Chapter 2

RELATED WORK

Knowledge Discovery in Databases, spatial data, non-spatial data, Data Mining, Spatial Data Mining, Geographic Information System, geoprocessing, and Geomatics are terms broadly used nowadays. This chapter presents an outline of these issues. The Geographic Information Systems are presented in Section 2.1. In Section 2.2 we describe related work about Knowledge Discovery in Databases and data mining. Finally, Section 2.3 describes the three types of spatial relations we propose to incorporate in our model to represent spatial data.

2.1 Geographic Information System (*GIS*)

A Geographic Information System is defined as a tool for the manipulation of geographic data [2]. The *GIS* performs a great diversity of functions; some of them are the compilation, verification, storage, retrieval, manipulation, update and visualization of geographic data. Additionally, one of its more important features is the inclusion of data analysis modules.

All these functions are applied by a *GIS* to geographic data, stored generally in a geographic database. The data processed and manipulated is *georeferenced*, that is, it is assigned to a specific location on the Earth's surface using a coordinate system.

The *GIS* can process data from several sources. For example, data collected from maps, images and photography, statistical data from mathematical analyses, and data from *CAD* systems (*Computer-Assisted Design*).

A *GIS* organizes and handles digital data stored generally in a geographic database. The databases are important in the *GIS* technology because they store geographic data with a structured form, allowing the data to be used for many tasks. Many *GIS* implement additional functionalities when they use database management systems (*DBMS*) to store and to handle all or some data in an independent subsystem.

The diversity of the uses of the *GIS* has generated the proliferation of a great variety of definitions of *GIS*. A user generally defines a *GIS* according to what he uses it for and his own experience and abilities. Some of these definitions are:

- A system for data processing designed for the production and/or visualization of maps.
- An information system to respond to questions about Earth's properties or soil types.
- A system for decision making support in situations of natural phenomena.
- An electronic positioning system to be used by terrestrial or marine transportation.

The *GIS* frequently is named according to its application field [2]. For example, when they are used to manage land's registries they are generally called Land Information Systems (*LIS*); in applications of municipal and natural resources they are important components of the Urban Information Systems (*UIS*), and Natural Resources Information Systems (*NRIS*) respectively. The term Automatic Mapping/Facility Management (*AM*/*FM*) is used by public maintenance companies, transportation agencies, and local governments for systems dedicated to the operation and maintenance of networks.

The *GIS*'s field (see Figure 2.1) involves many disciplines, applications, data types, and end users, for example:

- Disciplines: Computer Science, Cartography, Spatial Analysis, Topography, Hydrography, Statistic, Information Sciences, Planning, etc.
- Applications: Operation and maintenance of networks and other devices, administration of natural resources, highway planning, map production, urban analysis, planning analysis, etc.
- Data: Digital maps, digital images and photography, data satellites, video images, etc.
- Users: Planners, topographers, vulcanologists, geographers, environmentalist, engineers, etc.

