Knowledge Discovery in Spatial Databases Progress and Challenges

Junas Adhikary
School of Computing Science
Simon Fraser University
Burnaby, B.C., Canada V5A 1S6

Abstract

Spatial data, i.e., data related to objects that occupy space, are continuosly being collected for various applications ranging from remote sensing, geographical information systems (GIS) to computer cartography and environmental assesment and planing. The volume of data collected is so huge that it has become humanely impossible to do any intelligent data analysis. Even though very few methods have been proposed and applied to discover knowledge from spatial data, it is evident that the techniques from machine learning, database technology and statistics have to be explored and modified for further progress in this field. In this paper we intend to survey the progress and outline some challenges facing spatial data mining.

1 Introduction

Advancement in data collection technolgy, for e.g. barcode reading, remote sensing etc., has made data collection a common everyday routine. The data thus collected has grown exponentially and is now impossible for any human to intelligently analyse them. This brought forward an interesting problem which is known as database mining, knowledge discovery in databases, data archeology, data dredging, etc. [42, 30, 15]. Knowledge discovery in spatial databases, or *spatial data mining*, can be defined as the *discovery of interesting, implicit, and previously unknown knowledge* [19] from large spatial databases. Data mining in general has become a cross product of different fields of research including machine learning, database and statistics. Obviously, spatial data mining is no different since the goal of both are the same. A direct consequence of this, as can be expected, is that there are a number of issues coming forth from these fields of studies that relate to the progress and development of spatial data mining. There has been a considerable number of studies being done in machine learning (for e.g. [38, 39, 16]), databases (for e.g. [26, 1, 14]) , and statistics (for e.g. [18, 32, 10]) that somewhat lay the foundation for knowledge discovery in spatial databases.

Development of data mining is challenging to every field discussed above, escpecially due to the fact that the data considered in this case is very large, upto several megabytes. For spatial data mining in particular, advances in areas of research like spatial data structures (e.g. [44, 23]), spatial reasoning [31] etc., are also of great importance. Algorithms for spatial data mining has to be very efficient, anything above the $\mathcal{O}(n)$ complexity, where n is the number of spatial objects, can be questionable. Knowledge to be discovered from spatial data can be of various forms, for example, characteristic features, prominent structures or clusters, *how* and *why* do they exist, spatial associations etc.

Spatial data mining methods can be applied to extract interesting and regular knowledge from large spatial databases. In particular, it can be used for understanding spatial data, discovering relationships between spatial and non spatial data, data reorganization in spatial databases, capturing the general characteristics in fairly simple and concise manner, etc. This has wide application in Geographic Information Systems (GIS), image databases exploration, medical imaging, robot navigation etc. The purpose of this survey is to provide a global picture of the methods of spatial data mining, their strengths and weaknesses, how and when to apply them, and to determine what we have achieved so far and what are the challenges we face.

1.1 Spatial Data Mining Background

Until recently, the most common approach of analyzing spatial data has been statistical spatial analysis techniques (for e.g. [18]). Statistical analysis methods is a well studied area and therefore has a large number of algorithms with various optimization techniques. It handles numeric data very well and usually comes up with realistic models of spatial phenomena, although without any background knowledge. But the major disadvantage of this approach is the assumption that requires statistical independence among the spatially distributed data. This causes problems as many spatial data are infact interrelated, i.e., spatial objects are influenced by the neighbouring spatial objects. Regression models are used to alleviate this problem to some extent but it makes the whole modelling process more complicated and can only be done by experts with a fair amount of domain knowledge and statistical expertise. In other words, it is not the kind of techniqe that we want to present to the end users for spatial data mining. Furtermore, this approach can not model nonlinear rules very well and lastly although numeric values are handled well, symbolic values like names are handled poorly.

With the advent of data mining, reserachers proposed various methods for discovering knowledge from large databases, almost all of them concentrating on relational databases. These methods strived to combine the already mature areas like machine learning, databases and statistics. Studies like [42, 26, 1, 15] lay a foundation for spatial data mining. The study by Lu *et al.* [36] is probably the first one to extend data mining techniques to spatial databases. Machine learning techniqes *learning from examples* and *generalization and specialization* are widely used in such algorithms. It did not take long before statistical cluster analysis technique became noticeable and was modified for the use in spatial data mining. In the next section, we define some commonly used terms in spatial data mining.

1.1.1 Primitives of Spatial Data Mining

RULES: Various kinds of rules can be discovered from databases in general, for e.g. characteristic rules, discriminant rules, association rules, deviation and evolution rules etc. We adopt the definition of the rules in terms of spatial databases as given in Koperski and Han [34]. A *spatial charancteristic rule* is a general description of spatial data, for e.g., rule describing general price range of the houses in various geographic regions in a city. A *spatial discriminant rule* is a general description of the discriminating or contrasting features of a class of spatial data from other class(es), for e.g., comparison of price range of houses in differenct geographical regions. Finally, 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 databses, for e.g., rule associating price range of the houses with the nearby spatial features, like beaches.

THEMATIC MAPS: Thematic maps encode spatial objects in relation to a theme. Different thematic maps are analysed for discovering different rules, for e.g., we may want to look at *temperature* thematic map while analysing the general weather pattern of a geographic region. There are two ways to represent thematic maps: *raster* and *vector*. In the raster image form thematic maps have a pixel associated with an attribute value. For example, a map may have the height of the spatial objects coded as the intensity of the pixel (or the color). On the other hand, in the vector representation, an spatial object is represented by its geometry, most common being the boundary representations along with the thematic atributes. For example a park may be represented by the boundary points and corresponding elevation value (height).

COMPUTATIONAL GEOMETRY TERMS: Some concepts arising from computational geometry are frequently used in spatial data mining. They include:

- *Point:* Having x and y coordinate in 2-dimensional Eulidean space.
- *Closed Curve:* A boundary of points with no opening.
- *Polygon:* A simple closed curve having k sides and k vertices, also called k-gon.
- *Cluster:* A set of points.
- *Medoid:* Most centrally located ppoint in a cluster.
- *Centroid:* The point $(\sum_{i=1}^{n} x_i/n, \sum_{i=1}^{n} y_i/n)$ is centroid of n points (x_i, y_i).

IMAGE DATABASES: These are special kind of spatial databases where data almost entirely consists of images or pictures. These databases arise from remote sensing, medical imaging, etc.

1.1.2 Spatial Data Structures, Computations, and Queries

Algorithms for spatial data mining involve the use of spatial operations (computations) like, joins, overlays e.t.c. Therefore, efficient spatial access methods (SAM) and data structures for such computation is also a concern in spatial data mining. We will briefly introduce some of the prominent spatial data structures and spatial computations.

