ISSN:0975-9646
Manjula Aakunuri et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (4) , 2011, 1501-1504

Spatial Data Mining: A Recent Survey and New Discussions

Manjula Aakunuri[#], Dr.G.Narasimha[*], Sudhakar Katherapaka[$]

[#] *Department of Computer Science & Engineering, Jyothishmathi Institute of Tech & Sciences, Karimnagar, JNTUH,Hyderabad, AP, INDIA*

[*] *Department of Computer Science & Information Technology, JNTU College of Engineering, Kondagattu,Karimnagar, JNTUH ,Hyderabad, AP, INDIA*

[$] *Department of Computer Science & Engineering, Vivekananda Institute of Tech & Sciences, Karimnagar, JNTUH Hyderabad, AP, INDIA*

Abstract **–The main objective of the spatial data mining is to discover hidden complex knowledge from spatial and not spatial data despite of their huge amount and the complexity of spatial relationships computing. However, the spatial data mining methods are still an extension of those used in conventional data mining. Spatial data is a highly demanding field because huge amounts of spatial data have been collected in various applications, ranging from remote sensing, to geographical information systems (GIS), computer cartography, environmental assessment and planning, etc. The collected data far exceeded human's ability to analyze. Recent studies on data mining have extended the scope of data mining from relational and transactional databases to spatial databases. This paper summarizes recent works on spatial data mining, from spatial data generalization, to spatial data clustering, mining spatial association rules, etc. It shows that spatial data mining is a promising field, with fruitful research results and many challenging issues. The main aim of this paper shows the existing methods of clustering and association rules based on spatial data, i.e collected from large amount of spatial data bases.**

Keywords: **spatial data mining, clustering algorithms, knowledge discovery.**

I. SPATIAL DATA MINING

Spatial data mining is the application of data mining techniques to spatial data. Data mining in general is the search for hidden patterns that may exist in large databases. Spatial data mining is the discovery of interesting the relationship and characteristics that may exist implicitly in spatial databases. Because of the huge amounts (usually, tera-bytes) of spatial data that may be obtained from satellite images, medical equipments, video cameras, etc. It is costly and often unrealistic for users to examine spatial data in detail. Spatial data mining aims to automate such a knowledge discovery process. Thus it plays on important role in

1. Extracting interesting spatial patterns and features
2. Capturing intrinsic relationships between spatial and non spatial data
3. Presenting data regularity concisely and at higher conceptual levels and
4. Helping to reorganize spatial databases to accommodate data semantics, as well as to achieve better performance.

Spatial database stores a large amount of space related data, such as maps, preprocessed remote sensing or medical imaging data and VLSI chip layout data. Spatial databases have many features distinguishing them from relational databases. They carry topological and/or distance information, usually organized by sophisticated, multi dimensional spatial indexing structures that are accessed by spatial data access
methods and often require spatial reasoning, geometric computation, and spatial knowledge representation techniques.

A. Spatial Data Mining Structure:

The spatial data mining can be used to understand spatial data, discover the relation between space and the non space data, set up the spatial knowledge base, excel the query, reorganize spatial database and obtain concise total characteristic etc.. The system structure of the spatial data mining can be divided into three layer structures mostly, such as the Figure 1 show [2].The customer interface layer is mainly used for input and output, the miner layer is mainly used to manage data, select algorithm and storage the mined knowledge, the data source layer, which mainly includes the spatial database (camalig) and other related data and knowledge bases, is original data of the spatial data mining.

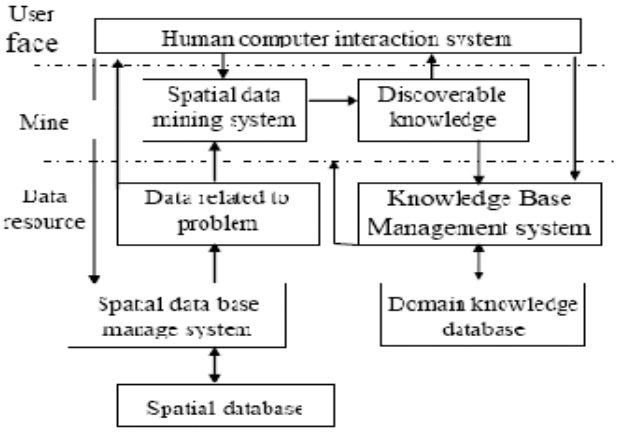

Fig.1 The systematic structure of spatial data mining

B. Primitives of Spatial Data Mining

1) *Rules:* There are several kinds of rules can be discovered from databases in general. For example characteristic rules, discriminant rules, association rules, or deviation and evalution rules can be mined [2]. A ***Spatial characteristic rule*** is a general description of the spatial data**.**