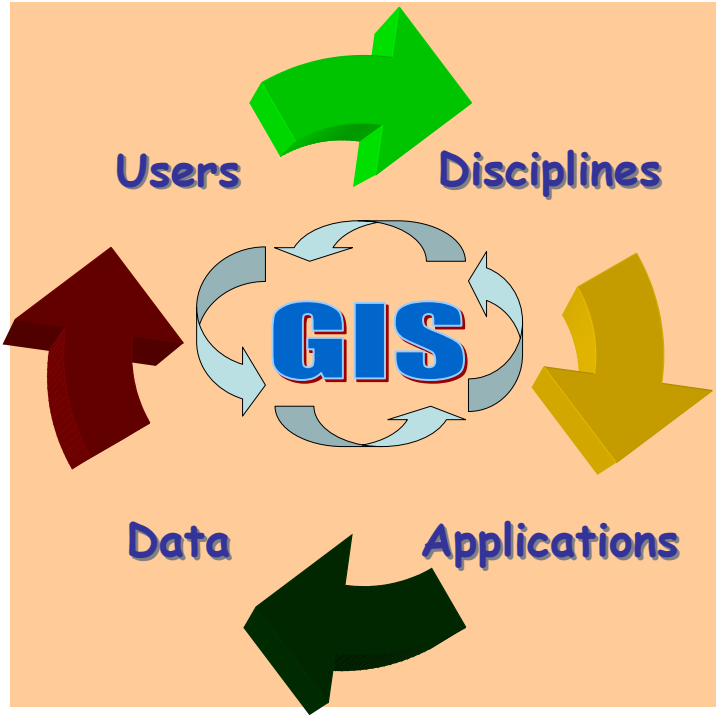

Figure 2.1. Geographic Information System.

Geographic Information Systems involve the uses of systems and science. Their use arise questions such as: How does a *GIS* user know that the results obtained are accurate? How can user interfaces be made readily understandable by novice users? M. F. Goodchild published in 1992 a paper where he argued that questions such as these and their systematic study constituted a science.

Geoprocessing study the fundamental issues arising from geographic information (i.e., creation, handling, storage and use of the information). The term *Geomatics*, the fusion of ideas from geosciences and informatics, is defined as the umbrella covering all fields that are today important for understanding and further developing information systems in this context [32, 33, 34].

A *GIS* offers its users greater capabilities to process datasets than those offered by manual systems. In a *GIS* database the data is stored in a structured form, unlike the manual systems where the data is stored in files, maps, and/or reports. The data can be recovered from geographic databases and processed faster and more safely than in manual systems.

We can classify the *GIS*'s users into two groups. In the first group, we find professional operators. They are people trained in some particular software and they know the capabilities of this technology. Many times these people do not use the results of their work, but they pass them on to the end users.

The second user group spends less time working with the *GIS*'s. They maintain geographic information in order to have tools which help them in the decision making process. They have few opportunities for extensive training in the *GIS* tools, and consequently the *GIS* must be simple and easy to handle.

2.2 Data Mining

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: data preparation, search for patterns over the data, evaluation and interpretation of discovered patterns, and refinement of the whole process as it is shown in Figure 2.2.

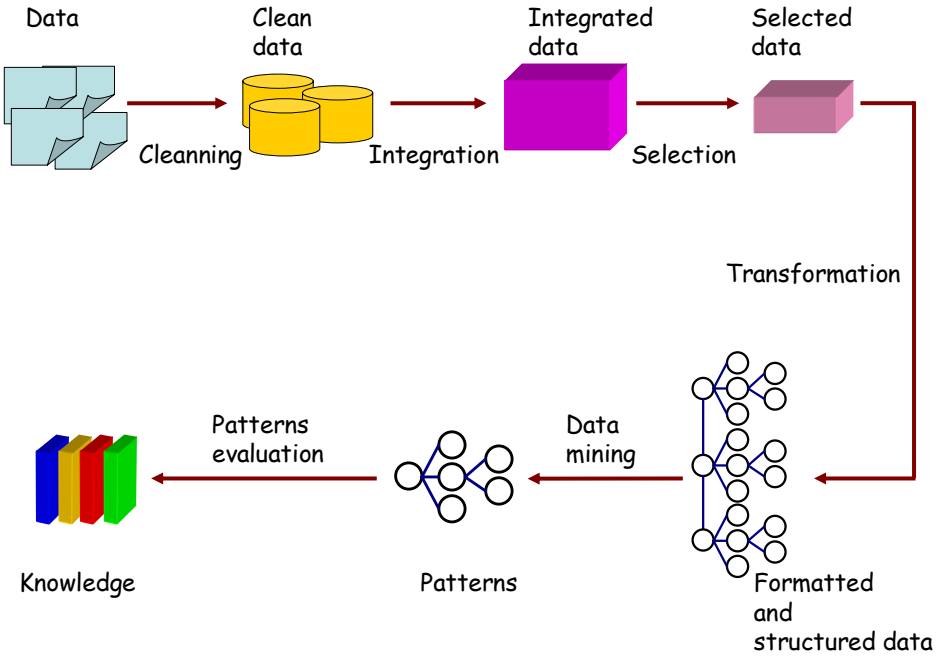

Figure 2.2. Knowledge Discovery in Databases.