Spatial Data Structures: Spatial data structure consists of points, lines, rectangles, etc. Most spatial data contain multudimensional point data. In order to access or to build indices for these data, multidimensional trees have been proposed. These include quad tress, k-d trees, R-tree and its variant R*-tree, etc. Readers are encouraged to see Samet [44] for excellent survey on this topic. One of prominent SAM much discussed in the literature recently is R*-tree developed by Beckmann *et al.* [6]. It is a modification of R-tree developed by Guttman [23]. In general, they are used to store points and minimum bounding rectangles (MBR). Combination of SAM and computational geometry in Spatial databases is discussed in [24].

Spatial Computations: *Spatial join* is one of the most expensive spatial operation. In order to make spatial queries (e.g. [5]) efficient spatial join has to be efficient as well. Brinkhoff *et al.* propose an efficient processing of spatial joins using R-Trees [8]. Another important spatial operation is *map overlay* (see [9] for details).

Spatial Query Processing: Optimization strategies for spatial query processing is outlined in Aref and Samet [5]. The authors propose an architecture for spatial database called SAND (spatial and nonspatial data) architecture, which is a model of extended relational database with spatial operations. This architecture provides both spatial and nonspatial components of spatial databse to participate in query processing and optimization.

1.2 Spatial Data Mining Architecture

Various architecture (model) has been proposed for data mining in general. They include Han's architecture for general data mining prototype DBLEARN/DBMINER [25], Holsheimer *et al*'s parallel architecture [29], Matheus *et al.*'s multicomponent architecture [37] etc. Almost all of these architectures have been used or extended to handle spatial data minign as well. Matheus *et al.*'s architecture seems to be very general and is used by other researchers in spatial data mining as well, including Ester *et al.* [13]. This architecture is comparable to others. We outlilne this architecture in Figure 1. In this architecture, user is given full control and background knowledge (e.g. concept hierarchies) resides in knowledge base. Data is focussed using query from user interface and interesting patterns are extracted. These patterns are then evaluated to discover knowldge.

1.3 Scope of Literature Survey

Spatial data mining as a research field is only about three years old. Therefore, it is not widely known or developed yet and we have only looked at conference reports or journals for the last 10 years at most. However, it is getting a lot of attention from researchers in various fields dealing with spatial data. We mainly surveyed *Knowledge Discovery in Databases (KDD) workshops*, *International Symposium on Spatial Databses*, related papers in *ACM SIGMOD Proceedings*, *IEEE Transactions on Knowledge and Data Engineering*, *VLDB conference proceedings*, and some other conferences on Geographic information Systems. Apart from conference reports, some excellent books on spatial data structures and computational geometry were also looked at.

1.4 Organization of the paper

The rest of the paper is organized as follows. In Section 2 we survey the methods for spatial data mining. We categorize the methods and discuss each in detail. Section 2.1 describes generalization based methods, Section 2.2 discusses clustering based methods, Section 2.3 discusses methods to explore spatial associations, Section 2.4 describes pattern recognition methods, and finally in Section 2.5 other interesting methods are outlined. We present future suggestions and directions in Section 3, and conclude in Section 4.

2 Methods for Knowledge Discovery in SD

Geograhical data in general have *thematic* and *spatial* data. Thematic data is alphanumeric data related to spatial objets, e.g. average temperature of a region. Spatial data on the other hand is

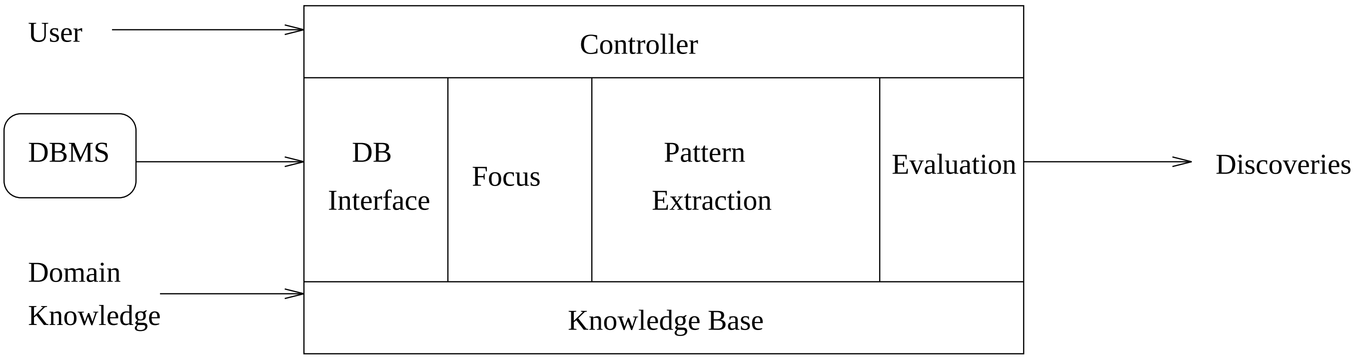

Figure 1: An architecture for a KDD system

described using two different properties, geometric and topological. For example, geometric properties can be spatial location, size etc. whereas topological properties can be adjacency (object A is *left of* object B), inclusion (object A is *included* in object B) etc. Thus, methods for discovering knowledge can be focussed on the thematic and/or the spatial properties of spatial objects in a spatial database.

As various rules can be mined from spatial databases, various algorithms has been proposed accordingly. They include generalization-based methods for mining spatial characteristic and discriminant rules [26, 36, 41], two-step spatial computation technique for mining association rules [34], aggregate proximity technique for finding characteristics of spatial clusters [33], spatial reasoning or topological inferencing to find spatial relationships [12, 20] etc. In the following sections, we catagorize and describe a number of algorithms (methods).

2.1 Generalization Based Knowledge Discovery

One of the effective methods of discovering knowledge has been shown to be the use of learning from examples technique with generalization. It was introduced by the machine learning community and has been used in machine learning algorithms (e.g. [39]). Their approach can not be directly adopted for large spatial databases because, 1) the algorithms are exponential in the number of examples, 2) it does not handle noise and inconsistent data very well. Han *et al.* [26] modify these techniques and give an attribute-oriented (as oopsed to tuple-oriented in machine learning algorithms) induction algorithm to mine knowledge from large relational databases. Later Lu *et al.* extend the technique to spatial databases. Thus, the assumptions that are made for relational databases are also carried to the spatial data mining.

