Chapter 1

INTRODUCTION

Due to the advances in data generation and recompilation we are facing a continuous growth in data collections. The analysis and interpretation of this data by manual techniques is sometimes a tough task; therefore, different methods have been proposed to help us transform it into useful knowledge. Knowledge discovery in databases (*KDD*) is defined as the non-trivial extraction of implicit, previously unknown, and potentially useful information from data [16]. This is an interactive and iterative process that involves several phases. The data mining phase is the nucleus of the process; it consists of the application of data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produces a particular enumeration of patterns over the data [12]. Data mining involves the integration of methods from different scientific fields like machine learning, database technology and statistics. The first approaches developed focused on the discovery of knowledge from relational data.

Nowadays, however, terms like geoprocessing and Geographic Information System are used widely in daily life. This is a result of the improvement in the human capabilities to create, manipulate, store and use data from phenomena on, above or below the earth's

surface. This data is known as spatial data. Spatial data mining focuses on the discovery of implicit and previously unknown knowledge in spatial databases [16]. Spatial data has many features that distinguish it from relational data such as the relationships among the objects, complexity, and the query language used to access them.

1.1 Motivation

As a result of the growth in the volume of spatial datasets and the necessity for tools to help us transform them into useful information, several approaches have been developed for knowledge discovery from spatial data: the principle in the *Generalization-based* methods [21, 35] is that data and objects often contain detailed information at primitive concept levels but sometimes it is desirable to summarize that information and present it at a higher concept level. *Clustering* [26, 28, 37, 40, 46] is the process of grouping physical or abstract objects into classes (clusters) of similar objects so that the members of a cluster are as similar as possible whereas members of different clusters differ as much as possible from each other. *Spatial associations* [13, 29] discover rules that associate one or more spatial objects with other spatial objects. In *Approximation and aggregation* methods [27] the idea is to analyze the characteristics of the clusters in terms of the features (objects) close to them. Aggregate proximity is the measure of closeness of the set of points in the cluster to a feature. *Mining in image and raster databases* [13, 14] can be viewed as another approach of spatial data mining. Some applications of this approach (based on images) are automatic recognition and categorization of astronomical objects; classification of stars, galaxies and other stellar objects. *Spatial Classification* [31] is the task of assigning objects to a set of

classes based on their attribute values. *Spatial Trend Detection* [10] describes the regular change of one or more non-spatial attributes of an object when moving away from a given starting object.

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 separated elements (in the real world they are related). In this context, we propose to use a graph-based representation since it provides the flexibility to describe these elements together and this is the motivation to explore the area of graph-based spatial knowledge discovery.

Our work is based on the hypothesis that if we create a graph-based model to represent together spatial and non-spatial data and if we use this model for generating a dataset composed of both type of data, then we can apply data mining techniques using this knowledge representation to spatial and non-spatial data at the same time and get descriptive patterns considering both kind of data about objects and the spatial relations among them.

1.2 Proposal

Our proposal 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, therefore, to enhance the data representativeness we propose to include in the model three relationship types (topological, orientation, and distance relations). Moreover, from a general point of view spatial database systems are relational databases plus a concept of spatial location and spatial extension. So, most KDD algorithms for spatial databases must make use of those neighborhood relationships because it is the main difference between KDD in relational database system and spatial database system.

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 [24, 45], 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. We argue that a third option is needed: a limited overlap. With this option we give 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.

1.3 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 approach includes the following issues:

- We proposed a new graph-based data representation for spatial data, non-spatial data and spatial relations among the spatial objects. We visualize two objectives for creating a data model with these characteristics. The first one is to create a unique graph-based dataset representing these related elements. The second one is to use this dataset to feed a graph-based mining system, so we can discover single patterns (involving these elements) that will help us to describe/understand the data, based on the premise that they are related elements in the real world.
- We proposed a new algorithm to discover substructures (patterns) using a limited overlap approach in the Subdue system. We visualize directly three motivations issues to propose the implementation of the new algorithm: search space reduction, processing time reduction, and specialized overlapping pattern oriented search.
- We designed and developed a prototype system implementing the proposed model. The prototype provides to the user a friendly graphical user interface for managing the spatial layers to work with, for creating spatial and non-spatial graphs, for mining those graphs (by calling the Subdue system) and for displaying the generated results.

1.4 Organization of the thesis

The thesis is structured in the following way: Chapter 1 presents the motivation, proposal, and contributions of the thesis. Related work is described in Chapter 2. In Chapter 3 we detail our graph-based model to represent together spatial data, non-spatial data, and spatial relations. Our graph-based data mining tool, the Subdue system, and the new limited overlap algorithm are described in Chapter 4. A prototype system implementing our model is presented in Chapter 5. Use-cases showing the applicability of our proposal are described in Chapter 6. Finally, conclusions and final remarks are commented in Chapter 7.