For example, a rule describing the general price range of houses in various geographic regions in a city is a spatial characteristic rule. A *discriminant rule* is general description of the features discriminating or contrasting a class of spatial data from other class(es) like the comparison of price ranges of houses in different geographical regions. A *spatial association rule* is a rule which describes the implication of one a set of features by another set of features in spatial databases. For example, a rule associating the price range of the houses with nearby spatial features, like beaches, is a spatial association rule.

2) *Thematic Maps:*Thematic map is map primarily design to show a theme, a single spatial distribution or a pattern, using a specific map type. These maps show the distribution of features over limited geography areas [2]. Each map defines a partitioning of the area into a set of closed and disjoint regions; each includes all the points with the same feature value. Thematic maps present the spatial distribution of a single or a few attributes. This differs from general or reference maps where the main objective is to present the position of the object in relation to other spatial objects. Thematic maps may be used for discovering different rules. For example, we may want to look at temperature thematic map while analyzing 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 pixels associated with the attribute values. For example, a map may have the altitude of the spatial objects coded as the intensity of the pixel (or the color). In the *vector representation*, a spatial object is represented by its geometry, most commonly being the boundary representation along with the thematic attributes. For example, a park may be represented by the boundary points and corresponding elevation values.

II. ALGORITHMS FOR SPATIAL DATA MINING IN KNOWLEDGE DISCOVERY

The algorithms for spatial data mining include generalization-based methods for mining spatial characteristics and discriminant rules [4, 5, 6], two-step spatial computation technique for mining spatial association rules [8], aggregate proximity technique for finding haracteristics of spatial clusters [7], etc. In the following sections, we categorize and describe a number of these algorithms.

A. Generalization-Based Knowledge Discovery

Generalization based mining is the concept of data from more than a few evidences from a concept level to its higher concept level and performing knowledge withdrawal on the widespread data (Mitchell, 1982).It assumes the survival of background knowledge in the form of concept hierarchies, which is either data-driven or assigned clearly by expert-knowledge [2]. The data can be articulated in the form of a generalized relation or data-cube on which many other operations can be performed to change generalized data into different forms of knowledge. A few of the multivariate statistical or arithmetic techniques such as principal components analysis, discriminant analysis, characteristic analysis, correlation analysis, factor analysis and cluster analysis are used for generalization based knowledge discovery (Shaw and Wheeler,1994).

The generalization-based knowledge discovery requires the existence of background knowledge in the form of concept hierarchies. Issues on generalization-based data mining in object-oriented databases are investigated in three aspects: (1) generalization of complex objects, (2) class-based generalization, and (3) extraction of different kinds of rules. An object cube model is proposed for class-based generalization, on-Line analytical processing, and Data Mining.

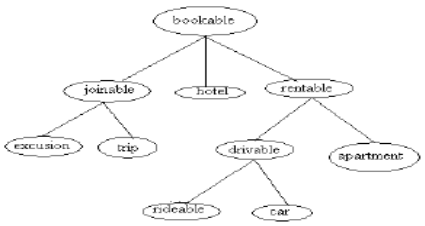

Fig.2 Example of Tourism use concept hierarchy

In the case of spatial database, there can be two kinds of concept hierarchies, non-spatial and spatial. Concept hierarchies can be explicitly given by the experts, or data analysis [7]. An example of a concept hierarchy for tourism 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 ridable and car can be generalized to concept driveable which in turn can be generalized to concept rentable, which also includes apartment. A similar hierarchy may exist for spatial data. For example, in generalization process, regions representing countries can be merged to provinces and provinces can be merged to larger regions. Attribute oriented induction is performed by climbing the generalization hierarchies and summarizing the general relationships between spatial and non-spatial data by (a) climbing the concept hierarchy when attribute values in a tuple are changed to the generalized values, (b) removing attributes when further generalization is impossible and (c) merging identical tuples. Lu et al. [35] presented two generalization based algorithm: *spatial-data-dominant and* non-*spatial-data-dominant generalizations*. Both algorithms assume that the rules to be mined are general data characteristics and that the discovery pricess is initiated by the user who provides a learning request (query) explicitly, in syntax similar to SQL. We will briefly describe both algorithms as follows:

1) *Spatial-Data-Dominant Generalization:* In the first step all the data described in the query are collected. Given the spatial data hierarchy, generalization can be performed first on the spatial data by merging the concept hierarchy. Generalization of the spatial objects continues until the spatial generalization threshold is reached. The spatial generalization threshold is reached when the number of regions is no greater than the threshold value. After the spatial-oriented induction process, non-spatial data are retrieved and analyzed -for each of the spatial objects using the attribute-oriented induction technique as described. The computational complexity of the algorithm is *O(NlogN)*, where N is the number of spatial objects.

2) *Non-Spatial-Data-Dominant Generalization:* In the second step the algorithm performs attribute-oriented induction on the non-spatial attributes, generalizing them to a higher (more general) concept level. For example, the precipitation value in the range (10 in, 15in) can be generalized to the concept wet. The generalization threshold is used to determine whether to continue or stop the generalization process. The third and the last step of the algorithm, neighboring areas with the same generalized attributes are merged together based on the spatial function adjacent *to*. For example, if in one area the precipitation value was 17 in., and in neighboring area it was 18 in. Both precipitation values are generalized to the concept very wet and both areas are merged.

III. CLUSTERING TECHNIQUES

Cluster analysis is a branch of statistics that has been studied extensively 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, like concept hierarchy. A similar approach in machine learning is known as *unsupervised* learning. Clustering algorithms used in statistics, like PAM or CLARA [1], are reported to be inefficient from the computational complexity point of view. As per the efficiency concern, a new algorithm called CLARANS (Clustering large Applications based upon RANdomized Search), was developed for cluster analysis. Experimental evidence showed that CLARANS outperforms the two existing cluster analysis algorithms, PAM (Partitioning Around Medoids) and CLARA (Clustering LARge Applications).

A. PAM (Partitioning Around Medoids)

Assuming that there are *n* objects, PAM finds *k* clusters by first finding a representative object for each cluster. Such a representative, which is the most centrally located point in a cluster, is called a *medoid*[1]. After selecting *k* medoids, the algorithm repeatedly tries to make a better choice of medoids analyzing all possible pairs of objects such that one object is a medoid and other is not. The measure of clustering quality is calculated for each such combination. The best choice of points in one iteration is chosen as the medoids for the next iteration. The cost of a single iteration is $O(k(n-k)^2)$. It is therefore computationally quite inefficient for large values of *n* and *k*.

B. CLARA (Clustering LARge Applications)

The difference between the PAM and CLARA algorithms is that the later one 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[1]. The idea is that if the sample is selected in a fairly random manner, then is correctly represents the whole dataset and therefore, the representative objects (medoids) chosen will be similar as if chosen from the whole dataset. CLARA draws multiple samples and outputs the best clustering out of these samples. CLARA can deal with larger dataset then PAM. The complexity of each iteration now becomes $O(kS^2+k(n-k))$, where S is the size of the sample.

C. CLARANS (Clustering large Applications based upon RANdomized Search)

CLARANS algorithm mix both PAM and CLARA by searching only the subset of the dataset and it does not confine itself to any sample at any given time [1]. 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 process can be presented as searching a graph where every node is a potential solution, i.e, a set of k medoids. The clustering obtained after replacing a single medoids is called the *neighbor* of the current clustering. The number of *neighbors* to be randomly tried is restricted by the parameter *maxneighbor*. If a better *neighbor* is found CLARANS moves to the neighbor's node and the process is started again, otherwise the current clustering produces a local *optimum*. If the local optimum is found CLARANS starts with new randomly selected node in search for a new local optimum. The number of local optima to be searched is also bounded by the parameter *numlocal.* CLARANS also enables the detection of outliers, e.g.. points that do not belong to any cluster.

Based upon CLARANS, two spatial data mining algorithms were developed: *Spatial dominant approach*, SD(CLARANS) and *non-spatial dominant approach*, NSD(CLARANS).