The generalization based knowledge discovery requires concept hierarches. In the case of spatial databases, there can be two kind of concept hierarchies, thematic and spatial. Attribute-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. Lu *et al.* give two generalization based algorithms, *spatial data dominant* and *non spatial data dominant* generalizations. Both algorithms assume that the rules to be mined are general data characteristics (characteristics rule) and that the discovery process is initiated by the user who provides a learning request (query) explicitly, in a syntax similar to SQL. Moreover, the spatial databases is assumed to consist of both spatial as well as nonspatial data. Concept hierarchies can be explicitly given by the experts or can be generated automatically by statistical data analysis. An example of the concept hierarchy for the concept *agriculture* is shown in Figure 2. As we ascend the concept tree, information becomes more and more general, but still remains consistent with the lower concept levels. For example, in Figure 2 both *jasmine* and *basmati* can be generalized to concept *rice* which in turn can be generalized to concept *grain*, which will also include *wheat*. We will briefly describe both algorithms as

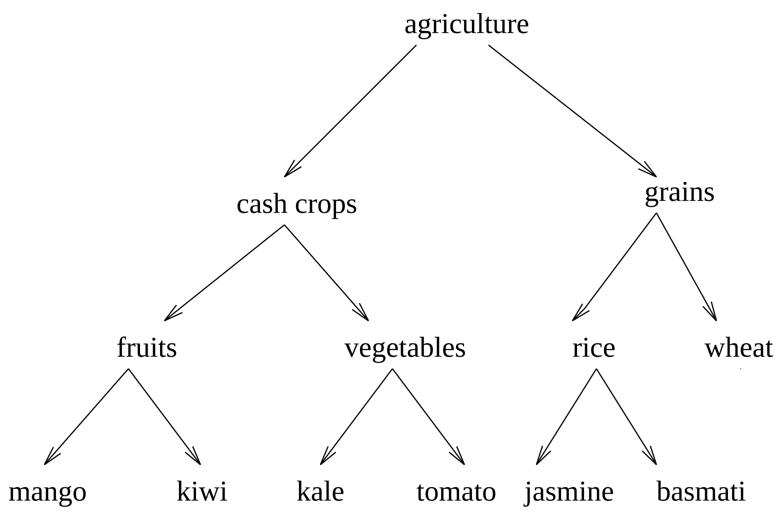

Figure 2: Example of *agriculture* concept hierarchy

follows:

Nonspatial-Data-Dominat Generalization: This method performs attribute-oriented induction on the nonspatial attributes, thus generalizing them to a higher (more general) concept level first and then merges corresponding spatial attributes later (using spatial merge and approximation). *Generalization threshold* is used to determine whether to continue or stop the generalization process. The computational complexity of this algorithm is proven to be $\mathcal{O}Nlogp$, where N is the number of spatial objects, and p the size of generalized relation.

Spatial -Data-Dominant Generalization: Given the spatial data hierarchy, generalization can be performed first on the spatial data and then generalizing their corresponding nonspatial attributes. Generalization of the spatial objects continue until the *generalization threshold* is reached. After this spatial-oriented induction process, nonspatial data are retrived and generalized for each of the spatial objects using attribute-oriented induction technique. The computaitnal complexity of the algorithm is again $\mathcal{O}Nlogp$, where N is the number of spatial objects and p the size of the generalized relation.

The above algorithms only find the characteristics of the spatial data, that is, only the characteristic and discrimant rules are mined. As mentioned in Section ??, other rules can be mined from databases as well. Interestingly, generalization and concept hierarchy is used in the study proposed by Koperski and Han [34], which mines spatial association rules. The authors modify the technique of mining multiple level association rules from large transaction databases to discover spatial association rules from large spatial databases. We will discuss this technique in Section 2.3.

2.1.1 Summary and Discussion

We discussed two generalization based algorithms that assumed the concept hierarchies to be given or generated automatically. However, as pointed out before, there may be cases where such hierarchies may not be present *a priori*. Another problem with previous algorithms is that the spatial components of the databses is explored by region merging, merging regions at lower levels of the concept hierarchy to form region(s) at higher lwvel of the hierarchy. BOth of these suggest that the quality and the interestingness of the characteristic rules mined is going to be very much dependent upon the given concept hierarchy(ies). In many cases such hierarchies are given by the experts and they may be not entirely appropriate. Therefore, we would like to find algorithms that does not need to use these hierarchies at all. We will describe such algorithm in the next section.

2.2 Methods Using Clustering

Cluster analysis is a branch of statistics that has been extensively studied for many years. The main advantage of using this technique is that interesting structures or clusters can be found directly from the data without using any background knowledge, for e.g. concept hierarchies, similar to *unsupervised learning* approach used in machine learning. We can exploit this result to use in the spatial data mining process as proposed by Ng and Han [41].

Ng and Han do not just pick up the cluster analysis technique from statistics and plug it into the spatial data minig algorithms. There are good reasons for not doing so, the most important being the fact that some researchers have reported its inefficiency and ineffectiveness. It is obvious that for the cluster analysis technique to be effective, there has to be some kind of similarity between the data objects, that is this technique may not produce "good" results when such similarity measure is absent. To counter this problem, Ng and Han use it just for the spatial data where there is natural notion of similarity, like Euclidean or Manhattan distance. As for the efficiency concern, Ng and Han develop a new algorithm for cluster analysis, called CLARANS (Clustering Large Applications based upon RANdomized Search). Based upon experimental evidence, it has been shown that CLARANS outperforms the two best known existing cluster analysis algorithms, PAM (Partitioning Around Medoids) and CLARA (Clustering LARge Applications). Then, they use CLARANS in spatial data mining algorithms, SD(CLARANS) and NSD(CLARANS). First, we will briefly describe the three cluster analysis algorithms.

PAM was developed by Kaufman and Rousseeuw [32]. It is one of the clustering algorithms based on partitioning, as the name suggests. Assuming that there are n objects, PAM finds k clusters by first finding a representative object for each cluster, which is called a *medoid*. That is why it is also known as the $k - medoid$ method. After selecting k *medoids*, the remaining $n - k$ objects are put into a cluster, the medoid of which is the most similar to the object. The measure of similarity or dissimilarity is given as inputs. The overall *quality* of the clustering is measured as the average dissimilarity between an object and the medoid of its cluster. The algorithm repeatedly tries to make a better choice of the medoid and to insert the non selected objects into one of the k clusters. The cost of each iteration is $\mathcal{O}(k(n-k)^2)$. It is therefore computationally inefficient for large values of n and k.