The data mining phase (the search for patterns) 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 [4, 12].

Data mining involves the integration of methods from different scientific fields such as machine learning, database technology, statistics, and visualization as shown in Figure 2.3. The first approaches developed focused on the discovery of knowledge from relational data.

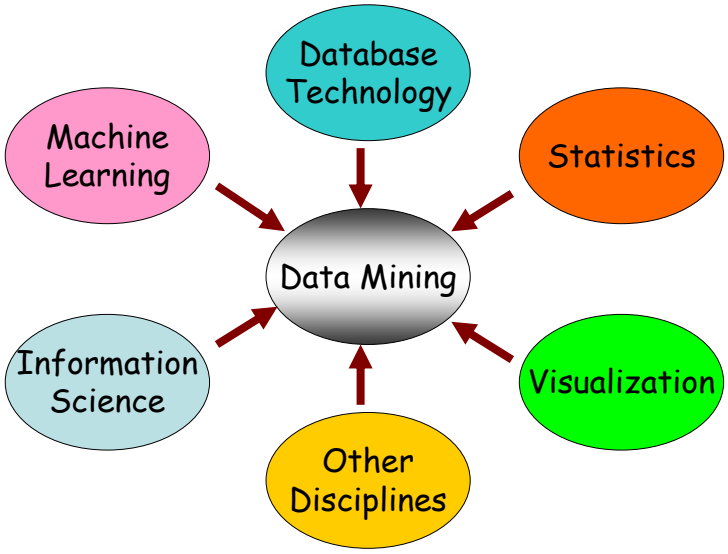

Figure 2.3. Data mining: integration of several fields.

Several architectures have been proposed for data mining [22, 25]. Figure 2.4 presents an architecture based on the proposal of Han et al. [23]. The *user* is the trigger of the entire process and receptor of the discovered knowledge. *Data* may be fetched from several sources such as files, databases, and data warehouses using a *data server* module. The *data mining engine* may use one or more data mining techniques for searching patterns from data. The significance, importance and interestingness of the found patterns are evaluated by the *pattern evaluation* module. The *data mining engine* and *pattern evaluation* modules may use background knowledge stored in a *knowledge database*. The role of the *graphical user interface* is to receive the user requirements and to deliver the generated results. Since this is an iterative and interactive process, the components may interact among themselves.

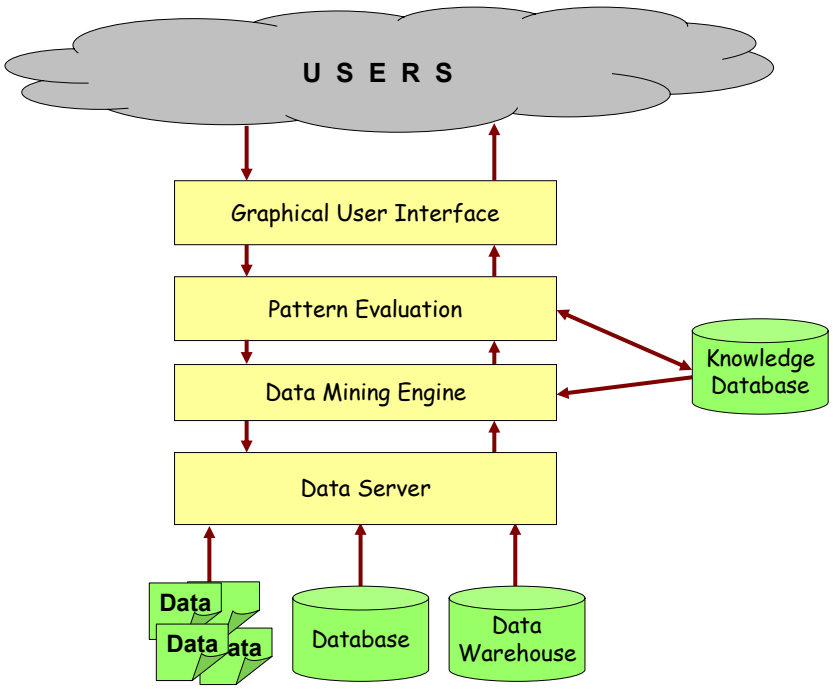

Figure 2.4. Architecture for a KDD system.

