www.ijecs.in
International Journal Of Engineering And Computer Science
Volume1 Issue 1 Oct 2012 Page No. 37-41

A Study On Spatial Data Clustering Algorithms In Data Mining

Sundararajan S, Dr.Karthikeyan S

Associate Professor and Head, Department of MCA, SNS College of Technology, Coimbatore, Tamil Nadu, India.
E-Mail: sundar_mtp@yahoo.co.in
The Director, Department of Computer Applications, Karpagam University, Coimbatore, Tamil Nadu
E-Mail: skaarthi@gmail.com

Abstract

The enormous amount of hidden data in large databases has produced incredible interests in the area of data mining. Clustering is an indispensable task in data mining to cluster data into significant subsets to obtain useful information. Clustering spatial data is a significant issue that has been broadly investigated to discover hidden patterns or useful sub-groups and has several applications like satellite imagery, geographic information systems, medical image analysis, etc. The spatial clustering approach is supposed to satisfy the necessities of the application for which the investigation is carried out. In addition the same clustering approach should be very efficient in processing data along with noise and outliers, since they are inherently present in spatial data. In recent times, several commercial data mining clustering approaches have been developed and their practice is increasing enormously to realize desired objective. Researchers are attempting their best efforts to accomplish the fast and effective algorithm for the abstraction of spatial data, which are clearly discussed in the literature. Each individual clustering approach has its own merits and demerits for processing multidimensional data and consequently in spatial clustering.

Keywords---, Clustering Algorithms, Knowledge Discovery in Databases (KDD), Spatial Data Mining.

I. INTRODUCTION

Clustering is the process of generating a group of objects based on certain similarity between the available data and is normally applied to huge datasets. In case of spatial data sets, clustering allows a generalization of the spatial constituent which permits for successful data mining. There are a several approaches developed for carrying out the process of clustering [1]. At the same time, there are three main categories of clustering approaches, which are partitional clustering, hierarchical clustering and locality-based clustering. Partitional clustering builds up a separation of the data in such a way that the objects inside a cluster are more comparable to each other than they are to objects in other clusters. The K-Means and K-Medoid approaches are certain of partitional clustering [2].

Hierarchical clustering carries out a series of partitioning operations. These can be completed bottom-up, by executing repeated combination of groups of data until certain pre-determined threshold is attained, or top-down, recursively separating the data until certain pre-determined threshold is attained. Hierarchical clustering is typically used in document and text analysis. Grid-based clustering is also hierarchical. Locality-based clustering approaches group objects depending on local relationships, and as a result the complete database can be scanned at a single pass. A few locality dependent approaches are density based, at the same time as others presume a random distribution.

Even though there are similarities among spatial and non-spatial clustering, bulky databases and spatial databases in specific, have distinctive requirements that generate special considerations for clustering algorithms.

- An obvious need considering the huge quantity of data to be managed is that algorithms be effective and scalable.
- Algorithms are required to be capable of recognizing irregular shapes, together with those with lacunae or concave sections and nested shapes.
- The clustering process must be insensitive to bulky amounts of noise.
- Accordingly, algorithms should not be responsive to the order of input. Specifically,

clustering outcome should be independent the order of data.

- No previous knowledge of the data or the amount of clusters to be generated should be required, and as a result no domain knowledge input should be required for the user.
- Algorithms must manage data with large numbers of features, that is, higher dimensionality [3].

Spatial data illustrates information associated with the space engaged by objects. The data consists of geometric information and can be either distinct or continuous. Discrete data possibly will be a single point in multi-dimensional space; on the other hand discrete spatial data is different from non-spatial data in that it has a distance feature that is used to locate the data in space. Continuous data spans a region of space. This data may perhaps include medical images, map regions, or star fields [4]. Spatial databases are database systems that handle spatial data. It is intended to manage both spatial information and the non-spatial characteristics of that data. With the purpose of providing improved and effective access to spatial data it is essential to develop indices. These indices are most useful when based on multi-dimensional trees. The structures for these indices comprise quad trees, k-d trees, R trees and R* trees. Data mining, or knowledge discovery in databases (KDD), is the method of investigating data to determine previously unidentified potential information. The objective is to show the regularities and relationships that are non-trivial. This is possible through an examination of the patterns that form in the data. Several algorithms have been developed by many researchers to carry out this spatial data mining, but the majority of these approaches are not scalable to very huge databases.