CLARA was proposed by Kaufman and Rousseeuw [32] as well. The difference between PAM and CLARA is that the latter is based upon *sampling*. Only a small portion of the real data is chosen as a representative of the data and *medoids* are chosen from this sample using PAM. The intuition is that if the sample is selected in fairly random manner, then it correctly represents the whole data set and therefore, the representative objects (medoids) chosen, will be similar as if chosen from the whole data set. CLARA draws multiple samples and outputs the best clustering out of these samples. As expected, CLARA can deal with larger data sets than PAM. The complexity of of each iteration now becomes $\mathcal{O}(kS^2 + k(n-k))$, where S is the size of the sample. The authors indicate through their experimental results that 5 samples of size $40 + 2k$ gives "good" results.

It is easy to realize that PAM searches for the best k medoids among a given data set whereas CLARA searches for the best k medoids among the selected sample of the data set. Let us suppose that object O_i is one of the medoids in the best k medoids. Thus, if during sampling O_i is not selected, then CLARA will never find the best clustering. This is exactly the tradeoff for the efficiency. But Ng and Han's proposed CLARANS tries to mix both PAM and CLARA by searching only the subset of the data set and not confine itself to any sample at the same time. While CLARA has a fixed sample at every stage of the search, CLARANS draws a sample with some randomness in each step of the search. The clustering obtained after replacing a single object is called the *neighbour* of the curent clustering. The number of neighbours to be tried is restricted by the parameter *maxneighbour*. Each iteration of the algorithm produces a local optimum. The number of local optima to be searched is also bounded by the parameter *numlocal*. CLARANS has been experimentally shown to be more effective than both PAM and CLARA. The authors claim that the computational complexity of every iteration in CLARANS is basically lineary propertional to the number of objects. This claim has been supported by Ester *et al.* in [13].

Based upon CLARANS, Ng and Han further develope two spatial data mining algorithms, in a fashion similar to the algorithms discussed earlier in this section: *spatial dominant approach*, SD(CLARANS) and *nonspatial dominant approach*, NSD(CLARANS). Both algorithm assumes that the user specifies the type of the rule to be mined and relevant data through a learning request similar to the experimental database mining prototype, DBMINER [26].

Algorithm SD(CLARANS)

In this spatial dominant approach, spatial component(s) of the relevant data items are collected and clustered using CLARANS. It should be mentioned that CLARANS is used to find the most natural number, k_{nat}, of clusters. One may ask, how is k_{nat} determined in the first place. It is indeed a very difficult and open question. The authors however, adopt a heuristic of detemining k_{nat}, which uses *silhoutee coeffiients* [1], introduced by Kaufman and Rousseeuw [32]. Each of the clusters thus obtained is processed by generalizing its nonspatial components using DBLEARN. Note that this algorithm differs from the spatial dominant generalization algorithm (without clustering), in that the latter requires the user to provide the spatial concept hierarchies. however, in this case, it can be said that SD9CLARANS) computes spatial hierarchy dynamically. The hierarchy thus found is more "data oriented" rather than "human orieted".

Algorithm NSD(CLARANS)

This nonspatial dominant approach first applies nonspatial generalizations and spatial clustering afterwards. DBLEARN is used to perform attribute-oriented generalizations of the nonspatial attributes and produce a number of generalized tuples. Then, for each such generallized tuple, all the spatial components are collected and clustered using CLARANS to find k_{nat} clusters. In the final step, the clusters thus obtained are checked to see if they overlap with eachother. If so, then the clusters are merged, and the corresponding generalized tuples are merged as well.

Depending upon the rules or the form of knowledge that user wants to discover, it may be better to choose one or the other of the above two algorithms. If the rules to find are nonspatial characterizations of spatial attributes, then SD(CLARANS) has an edge. This is because NSD(CLARANS) separates the objects into different groups before clustering which may weaken the inter object similarity, or cluster tightness. On the other hand, NSD(CLARANS) is suitable if the spatial clusters within groups of data that has been generalized nonspatially is sought. However, both algorithms arrive at the same result (or rules).

CLARANS on Large Spatial Databases

Ester *et al.* [13] point out some of the drawbacks of the CLARANS clustering algorithm presented by Ng and Han [41]. First of all, it is assumed that the objects to be clustered are all stored in fast memory at the same time. This assumption may not be valid for large databases and disk-based methods are required. Secondly, the efficiency of the algorithm can be improved substantially by modifying the *focusing* component of the algorithm (see architecture in Figure 1).

The first drawback is allevieted by integrating CLARANS with efficient spatial access methods, like R*-tree. R*-tree provides an excellent interface for the queries to the large spatial databases. Most importantly, it supports the focusing techniques that Ester *et al.* propose to reduce the cost of implementing CLARANS. Each node n R*-tree can be thought of as a disk page and *centroid query* is defined which returns the most central object of a disk page. The authors further show that the most computationally expensive step of CLARANS is calculting the total distances between the two clusterings. They propose two approaches to reduce the cost of this step. One is to reduce the number of objects (n) to consider [2], and the other is to restrict the access to certain objects that do not actually contribute to the computation. The authors further give three different focusing technique which tries to exploit the above two approaches: *focus on representative objects*, *focus on relevant clusters*, and *focus on a cluster*.

Ester *et al.* apply the above clustering techniques to a large protein database to find a segmentation of protien surfaces so as to facilitate the so called *docking queries*. The way a protein molecule interacts with other biomolecules is called *docking*. They evaluated CLARANS with and without focusing technique, focusing on representataives, with respect to effectiveness and efficiency. They report that the effectiveness decreased from 1.5% to 3.2% whereas the efficiency increased by factor ranging from 48 to 158. The measure of effectiveness used is the average distance of the resulting clustering whereas the measure of efficiency used is the CPU run time. All the experiments were done varying the parameter *maxneighbour* (cf. CLARANS description above).

[1] It is a propoerty of an object that specifies how much truly does the object belong to the clusterthat it is in

[2] Since the operation is $\mathcal{O}(n)$.