2.2.1 Spatial Data Mining

Climate change, natural risk prevention, human demography, deforestation, and natural resources atlas are examples from a large variety of issues arisen as a result of the interaction among people and their natural environment, the Planet Earth. Data generated from those issues are known as spatial data. Spatial data mining methods focuses on the discovery of implicit and previously unknown knowledge in spatial databases [16].

Spatial data have many features that distinguish them from relational data. For example, the spatial objects may have topological, distance, and direction information, the complexity,

and the query language used to access them. Different approaches have been developed for knowledge discovery from spatial data such as Generalization-based methods [21, 35], Clustering [26, 28, 37, 40, 46], Spatial associations [13, 29], Approximation and aggregation [27], Mining in image and raster databases [13, 14], Classification learning [31], and Spatial trend detection [10]. In the following subsections we present a brief description of these spatial data mining approaches.

2.2.1.1 Generalization-based Method

Generalization has been shown to be one of the effective methods of discovering knowledge. It was introduced by the machine learning community and it is based on learning from examples. Generalization-based knowledge discovery requires concept hierarchies (given explicitly by experts or generated automatically). In the case of the spatial databases, there can be two types of concept hierarchies:

- Thematic hierarchies. We can generalize tomatoes and bananas as fruits, fruits and vegetables as cash crops.
- Spatial hierarchies. We can generalize some geographic points as a country or a region.

The approach introduced by the machine learning community (tuple-oriented) cannot be directly adopted for large spatial databases because it does not handle very well the noise and inconsistent data, and the algorithms are exponential in the number of examples. Han et al. [21] present a modified technique named attribute-oriented induction for mining

relational data. Lu et al. [35] extended the attribute-oriented induction technique to spatial databases. Attributed-oriented induction is performed by climbing the generalization hierarchies and summarizing the general relationships between spatial and non-spatial data at a higher concept level. The authors present two generalization based algorithms:

- **Non-spatial data dominant.** This method performs attribute-oriented induction on the non-spatial attributes, first generalizing them to a higher concept level and later merges corresponding spatial attributes (using spatial merge and approximation).
- **Spatial data dominant.** Given the spatial data hierarchy, generalization can be performed first on the spatial data and then generalizing their corresponding non-spatial attributes.

Both algorithms assume that the rules to be mined are general data characteristics (characteristics rules) and that the discovery process is initiated by the user who provides a learning request. A disadvantage in this approach is the case where a hierarchy may not exist or the hierarchy given by the experts may not be entirely appropriate in some cases. The quality of mined characteristics is highly dependent on the structure of the hierarchy.

2.2.1.2 Clustering

Clustering is the process of grouping physical or abstract objects into classes of similar objects. Clustering analysis helps construct meaningful partitioning of large set of objects. It identifies clusters, or densely populated regions, according to some distance measurement in a multidimensional dataset. Given a set of multidimensional data points, the data space is usually not uniformly occupied by the data points. Data clustering

identifies the sparse and the crowded places, and hence discovers the overall distribution patterns of the dataset. We can classify clustering algorithms in four main approaches: Partitioning, Hierarchical, Locality-based and Grid-based algorithms.

- Partitioning algorithms partition a database of *n* objects into a set of *k* clusters which are represented by the gravity of the cluster (*k*-means algorithms) or by one representative object of the cluster (*k*-medoid algorithms). These algorithms use a two-step procedure. First, they determine *k* representatives, next, assign each object to the cluster with its representative closest to the considered object.
- Hierarchical clustering algorithms decompose the database into several levels of partitioning which are usually represented by a *dendrogram*. The algorithm iteratively splits the database into smaller subsets until some termination condition is satisfied. The *dendrogram* can either be created *top-down* (divisive) or *bottom-up* (agglomerative).
- Locality-based clustering algorithms group neighboring data elements into clusters based on local conditions and therefore allow the clustering to be performed in one scan of the database.
- Grid-based algorithms quantize the space into a finite number of cells and then do all operations on the quantized space. They frequently use hierarchical agglomeration as one of their processing phases.