On the other hand, the very large size of spatial databases also needs added approaches for manipulating and cleaning the data with the purpose of preparing it for analysis. This paper describes various spatial clustering algorithms with its merits and drawbacks. This will help the researchers to develop a better clustering algorithm to cluster the spatial data.

II. REVIEW OF LITERATURE

Clustering spatial data is a distinguished difficulty that has been investigated extensively. Grouping similar data in large two-dimensional spaces to discover hidden patterns or meaningful sub-groups has several applications like satellite imagery, marketing, geographic information systems, medical image analysis, computer visions, etc. Even though several techniques have been proposed, only a few techniques have taken physical obstacles into consideration that may possibly have considerable consequences on the efficiency of the clustering. Considering these constraints throughout the clustering process is costly and the representation of the constraints is paramount for better performance. Zaiane et al., [5] analyzed the difficulty of clustering in the existence of constraints for instance physical obstacles and establish a new approach to represent these constraints by means of polygons. This author also developed an approach to prune the search space and decrease the number of polygons to experiment during clustering. This author devised a density-dependent clustering approach, DBCluC, which exploits the advantage of this constraint modeling to effectively cluster data objects at the same time considering all physical constraints. The algorithm can identify clusters of random shape and is not sensitive to noise, the input order and the complexity of constraints. The average running complexity is $O(NlogN)$ where N represents the number of data points.

A Particle Swarm Optimization (PSO) technique is attempted for providing solution for Spatial Clustering with Obstacles Constraints (SCOC). In this approach, Xueping Zhang et al., [6] initially used PSO to get obstructed distance, and then developed the PSO K-Medoids SCOC (PKSCOC) in order to cluster spatial data with obstacles constraints.

Spatial Clustering with Obstacles Constraints (SCOC) has turned out to be a new topic in Spatial Data Mining (SDM). Xueping Zhang et al., [7] developed an Improved Ant Colony Optimization (IACO) and Hybrid Particle Swarm Optimization (HPSO) technique for SCOC. In this approach, the author first exploited IACO to acquire the shortest obstructed distance, which is an efficient technique for random shape obstacles, and then the author developed a novel HPKSCOC dependent on HPSO

and K-Medoids to cluster spatial data with obstacles, which not only provide concentration on higher local constringency speed and stronger global optimum search, however also get down to the obstacles constraints.

The spatial join is a task that integrates two sets of spatial data on the basis of their spatial associations. The cost of spatial join possibly will be extremely huge because of the bigger sizes of spatial objects and the computation-intensive spatial operations. In spatial join processing, a general technique to diminish the I/O cost is to partition the spatial objects into clusters and then allot the processing of the clusters in such a way that the number of times the similar objects to be loaded into memory can be considerably reduced. Jitian Xiao et al., [8] proposed a match-dependent method to partition a huge spatial data set into clusters, which is generated depending on the maximal match on the spatial join graph.

Clustering is a discovery process in data mining and can be utilized to group together the objects of a database into significant subclasses which serve as the basis for other data analysis approaches. The main concentration is on managing with a collection of spatial data. In case of the spatial data, the clustering problem become that of discovering the thickly populated regions of the space and consequently clustering these regions into clusters in such a way that the intracluster similarity is increased and the intercluster resemblance is reduced. Jong-Sheng Cherng et al., [9] developed an innovative hierarchical clustering algorithm that utilizes a hypergraph to denote a set of spatial data. This hypergraph is primarily build from the Delaunay triangulation graph of the data set and can properly obtain the associations between sets of data points. Two stages are in this clustering approach to discover the clusters in the data set.

With the continuous increasing of geo-spatial data scale, the mounting complication of spatial operation, conventional single GIS mode can no longer satisfy new requirements of the bulky geo-spatial data operation. To increase parallel equipment computational capability is the present hot research area. The parallel task division and geo-spatial data partition are put on the strategy study as the precondition of GIS additional performance. Zhanlong Chen et al., [10] is facing to the better performance parallel GIS operation requirements and developed a spatial data partitioning approach depending on the minimum distance clustering, understanding load balance when partitioning spatial data. Developing a new approach to fix the clustering centers depending on K-Means approach, the centers arranged based on the ascending x coordinate sort order and distributed smoothly in the space.

