the data (i.e., patterns). The discovery algorithm follows a computationally constrained beam search. The algorithm begins with the substructure matching a single vertex in the graph. On each iteration, the algorithm selects the best substructure and incrementally expands the instances of the substructure. An instance of a substructure in the input graph is a subgraph that matches (graph theoretically) that substructure.

A special feature named overlap has a primary role in the substructures discovery process and consequently a direct impact over the generated results. However, it is currently implemented in an orthodox way: all or nothing. If we set overlap to true, Subdue will allow the overlap among all instances sharing at least one vertex. On the other hand, if overlap is set to false, Subdue will not allow the overlap among instances sharing at least one vertex. So, we propose a third approach: limited overlap, which gives the user the capability to set over which vertices the overlap will be allowed (vertices representing remarkable elements that refer, for instance, to a spatial object in a spatial database or to some characteristic defining a particular topic of a dataset). We visualize directly three motivations issues to propose the implementation of the new algorithm: search space reduction, processing time reduction, and pattern oriented search.

Contribution

The contribution to the discovery knowledge in the spatial data domain, described in this dissertation, is the development of a new approach for spatial data modeling and mining using a graph-based representation. This contribution includes the following results:

- A new graph-based data representation for spatial, non-spatial data and spatial relations.
- A new algorithm to discover substructures using a limited overlap approach in the Subdue system.
- A prototype system implementing the proposed model.