Classification of clustering algorithms is neither straightforward, nor canonical. Some algorithms perform clustering by combining techniques from these approaches. Important issues in clustering algorithms include the following properties:

- The algorithm must be efficient (time complexity).
- Ability to handle noise (outliers).
- Non-sensibility to data input order.
- A Priori knowledge and parameters tuning.
- Ability to find clusters of arbitrary shape.
- Scalability to large databases.
- Ability to work with high dimensional data.

2.2.1.3 Spatial Associations

A spatial association rule is a rule which describes the implication of one or a set of features by another set of features in spatial databases [29]. An example of a spatial association rule is "if the company is close to México City then is a big company".

A spatial association rule is of the form $X \rightarrow Y$, where X and Y are sets of spatial or non-spatial predicates. There are various kinds of spatial predicates that could constitute a spatial association rule. Some examples are topological relations such as *intersects*, *overlaps*, *disjoint*; and spatial orientations such as *left_of*, *west_of*.

Koperski and Han [29] developed an algorithm for spatial associations rules in spatial databases. They use the concepts of minimum support and minimum confidence introduced by Agrawal et al. [1] to develop association rules from large transactional databases. The support of a pattern A in a set of spatial objects S is the probability that a member of S satisfies pattern A, and the confidence of $A \rightarrow B$ is the probability that a pattern B occurs if

pattern *A* occurs. A user or an expert may assign thresholds to confine the rules to be discovered to be strong ones.

Although many spatial association rules may exist in large databases, some of them may occur rarely or may not hold in most cases. In addition, such rules are usually not 100% accurate and may contain non trivial knowledge.

The authors employ a method which uses a top-down progressive deepening search technique. The technique firstly searches at a high concept level for large patterns and strong implications relationships among the large patterns at a coarse resolution scale. Then only for those large patterns, it deepens the search to lower concept levels. Such deepening search process continues until no large patterns can be found. The search employed for large patterns at high concept levels is applied at a coarse resolution scale efficiently by using approximate spatial computation algorithms such as *R-trees* [20] or *plane-sweep* [39] techniques operating on minimum bounding rectangles (*MBR*). Only the candidate spatial predicates, which are detailed reviewed, will be computed by refined spatial techniques. Such multiple-level approach saves much computation because it is very expensive to perform detailed spatial computation for all possible spatial association relationships.

2.2.1.4 Approximation and Aggregation

Clustering algorithms are effective and efficient methods for answering questions such as: where are the clusters in the spatial database? In some cases it is also important to answer the question why the clusters are there. We can rephrase the question as: what are the

characteristics of the clusters in terms of the features (objects) that are close to it? Knorr and Ng [27] present a study based on this question.

The aggregate proximity is the measure of closeness of the set of points in the cluster to a feature as opposed to the distance between a cluster boundary and the boundary of a feature. Finding aggregate proximity relationships is not as simple as it may seem. There are three reasons:

- The sizes and shapes of the cluster and the features may vary greatly.
- There may be a very large number of features to examine.
- Even if a suitable feature (i.e., polygon) is found to describe the shape of the cluster of points, it is inappropriate to simply report those features whose boundaries are closest to the cluster's boundary, because the distribution of points in a cluster may not be uniform.

The authors propose the use of computational geometry concepts to find out the characteristics of a given cluster in terms of the features close to it. They present the algorithm *CRH* (where *C* is for encompassing circle[1], *R* for isothetic rectangle[2], and *H* for convex hull[3]) which uses concepts as filters to reduce the candidate features at multiple levels. In general, they collect a large number of features from multiples sources (i.e.,

[1] A circle that encloses a set of *n* points. Not necessarily minimum bounding.

[2] A rectangle that is orthogonal to the coordinate axes.