Lin Xiao-ping et al., [11] proposed a spatial clustering approach using field model to cluster data. Spatial clustering is typically used in Spatial Data Mining. Spatial clustering techniques in currently have noticeable limitations in solving spatial clustering difficulties. On the base of evaluating and investigating the potential of each spatial clustering approach, a new approach based on field model is developed based on the individuality of spatial clustering (difficulty in cluster shape and multi-scale-property in cluster structure), in order to provide solution for all kinds of problems of current spatial clustering research which clusters data with object model and represents continuous space with discrete entities, and enabling it be modified to representation of difficult spatial clustering structure.

Since the current world is moving along with budding technologies, the accumulation of data will no longer be in the conventional single dimensional raw data. The relational data can be additional recognized with spatial and non-spatial association. This is to observe whether clustering technique has any role to take part in spatial data mining. Thirumurugan et al., [12] discussed spatial clustering based on statistical method of analysis for determining the knowledge which is encapsulated in the spatial database. This study shows the importance of the spatial clustering approach accomplished through approaches for instance, PAM and CLARA and permits to come across the restrictions of PAM technique.

GML is an application of XML in geographic information system, utilized to accumulate spatial data. Genlin Ji et al., [13] developed an approach for SCTR-GML for spatial clustering in GML data. Evaluated against spatial clustering approaches, SCTR-GML clusters spatial objects depending on the spatial topological associations, at the same time as the approaches like DBSCAN simply group the

spatial objects that are close to each other into one cluster. The SCTR-GML approach initially generates indices for the spatial objects provided by GML, evaluates all the topological associations comprising contain, intersect, adjacent and also developed a novel technique to determine the resemblance among spatial objects which requires the communication of users. The objects in one cluster possibly will not be close to each other, but they have resemblance in the spatial topological associations.

The I/O cost of spatial join processing may well be extremely high because of the huge sizes of spatial objects and the large quantity of spatial objects involved. Spatial joins are typically carried out by the filter-and-refinement approach. Even though there exists several approaches for realizing the filter step of the join processing for huge spatial data sets, not several research has been carried out to enhance the performance of the refinement phase. By grouping the output of the filter phase, the total number of times can be reduced that spatial objects are constantly loaded throughout the refinement phase, as a result to reduce the I/O cost of the refinement phase. A multilevel data partitioning technique is proposed by Xiao, J et al., [14] to partition objects into clusters for spatial join processing. At any time the number of objects is higher than a fixed threshold, consider the hundred, the objects will be clustered through a multilevel scheme, i.e., initially coarsening, subsequently partitioning and at last uncoarsening back to the original object sets, which can be additionally partitioned with the use of the known partitioning techniques.

The attributes of the nearest spatial data objects are constantly related or connected to each other, Yaqin Wang et al., [15] concentrated the problem of determining spatial associations in spatial data through the recognition of clusters depending on spatial neighborhood associations. The author presented an effective clustering approach called SPANBRE that produces high quality clusters in $O(nlogn)$ time and in $O(n^2)$ message complication. SPANBRE is type of agglomerative hierarchical technique. By exploiting the sequence data structure, SPANBRE ignores the complicated spatial join operation. SPANBRE also carry out an optimization strategy for clustering splitting and merging to accomplish high clustering quality.

III. PROBLEMS AND DEFINITIONS

In general, data modeling lay clustering in a historical perception rooted in mathematics, statistics, and numerical analysis. The constraints which denote the clusters are either computed on the basis of quality criterion or cost functions or, on the other hand, they are obtained by local search algorithms which are not essentially following the gradient of a global quality criterion.

The limitation met by the clustering method is mainly because of its deficiency in predicting the similarity when the spatial dataset is used for clustering. Also, the interdistance of the clusters needs to be increased and the intradistance for the cluster needs to be decreased.

To overcome these difficulties and satisfying those needs, a new clustering algorithm should be developed which focuses on reducing the dimensions of original dataset. This can be achieved by gathering only the important components from the whole dataset which can be performed by means of using techniques like Principle Component Analysis, etc.