IV. OBJECT ORIENTED DATABASE

Object oriented databases are based on the object oriented programming paradigm, each entity is considered as an object. Data and code relating to an object are encapsulated into a single unit. Each object has associated with it the following:

a. Set of variables that describes the object.
b. Set of messages that the object can use to communicate with other objects, or with the rest of the database system.
c. Set of methods, where each method holds the code to implement a message.

Objects that share a common set of properties grouped into an object class. Each object is an instance of its class. Object classes can be organized into class/subclass hierarchies so that each class represents properties that are common to objects in that class.

Spatial data mining is even younger since data mining researches first concentrated on data mining in relational databases. Many spatial data mining methods we analyzed actually assume the presence of extended relational model for spatial databases but it widely believed that spatial data or not handled well by relational databases. As advanced database systems, like object oriented, deductive and active databases are being developed.

V. PARALLEL DATA MINING

In recent years, there is an increasing interest in the research of parallel data mining algorithms. In parallel environment, by exploiting the vast aggregate main memory and processing power of parallel processors, parallel algorithms can have both the execution time and memory requirement issues well addressed. However, it is not trivial to parallelize existing algorithms to achieve good performance as well as scalability to massive data sets. First, it is crucial to design a good data organization and

decomposition strategy so that workload can be evenly partitioned among all processes with minimal data dependence across them. Second, minimizing synchronization and/or communication overhead is important in order for the parallel algorithm to scale well as the number of processes increases. Workload balancing also needs to be carefully designed. Last, disk I/O cost must be minimized.

VI. FUTURE DISCUSSIONS

Data mining in Spatial Object Oriented Databases: How can the object oriented approach be used to design a spatial database and how can knowledge be mined these databases? It is an important question since many researchers have pointed out that Object Oriented Database may be a better choice for handling spatial data rather than traditional relational or extended relational models. For example, rectangles, polygons, and more complex spatial objects can be modeled naturally in object oriented database.

Parallel Data mining: Due to the high volume of spatial data used during the computations mining using parallel machines or distributed farms of workstations can accelerate significantly the work. We expect that parallel knowledge discovery will be a growing research issue in both relational and spatial data mining.

Alternative Clustering Techniques: Another interesting feature direction is the clustering's of possible overlapping object like polygons as oppose to the clustering of points. Clusters can also maintain additional information about each object they contain, which can be the degree of membership. In this way, fuzzy clustering techniques can be used to accommodate object having the same distance from the medoid.

VII. CONCLUSION

We have shown that spatial data mining is a promising field of research with wide applications in GIS, medical imaging, remote sensing, robot motion planning, and so on. Although, the spatial data miming ground is pretty young, a number of algorithms and techniques have been planned and 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 led as to future directions and suggestions for the spatial data mining field in general. We believe that some of the suggestions that we mentioned have already been thought about by researchers and work may have already started on them. But what we hope to achieve is to give the reader a general perspective of the field

REFERENCES

[1] Krzysztof Koperski.; Junas Adhikary.; and Jiawei Han. Spatial Data Mining: Progress and Challenges Survey Paper, School of Computer Science Simon Fraser University Burnaby, B.C.Canada V5A IS6.

[2] M.Hemalatha.M; Naga Saranya.N. A Recent Survey on Knowledge Discovery in Spatial Data Mining, IJCI International Journal of Computer Science, Vol 8, Issue 3, No.2, may,2011.

[3] Jianwei Li;, Ying Liu;, Wei-keng Liao;, Alok Choudhary. Parallel Data Mining Algorithms for Association Rules and Clustering.

[4] Matheus C.J.; Chan P.K.; and Piatetsky-Shapiro G.1993. Systems for Knowledge Discovery in Databeses,IEEE Transactions on Knowledge and Data Engineering 5(6):903-913.

[5] R Agrawal and R Srikant. Fast Algorithms for Mining Association Rules. In Proc. Of Very Large Databases, may 1994.

[6] J. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[6] S.Shekhar and S.Chawla. Spatial Databases: A Tour. Pretice Hall (ISBN 0-7484-0064-6), 2003.

[7] R. Ng and J. Han. (1994) Effective and Efficient Clustering Methods for Spatial Data Mining, Technical Report 94-13, University of British Columbia.

[8] H. Samet. (1990) The Design and Analysis of Spatial Data Structures, Addison-Wesley.