[3] This is the unique, minimum bounding convex shape enclosing a set of points.

maps) and feed them along with the cluster to the algorithm *CRH* and discover knowledge about spatial relationships.

Approximation by circles and then by rectangles is used to eliminate features that have large aggregate distance to the cluster. After these filters, the algorithm calculates the aggregate proximity of points in the cluster to the convex boundary of each feature that passed through the previous filters. In the last step, the algorithm reports the features with the best aggregate proximities showing the minimum and maximum distances of points in the cluster to the feature, average distance, and percentages of points located in the distance less than specified threshold.

2.2.1.5 Mining an Image Database

Knowledge mining from image databases can be viewed as a special case of spatial data mining. For example, Fayyad et al. present a system [14] to identify and categorize volcanoes on the surface of Venus from images taken by the Magellan spacecraft. Three basic components are implemented in the system: data focusing, feature extraction, and classification learning. The first component increases the overall efficiency of the system by first identifying the portion of the image being analyzed that is most likely to contain a volcano. They compare the intensity of the central pixel of a region to the estimated mean background intensity of its neighborhood pixels. The second component extracts interesting features from the data. The final component uses training examples provided by the experts to create a classifier that can discriminate between volcanoes and false alarms. For this task the authors implement decision trees.

Other studies of mining in image and raster databases are: Second Palomar Observatory Sky Survey [15], it uses decision trees for the classification of galaxies, stars and other stellar objects. Stolorz and Dean [43] proposed a system for detecting earthquakes from space. They combined methods of statistical interference, massively parallel computing, and global optimization to build the system that analyze tectonic activities with sub-pixel resolution over a large area. Stolorz et al. [44] and Shek et al. [41] carry out studies about fast spatio-temporal data mining from geophysical datasets; they described CONQUEST, a distributed parallel querying and analysis mining tool.

2.2.1.6 Classification Learning

Spatial classification has as objective to find rules that divide a set of objects into a number of groups, where objects in each group belong mostly to one class. Many types of information can be used to characterize spatial objects. We can classify such information into non-spatial attributes of objects, spatially related attributes with non-spatial values, spatial predicates and spatial functions.

Each of these categories may be used to extract both for class label attributes (attributes that divide data into classes) and predicting attributes (attributes on whose values the decision tree is branched). It is possible to use aggregate values for some of these attributes.

Koperski et al. [31] proposed and evaluated a method for classification of spatial objects. The method enables classification of spatial objects based on aggregate values of non-

spatial attributes for neighboring regions. Spatial relations between objects on the map are taken into a count, which may be represented into the form of predicates.

2.2.1.7 Spatial Trend Detection

Spatial trend detection is defined as a regular change of one or more non-spatial attributes in the neighborhood of some object in a database [10]. An example of spatial trend is as "moving away from downtown Puebla, the price of land decreases".

Neighborhood paths starting from some point *x* are used to model the movement and a regression analysis is performed on the respective attribute values for the objects of a neighborhood path to describe regularity of change. In the regression analysis the distance from *x* is the independent variable and the difference of the attribute values are the dependent variable(s). There are two types of trends: global trends and local trends. In the first one, the existence of a global trend for a start object *x* indicates that if considering all objects on all paths starting from *x* the values for the specified attribute(s) in general tend to increase or decrease with increasing distance. Local trends exist only in a particular direction.

2.3 Spatial relations

The explicit location and extension of objects define implicit relations of spatial neighborhood. Therefore, the information about the neighborhood of spatial objects constitutes a valuable element that must be considered in the mining task. In the following

subsection we will present the Neighborhood Graphs, Neighborhood Paths and Neighborhood indices concepts that introduce us to the three types of spatial relations we propose to include in our model to represent together spatial data.

2.3.1 Neighborhood Graphs, Neighborhood Paths and Neighborhood Indices

Martin Ester et al. [9, 11] introduce the concept of neighborhood graphs for explicitly representing those implicit neighborhood relations relevant for the KDD tasks. He claimed that attributes of the neighbors of some object of interest may have an influence on the object and therefore have to be considered as well. He commented that the efficiency of many KDD algorithms for spatial database systems depends heavily on an efficient processing of these neighborhood relationships. Neighborhood graphs may cover one of the following neighborhood relations:

- Topological.
- Distance.
- Direction.