2.3 Methods Exploring Spatial Associations

Most of the knowledge discovery has been focussed into discovery of characteristc rules in spatial databses. All the methods that we disccused previously only find characteristic rules that characterizes, for e.g., a spatial object according to its nonspatial attribute. But in many situations we want to discover spatial assocoation rules, rules that asociate one or more spatial object with other one or more spatial objects. A spatial association rule is of the form $X \rightarrow Y$, where X and Y are sets of spatial or nonspatial predicates. The examples of spatial predicates are *close_to, near_by, adjacent_to* etc. The following rule, for example, is a spatial association rule (confidence of the rule is in brackets):

$$is\_a(x, school) \rightarrow close\_to(x, park) \ (80\%)$$

There are various kinds of spatial predicates that could constitute a spatial association rule. Some examples are: topological relations like *intersects, overlap, disjoint, etc.*; spatial orientations like *left_of, west_of, etc.* We will take a closer look into how these associations can themselves be used to infer some non-trivial knowledge about the spatial objects later in the section (cf. Section 2.3.1). To our knowledge, only one study has been done to explicitly develop methods for spatial association rules in spatial databases, which is Koperski and Han [34]. We will discuss their study first in this section.

Koperski and Han use the concepts of *minimum support* and *minimum confidence* introduced by Agrawal *et al.* [1] in their study to develop association rules from large transaction databases. The intuition behind this is that in large databases, there may exist a large number of associations between objects but most of them will be applicable to only a small number of objects. We would like to filter out those associations using the minimum support and minimum confidence as thresholds. Informally, we can define the support of a pattern A in a spatial database S to be the likelyhood of the occurance of pattern A in S and the confidence of rule $X \rightarrow Y$ to be the likelihood that the pattern Y occurs whenever X does. These thresholds define the *strong* rules. Koperski and Han propose a top-down, progressive deepening search method for mining spatial association rules.

The method proposed first searches for the associations at a high concept level for large patterns using coarse spatial computations. Then, the search deepens for only those large patterns in a more finer resolution, that is, detailed spatial computations. It is an recursive process that terminates when no larger patterns can be found. The concept levels are determined using the provided hierarchies for spatial and nonspatial predicates. The authors thus introduce a novel *two step spatial computation* technique for optimization during the search for associations. At the high concept level, efficient spatial computation algorithms and structures like R-trees and planesweep techniques are used where as more detailed and finer, but more expensive, spatial computations (e.g. multi step joins) are applied at lower concept levels to those patterns that are promising. This saves a lot of computation since there are much fewer spatial associations relationships left at the lower concept levels, after the filtration using minimum support and minimum confidence at higher levels. The thresholds, minimum support and minimum confidence, can be different at each level of description (or filtration) since the same thresholds may not find interesting associations at the lower concept levels [3].

Algorithm for Multiple Level Spatial Association Rules

The first step of the algorithm collects all the task relevant data, $Task\_Relevant\_DB$, from the spatial databases, which is assumed to be an extended relationsl following SAND architecture (see [4]). Then, some efficient spatial computations are performed as mentioned above to extract spatial associations at high level, for example, predicate *gclose_to* (generalized close_to), which could encompass exact spatial predicates like *adjacent_to, intersects etc.* These prediates are stored in an extended relational database $Coarse\_predicate\_DB$. Each predicate in $Coarse\_predicate\_DB$ is cheked with the threshold for the top level to filter out those which are not very promising. This results in a database of large $1 - predicates$ ($Large\_Coarse\_predicate\_DB$), where a single predicate is called *1-predicate*. The conjunction of k such predicates is called $k - predicate$. Spatial association rule at the first level can be generated from this table. This table is further processed using finer spatial computations to produce $Fine\_predicate\_DB$. This table is used to produce large $k - predicates$ and generate association rules at multiple concept levels. At each concept level, start with large $1 - predicate$ and then iteratively

[3] See Han and Fu [27] for detailed discussion on the rationale behind the multiple level thresholds for mining multiple level association rules in large transaction databases.

generate large $k - predicates$ till no larger ones can be found.

The computational complexity of the algorithm is $\mathcal{O}(C_c \times n_c + C_f \times n_f + C_{nonspatial})$ [34], where C_c and C_f are average costs of computing each spatial predicate at a coarse and fine resolution level respectively, n_c is the number of predicates that are coarsely computed, n_f is the number of predicates that are finely computed, and $C_{nonspatial}$ is the total cost of generating rules from the predicate databses. Observe that C_c is much smaller than C_f but n_c is more efficient that n_f.

The above algorithm, especially the *two step computation* technique, is a novel approach towards mining spatial association rules at multiple levels. However, it demands background knowledge in the form of concept hierarchies and expects the user to know exactly the form of the rule s/he wants. It may be a good idea to work towards integrating this technique with clustering methods to avoid the necessity by the user to provide the concept hierarchies for spatial and nonspatial attributes.

2.3.1 Topological Inferencing

In the previous section, we discussed how the topological relations like *intersects, overlaps, etc.* are used as spatial predicates in the method for mining association rules. Spatial reasoning [31] is one whole area of research in itself where researchers are trying to solve problems dealing with objects occupying space, in general. Spatial reasoning is attractive because spatial information which has not been explicitly recorded and which is otherwise not immediately apparent from the raw data can be inferred. Besides its wide applications in Geographic Information Systems (GIS), CAD/CAM etc, we think that it can be used in the association rule mining from spatial databases. For example, it can be used during the coarse spatial computation stage (in Koperski and Han's [34]) to infer the generalized relationships between the spatial objects.

Topological relations are those which are invariant under topological transformations like, rotation, translation etc. Allen's paper [2] give a complete relation between 1-dimensional intervals which is widely used as a basis for extensions to higher dimensions. Researchers like Egenhofer [12] have given a formal model to combine topological knowledge. The author has also derived a complete set of topological relations from the combination of two topological relations between two n-dimensional objects. This can be used to derive the topological relationship $R_z(A, C)$ when the topological relations $R_x(A, B)$ and $R_y(B, C)$ are given. This is a powerfull method which can be used for intelligent retreival of spatial data and to optimize spatial queries [12]. Some researchers have also presented spatial logic that can be used for spatial reasoning, for e.g., Cui *et al* in [11]. This logic can be used to deductively mine knowledge about spatial objects.

Gringi *et al.* [20] take Egenhofer's work [12] further and discuss the computational complexities and technicalities involved in using his spatial relations for topological inferance. Basically, they give different levels of topological resolutions: high resulution being Egenhofer's eight binary relations betwen two planar regions and its coarser situations being lower resolutions. For example, high resolution relations *disjoint* and *meet* is combined to form a medium resolution *disjoint*. This idea is very similar to the one that was used in generalizing spatial relations in the algorithm to find multiple level association rules.