IV. CONCLUSION

Spatial data mining is the finding of useful associations and characteristics that may well exist implicitly in spatial databases. Spatial data mining concentrates on automating such a knowledge discovery process. It plays an essential role in (i) obtaining interesting spatial patterns and characteristics; (ii) capturing inherent associations among spatial and non-spatial data; (iii) presenting data reliability concisely and at conceptual levels; and (iv) assisting in reorganizing spatial databases to accommodate data semantics, in addition to accomplish enhanced performance. Spatial data clustering is a most important constituent of spatial data mining and is implemented as such to retrieve a pattern from the data objects distribution in a particular data set and as mentioned earlier it has several applications like satellite imagery, geographic information systems, medical image analysis, etc.. This paper presents various techniques available for spatial data clustering. This will assist

the researchers to come up with new techniques to cluster the spatial data effectively.

REFERENCES

[1] Ester, Martin, Alexander Frommelt, Hans-Peter Kriegel, and Jörg Sander, "Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support", Data Mining and Knowledge Discovery, Vol. 4, Pp. 193-216, 2000.

[2] Ester, Martin, Hans-Peter Kriegel and Jörg Sander, "Spatial Data Mining: A Database Approach", Proceedings of the 5th International Symposium on Spatial Databases (SSD 97), Berlin, Germany, 1997.

[3] Agrawal, Rakesh, Johannes Gehrke, Dimitrios Gunopulos and Prahhakar Raghavan, "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications", Proceedings of the ACM-SIGMOD International Conference on Management of Data, Seattle, Washington, 1998.

[4] Samet, Hanan, "Spatial Data Models and Query Processing", In Modern Database Systems: The Object Model, Interoperability and Beyond, Addison Wesley/ACM Press, Reading, MA, 1994.

[5] Zaiane, O.R.; Chi-Hoon Lee, "Clustering spatial data in the presence of obstacles: a density-based approach", Proceedings. International Database Engineering and Applications Symposium, Pp. 214 – 223, 2002.

[6] Xueping Zhang; Fen Qin; Jiayao Wang; Yongheng Fu; Jinghui Chen, "Clustering Spatial Data with Obstacles Constraints by PSO", Fourth International Conference on Fuzzy Systems and Knowledge Discovery, 2007.

[7] Xueping Zhang; Qingzhou Zhang; Zhongshan Fan; Gaofeng Deng; and Chuang Zhang; "Clustering Spatial Data with Obstacles Using Improved Ant Colony Optimization and Hybrid Particle Swarm Optimization", Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008.

[8] Jitian Xiao, "Clustering Spatial Data for Join Operations Using Match-based Partition", International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, Pp. 471 - 476 , 2005.

[9] Jong-Sheng Cherng; Mei-Jung Lo, "A hypergraph based clustering algorithm for spatial data sets", Proceedings IEEE International Conference on Data Mining, Pp. 83 – 90, 2001.

[10] Zhanlong Chen; Liang Wu; Dingwen Zhang, "Spatial data partitioning based on the clustering of minimum distance criterion", International Conference on Computer Science and Service System (CSSS), Pp. 2583 – 2586, 2011.

[11] Lin Xiao-ping; Mao Zheng-yuan; Liu Jian-hua, "A Spatial Clustering Method by Means of Field Model to Organize Data", Second WRI Global Congress on Intelligent Systems (GCIS), Pp. 129 – 131, 2010.

[12] Thirumurugan, S.; Suresh, L., "Statistical spatial clustering using spatial data mining", IET International Conference on Wireless, Mobile and Multimedia Networks, 26 – 29, 2008.

[13] Genlin Ji; Jianxin Miao; Peiming Bao, "A Spatial Clustering Algorithm Based on Spatial Topological Relations for GML Data", International Conference on Artificial Intelligence and Computational Intelligence, Pp. 298 – 301, 2009.

[14] Xiao, J.; Zhang, Y.; Jia, X., "Multilevel data clustering for spatial join processing", International Symposium on Database Applications in Non-Traditional Environments, Pp. 218 – 225, 1999.

[15] Yaqin Wang; Yue Chen; Minggui Qin; Yangyong Zhu, "SPANBRE: An Efficient Hierarchical Clustering Algorithm for Spatial Data with Neighborhood Relations", Fourth International Conference on Fuzzy Systems and Knowledge Discovery, Pp. 665 – 669, 2007.