These relations are called binary relations since we can determine spatial relations between pairs of objects.

A neighborhood graph G for some spatial relation neighbor is a graph where nodes are objects in the database and edges between nodes n_1 and n_2 represent the fact that the relationship neighbor (n_1, n_2) holds. The neighborhood path in the neighborhood graph G is

a set of nodes directly connected through the edges of the graph. The neighborhood index is a data structure that allows for an efficient execution of the operations for the construction of a graph and for browsing and expansion of paths. It stores all neighbors for the objects in a database.

2.3.2 Topological Relations

Topological relations are those relations which are invariant under linear transformations, i.e., if both objects are rotated, translated or scaled simultaneously the relations are preserved. They present a definition of topological relations derived from the nine intersection model [6, 7, 8].

In the model the topological relations between two objects are defined in terms of the intersections of the interiors, boundaries and exteriors of the objects (see Figure 2.5). The interior of an object consists of points that are in the object but not on its boundary, and the exterior consists of those points that are not in the object (its complement). The Figure 2.6 shows the nine intersection model of two objects A and B: object *A's* interior (A^{o}), object *A's* boundary (∂A) and object *A's* exterior ($\bar{A}$) with object *B's* interior (B^{o}), object *B's* boundary (∂B) and object *B's* exterior ($\bar{B}$).

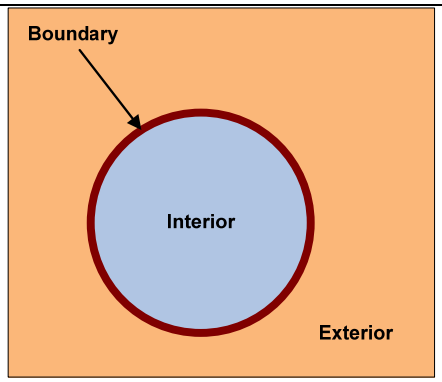

Figure 2.5. Example of the interior, boundary and exterior of a circle.

$$\begin{pmatrix} A^{\circ} \cap B^{\circ} & A^{\circ} \cap \partial B & A^{\circ} \cap B^{-} \\ \partial A \cap B^{\circ} & \partial A \cap \partial B & \partial A \cap B^{-} \\ A^{-} \cap B^{\circ} & A^{-} \cap \partial B & A^{-} \cap B^{-} \end{pmatrix}$$

Figure 2.6. Nine intersection model.

In Figure 2.7 we present the topological relations between two objects:

- Disjoint. The boundaries and interiors of the objects do not intersect.
- Contains. The interior and boundary of one object is completely contained in the interior of the other object.
- Inside. The opposite of contains.
- Equal. The two objects have the same boundary and interior.
- Touch. The boundaries of the objects intersect but the interiors do not intersect.
- Covers. The interior of one object is completely contained in the interior of the other object and their boundaries intersect.
- CoveredBy. The opposite of covers.
- Overlap. The boundaries and interiors of the two objects intersect.

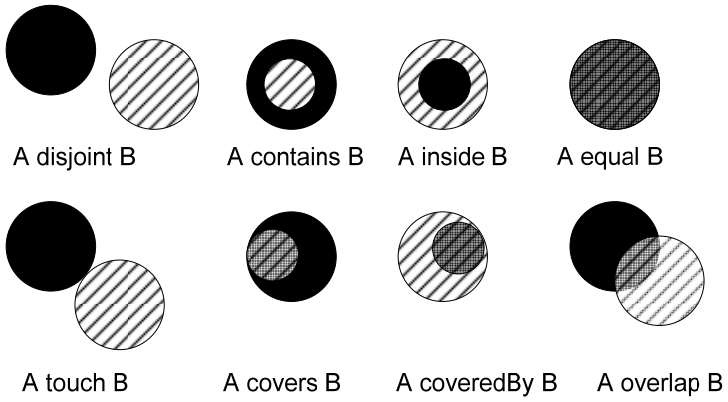

Figure 2.7. Topological Relations.