2.4 Using Approximation and Aggregation

We discussed a clustering algorithm CLARANS in Section 2.2. The algorithm is an effective and efficient method of finding *what* the clusters are in the spatial database. But perhaps the more interesting result would be to find out *why* the clusters are there in the first place? Knorr and Ng in [33] present a study motivated by this question. This question can be rephrased as what are the characteristics of the clusters in terms of the features that are close to it? The problem they face is the one of meeasuring aggregate proximity because statements like *90% of the houses ina cluster are close to feature F* is more informative and interesting than statements like *one house is close toa certain feature F.*

One may ask why are not they simply using the n nearest neighbour searches using structures like k-d trees, R-trees and its varants, Voronoi diagrams etc. It turns out that such structures are unable to perform the search needed for their purpose. For example, the distance between the cluster and a feature is measured as the distence between the boundaries, not between the points, like centroids. Furthermore, the cost of building and maintaining the indices for such structures are prohibitive given the fact that such indices may not be used frequently. So, then why not use the map overlay techniques developed in GIS studies? The

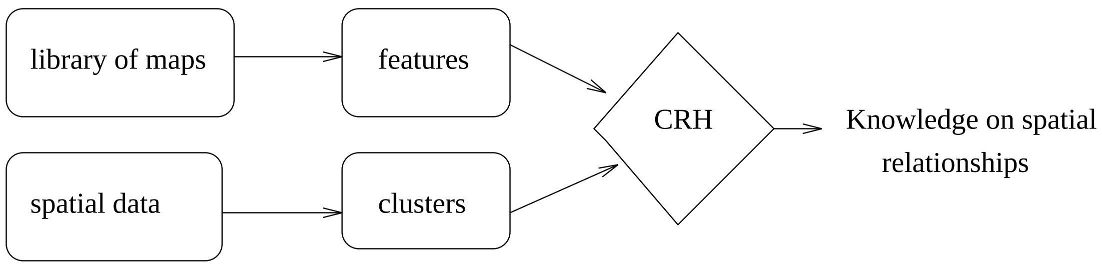

Figure 3: Using CRH for knowledge discovery in spatial databases

problem here is that it can only identify the nearby features but the knowledge about aggregate proximity becomes a problem. Therefore, the authors propose the use of computational geometry concepts [43] to find out the caracteristics of a given cluster in terms of the features close to it. Furthermore, they describe an algorithm CRH (where C is for encompassing circle, R for isothetic rectangle, and H for convex hull [4]) which uses such concepts as filters to reduce the candidate features at multiple levels. In short, they collect a large number of features from multiple maps and feed them along with the cluster to the algorithm CRH and discover knowledge about spatial relationships as shown in Figure 3.

Knorr and Ng evaluate various computational geometry algorithms for distance computation, and shape descriptions and overlap computations. Taking into account the problem of data distribution in a cluster and various size and shapes of the features, the authors choose a technique for computing the distance between a cluster point and feature boundary. As for the shape description and overlap, the authors choose minimum bounding structures (along with convex hulls). They use these to develop a multiple filtering approach, with the filters set up in an increasing order of accuracy but decreasing order of efficiency. That is, filters that are applied earlier are more efficient but more coarser than the later ones.

Algorithm CRH

The algorithm CRH first uses the encompassing circle filter to the large number of features, features that are most promising are then passed to the isothetic rectangle filter. These two filters eliminate a large number of features and only a small number are passed to the final convex hull filter. CRH then calculates aggregate proximity of points in the cluster to the convex boundary of each feature, upon which the features are ranked. The ranking feature can be thought of as a threshold of number of features wanted in the discovered knowledge. Also, each filter has its own threshold, which is the minimum umber of features to pass on to the next filter. When the number of fetures found lie below the threshold, cluster is enlarged to encompass more features to pass the threshold limit. Shape enlargement can be acheived by *linear* policy - enlarge the shape by constant abount, *bisection* policy - perform increment logarithmically. But the problem with these is that a feature may have to be tested for overlap with a cluster many times. *Memoization* can be used to avoid this by storing the distance between each feature and the cluster the first time the intersection test is performed. Depending upon the shapes, cirlces, rectangles or convex hulls, the minimum distance between the circumferences, the boundaries of the rectangles, or the boundaries of the polygons respectively are stored.

The algorithm CRH is experimentally reported to have response time of less than two seconds for processing 50,000 features. It is an impressive result. Furthermore, it is empirically shown to be scalable nd *memoization* policy is found to be the most consistent and efficient of all the shape enlargement policies. Overall, this method seems to be efficient and the expressiveness of the knowledge discovered is non-trivial and interesting. But computational complexity of the algorithm is not clear in the paper. It may be a good idea to do so because it will make it clear whether the algorithm is efficient and scalable as it is claimed to be.

[4]Convex Hull is the minimal, simple closed curve of a set of points such that a line connecting any two points of the set always lies on the interior of the boundary.

2.5 Pattern Recognition Method

Knowledge mining from Image databases can be said to be a special case of spatial data mining. There has been some studies, led by Fayyad *et al.* [14, 45], on the automatic recognition and catagorization of sky objects, like volcanoes, in Venus by processing the pictures taken by Magellan spacecraft. Their method is based upon pattern recognition and machine learning techniques. Magellan transmitted more than 30,000 high resolution images of the surface of Venus from different angles. The authors present a system in [14] for identifying small volcanoes on the surface of Venus.

Their system is composed of three basic components: *data focusing* technique, *feature extraction* and *classifiaction learning*. Like all other data focusing techniques, the first component increases the overall efficiency of the system by first identifying the portion of the image being analysed that is most likely to have a volcano. This is acheived by comparing the intensity of the central pixel of a region to the estimated mean background intensity of its neighbourhood pexels. Then the final task is to discriminate between volcanoes and other objects looking like volcanoes. Such "false alarms" are caused by objects on the surface of Venus causing intensity deviations [7]. The second component extracts interesting features from the data and the final component uses training examples (both positive and negative) provided by the experts to create a classifier that can discriminate between volcanoes and "false alarms".

Smyth *et al.* in [45] discuss issues on knowledge discovery when the absolute ground truth is not available, using the above problem as a case study. The paper's main contribution is the modelling and treatment of subjective label information given by the experts using probabilistic models. The research is important because it concludes that it is possible for the kowledge discovery methods to be modified to handle the lack of absolute ground truths.

2.6 Other Methods

The problem introduced by Fayyad *et al.*'s [14] has been followed up by other researchers as well. One interesting study is done by Bell *et al.* [7] which proposes a general method for knowledge discovery in spatial databases based upon evindence theory [21] taking the above image database mining problem as a case study. In the study the authors describe an extension of general framework for database mining in relational databases based on evidential theory [3] to mine knowledge from spatial databases.

Evidential reasoning is a generalization of conventional probability in the sense that it does not make any assumptions about the independence of data being analysed. Therefore, it may be a better choice to model the data like Venus pictures, where pixels may be interrelated, rather than using probabilistic model like Bayesian methods, for example. Evidential theory provides a method to combine evidences gathered from different sources to produce a single measure of uncertainty. Thus it is claimed to be a better mehod to reason about spatial data in the presence of uncertainty. The combining of evidences is done using a techniqe based upon Dempster-Shafer theory. Informally, this theory can be regarded as a generalization of the conventional probability theory, where the probabilities are fixed and known in advance, to the case where only the upper and lower bounds on probabilities are available [21].

Bell *et al.* [7] give an example of how this method can be applied to image databases. But the study does not include the computational complexity or the broader issues like scalability, effectiveness etc. Therefore, based just upon this study, it is not entirely clear whether such techniques can be used in general for spatial data mining.

3 Future Directions

As we mentioned earlier data mining is a very young field going back not more than year 1990. Spatial data mining is even younger since data mining researchers first concentrated on "attribute" data mining in relational databases. Many spatial data mining methods we saw are actually assuming the presence of extended relational model for spatial databases. But it is widely believed that spatial data are not handled well by relational databases. As advanced database systems, like Object-Oriented (OO), deductive, and active databases develop, methods for spatial data mining should also follow these paradigms.

DATA MINING IN SPATIAL OBJECT-ORIENTED DATABASES: How can OO approach be used to design an Spatial database [46] and how can knowledge mined from these databases? It is an important question since many researchers have pointed out that OO database may be a better choice for handling Spatial data rather than traditional relational or extended relational models. For example, rectangles, polygons etc. can be modelled naturally in OO database. OO database techniques are maturing, OO knowledge representaion techniques have beenproposed (e.g. Mohan and kashyap [40]), and efficient SAM, like R*-trees ca be used to make OO database more efficient in access and retreival. Therefore, exploiting OO technology in data mining is an area with enormous potential. Techniques for generalizations of complex data objects, methods and class hierarchies has been studied by Han *et al.* [28].

MINING UNDER UNCERTAINTY: Use of evidential reasoning should be explored in the mining process for image databases and other databases where uncertainty modelling has to be done. As mentioned in Bell *et al*'s [7], evidential theory can model uncertainty better that traditional probabilistic models, like Bayesian methods. It can be compared with the rough set and fuzzy set approach (Ref ??).

MULTIDIMENSIONAL RULE VISUALIZATION: Discovering knowledge is not enough because it has to presented in a manner that the user can understand easily. One of the most efective way of digesting the rules discovered is through graphical visualizations. Humans are very good at interpreting visual data and scenes. This fact should be exploited in the data mininng process. Multidimensional data visualization has been studied, e.g. Lee *et al.* [35], but multidimensional rule visualization is still an immature area. Spatial data mining can burrow some well-developed visualization techniques from computer graphics in this case.

ALTERNATIVE CLUSTERING TECHNIQUES: More research needs to be done to explore other clustering techniques, for e.g., one proposed by Fisher in [17], to be used in spatial data mining. Fisher proposes a method of generating hierarchical clustering based upon iterative optimization technique. The optimization can be done in the background while not so optimal clustering are being used by the mining system. This may provide greater flexibility than provided by other clustering methods. Improvements on the existing clustering used for spatial data mining like CLARANS may be possible. Heuristics like iterative reapair may help clustering algorithms to become more efficient without sacrificing its effectiveness.

Another clustering related future direction is study of clusterings with overlapping objects. Clusters can maintain additional information about each object it contains, which is its degree of membership. In this way, fuzzy clustering techniqes can be used to accomodate objects having same distance from the medoid, for example.

MINING SPATIAL DATA DEVIATION AND EVOLUTION RULES: One obvious extension to current work in spatial data mining is to study discovery of data devation and evolution rules. These rules may have its use in medical imaging, where one would like to find out how certain features are deviating from the norm or how they are evolving in time. This is just one example, other applications include, discovering and predicting weather patterns of geographic regions, land use planning, etc.

Following were suggested by Han in recent conference tutorial on data mining [25]. We believe that they are very important observations and future directions that should be followed up for the advancement of spatial data mining.

USING MULTIPLE THEMATIC MAPS: We discussed generallization based methods which used a single thematic map during generalizations. Various applications demand that an spatial data mining be conducted using multiple thematic maps. This would involve not only clustering but also spatial computations like map overlay, spatial joins etc. For example to extract general weather patterns, it may be better to use *temperature* and *precipitation* thematic maps and generalization carried out in both.

INTERLEAVED GENERALIZATION: To extend the generalization based methods, it is interesting to consider interleaving spatial and nonspatial generallizations so as to get the results in more efficient manner. Efficiency can be acheived because usually spatial operations, like joins and overlays, are more expensive than nonspatial computations. Thus, by first generalizing the nonspatial component and minimally using spatial generalizations may save a lot of computation time.

GENERALIZATION USING TEMPORAL SPATIAL DATA: This relates to the point we raised on discovery of data evolution rules earlier in this section. This involves generalization over a sequence of data maps collected during different time intervals. Then, comparison or summarization can be done to discover data evolution regularities.

4 Limitations and Conclusions

Despite our effort to do a comprehensive survey, we feel that we may have left out some matters of interest due to various reasons including time constraints. We also mainly focussed on North American published conference reports and journals. We are not sure whether there are promising literature outside the scope of our survey, also because of the fact that spatial data mining field is very young.

We have shown that spatial data mining is a promising field of reesarch with wide applications in GIS, medical imaging, robot motion planning, etc. Although, the field is quite young, a number of algorithms and techniques have been proposed to discover various kinds of knowledge from spatial data. We surveyed existing methods for spatial data mining and mentioned their strengths and weaknesses. This then led us to the future directions and suggestions for the spatial data mining field in general. We beleive that some of the suggestions that we mentioned has already been thought about by researchers and work may have already started on them. But what we hope to acheive is to give reader a unified perspective on the field as a whole.

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining associations rules between sets of items in large databases. *Proceedings to 1993* ACM-SIGMOD, pages 207–216, 1993.

[2] J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26:832–843, 1983.

[3] S. S. Anand, D. A. Bell, and J. G. Hughes. A general framework for database mining based on evidential theory. *Internal Report*, Dept. of Inf. Sys., Univ. of Ulster at Jordanstown, 1993.

[4] W. G. Aref and H. Samet. Extending dbms with spatial operations. In Gunther and Schek [22].