2.3.3 Distance Relations

Distance relations compare the distance between two objects with a given constant using arithmetic operators such as <, >, =. The distance between two objects is defined as the minimum distance between them (i.e., select all elements inside a radio of 50 km from a "*x*" point). Figure 2.8 shows two examples of this type of relation; in the figure we are representing how close and how far two objects are each other.

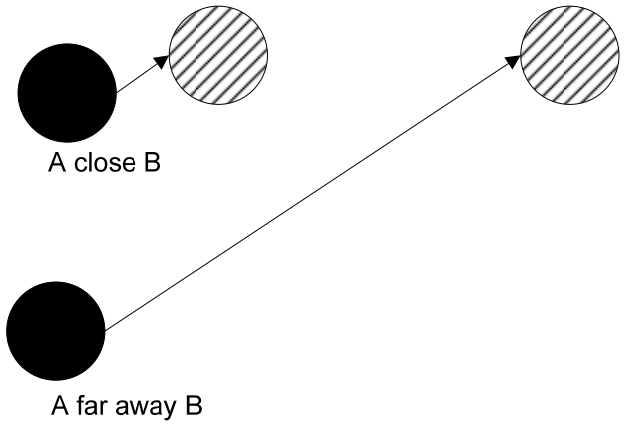

Figure 2.8. Distance Relations.

We propose to use the Euclidian distance which is defined as the straight line distance between two points. In a plane with *p1* at (x_1, y_1) and *p2* at (x_2, y_2), the Euclidian distance is $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$.

2.3.4 Direction Relations

The direction relation ***R*** between two spatial objects *A* and *B*, (*A* ***R*** *B*) is defined using one representative point of object *A* and all the points of the destination object *B*. It is feasible to define several possibilities of direction relations depending on the number of points that are considered in the source and destination objects. The representative point of a source object may be the center of the object or a point on its boundary. This representative point is used as the origin of a virtual coordinate system and its quadrant defines the direction.

The direction relations between two objects are *North_of*, *South_of*, *East_of*, *West_of*, *Northeast_of*, *Northwest_of*, *Southeast_of*, and *Southwest_of*. In Figure 2.9 we show some examples of direction relations, for instance, object *D* is South of object *C* and East of object *A*.

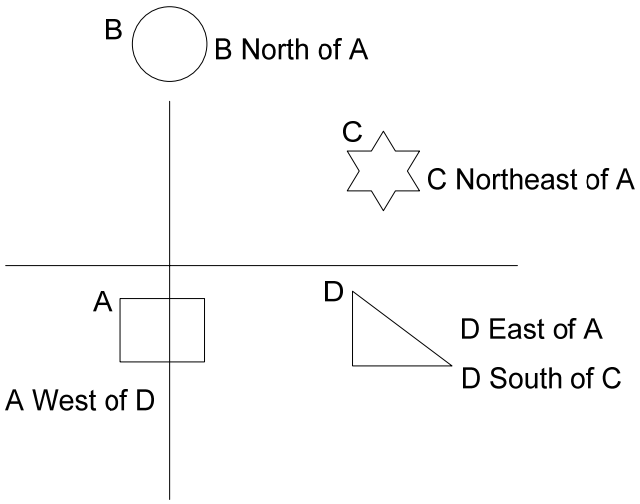

Figure 2.9. Direction Relations.

2.4 Conclusion

Data mining is a younger and promissory research field. Many of the data mining approaches developed for knowledge discovery in relational databases were extended to the spatial databases domain. In this chapter we presented several approaches such as generalization, clustering, spatial association, and spatial classifications. We have described three types of spatial relations that will be included in our model to represent as a unique graph-based dataset spatial data, non-spatial data and spatial relations.

In the next chapter we will describe the model. First, we will talk about the graph-based knowledge discovery, next we will detail our methodology, and finally we will present an example showing its applicability and generated results.