[5] W. G. Aref and H. Samet. Optimization strategies for spatial query processing. *Proceedings of the* 17^{th} *Int'l Conference on Very Large Databases (VLDB'91)*, September 1991.

[6] N. Beckmann, H.-P. Kriegal, R. Schneider, and B. Seeger. The r*-tree: An efficient and robust access method for point and rectangles. *Proceedings to ACM-SIGMOD 90*, 1990.

[7] D. A. Bell, S. S. Anand, and C. M. Shapcott. Database mining in spatial databases. *International Workshop on Spatio-Temporal Databases*, 1994.

[8] T. Brinkhoff, H.-P. Kriegel, and B. Seeger. Efficient processing of spatial joins using r-trees. *Proceedings of SIGMOD-93*, May 1993.

[9] P. A. Burrough. *Principles of Geographic Information Systems for Land Resources Assesment.* Oxford University Press, 1986.

[10] D. K. Y. Chiu, A. K. C Wong, and B. Cheung. A statistical technique for extracting clasificatory knowledge from databases. In Piatetsky-Shapiro and Frawley [42].

[11] Z. Cui, A. G. Cohn, and D. A. Randell. Qualitative and topological relationships in spatial databases. *Proceedings to SSD'93*, 1993.

[12] M. J. Egenhofer. Reasoning about binary topological relations. In Gunther and Schek [22], pages 143–160.

[13] M. Ester, H.-P. Kriegel, and X. Xu. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. *Proceedings of the* 4^{th} *Int'l Symposium on Large Spatial Databases (SSD'95)*, September 1995.

[14] U. M. Fayyad and P. Smyth. Image database exploration: Progress and challenges. *Proceedings of AAAI-93 workshop on KDD*, 1993.

[15] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press, Menlo Park, CA, 1995.

[16] D. Fisher. Improving inference through conceptual clustering. *AAAI-87*, pages 461–465, 1987.

[17] D. Fisher. Iterative optimization and simplification of heirarchical clustering. Technical report cs-95-01, Vandelbilt University, Nashville, TN, 1995.

[18] S. Fotheringham and P. Rogerson. *Spatial analysis and* GIS. Taylor and Francis, 1994.

[19] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. Knowledge discovery in databases: An overview. In Piatetsky-Shapiro and Frawley [42].

[20] M. Gringi, D. Papadias, and C. Papadimitriou. Topological inference. *To appear in IJCAI-95*, 1995.

[21] J. Guan and D. Bell, editors. *Evidence Theory and its Applications, vol. I.* North-Holland, 1991.

[22] O. Gunther and H. J. Schek, editors. *Advances in Spatial Databases.Second Symposium, SSD'91 Proceedings.* Springer-Verlag, 1991.

[23] R. Guttman. A dynamuc index structure for spatial searching. *Proceedings of 1984 SIGMOD*, pages 47–57, 1984.

[24] R. Schneider H.-P. Kriegal, T. Brinkhoff. The combination of spatial access methods and computational geometry in geographic database systems. In Gunther and Schek [22].

[25] J. Han. From database to knowledge-base systems: An evolutionary approach. *Conference Tutorial at ICDE'95*, 1995.

[26] J. Han, Y. Cai, and N. Cercone. Knowledge discovery in databases: An attribute-oriented approach. *Proceedings of 1992 Int'l Conference on Very Large Data Bases (VLDB'92)*, August 1992.

[27] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. *Proceedings of 1995 Int'l Conference on Very Large Data Bases (VLDB'95)*, September 1995.

[28] J. Han, S. Nishio, and H. Kawano. *Knowledge Discovery in Object-Oriented and Active Databases.* In F. Fuchi and T. Yokoi (eds), Knowledge Building and Knowledge Sharing, Ohmsha/IOS Press, 1994.

[29] M. Holsheimer and M. Kersten. Architectural support for data mining. *Technical Report CS-R9429, CWI*, 1994.

[30] M. Holsheimer and A. Siebes. Data mining: The search for knowledge in databases. *Report CS-R9406*, CWI, Amsterdam, The Netherlands, 1994.

[31] A. Kak. Spatial reasoning. *AI Magazine*, 9(2), 1988.

[32] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis.* John Wiley & Sons, 1990.

[33] Ed Knorr and Raymond T. Ng. Applying computational geometry concepts to discovering spatial aggregate proximity relationships. *Submitted to ICDE'96 for Review*, 1995.

[34] K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. *Proceedings of* 4^{th} *Int'l Symposium on Large Spatial Databases (SSD95)*, August 1995.

[35] H.-Y. Lee, H.-L. Ong, and K. S. Sodhi. Visual data exploration. *Unpublished Report, Information Technology Institute, Singapore*, 1995.

[36] W. Lu, J. Han, and B. C. Ooi. Discovery of general knowledge in large spatial databases. *Proceedings of 1993 Far East Workshop on Geographic Information Systems (FEGIS'93)*, June 1993.

[37] C. J. Matheus, P. K. Chan, and G. Piatetsky-Shapiro. Systerms for knowledge discovery in databases. *IEEE TKDE'93*, 1993.

[38] R. S. Michalski, J. M. Carbonnel, and T. M. Mitchell, editors. *Machine Learning: An Artificial Intelligence Approach.* Morgan Kaufmann, Los Altos, CA, 1983.

[39] T. M. Mitchell. Generalization as search. *Artificial Intelligence*, 18:203–226, 1982.

[40] L. Mohan and R. L. Kashyap. An object-oriented knowledge representation for spatial information. *IEEE Transactions on Software Engineering*, pages 675–681, 1988.

[41] R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. *Proceedings of 1994 Int'l Conference on Very Large Data Bases (VLDB'94)*, September 1994.

[42] G. Piatetsky-Shapiro and W. J. Frawley, editors. *Knowledge Discovery in Databases.* AAAI/MIT Press, Menlo Park, CA, 1991.

[43] F. Preparata and M. Shamos, editors. *Computational Geometry.* Springer-Verlag, New York, 1985.

[44] H. Samet, editor. *The Design and Analysis of Spatial Data Structures.* Addison-Wesley, 1990.

[45] P. Smyth, M. C. Burl, U. M. Fayyad, and P. Perona. Knowledge discovery in large image databases: Dealing with uncertainties in ground truth. *Proceedings of AAAI-94 workshop on KDD*, 1994.

[46] P. van Oosterom and J. van den Bos. An object-oriented approach to the design of geographic information system. *In proceedings to SSD'89*, 1989.