International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue 3, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

Finding Water Region in Aurangabad City Using SCLD Method in Spatial Data Mining

Varsha Kundlikar, Meghna Nagori

Abstract **- Spatial data mining is the discovery of interesting relationships and characteristics that may exist implicitly in spatial databases. In this paper authors propose a system that will find water regions in Aurangabad city situated in Maharashtra of India using clustering method called SCLD (Spatial Clustering in Large Database), whose aim is to identify spatial structures that may be present in the large spatial data as well as to identify different types of area where stored water used for touristy purpose, dirty water that is not properly disposed.**

Keywords **-- Spatial Data Mining, Clustering Algorithm, SCLD.**

I. INTRODUCTION

Spatial data mining is the discovery of interesting relationships and characteristics that may exist implicitly in spatial databases. Because of the huge amounts (usually, terabytes) 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 an important role in

1. Extracting interesting spatial patterns and features,
2. Capturing intrinsic relationships between spatial and nonspatial 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.

This paper is based on finding water region of stored water used for touristy purpose, dirty water that is not properly disposed. In Aurangabad city stored water is used for touristy place, swimming pools, water fountain etc. one of famous place is Panchakki in Aurangabad. There are some areas which are having water disposal problem of dirty water, waste water, and rainy water. Drainages are not properly made.

Using this paper author can help to municipal corporation to find and improve the areas where waste water and rainy water stored and that may reasons for many diseases.

II. APPROACH

Cluster Analysis is a branch of statistics that, in the past three decades, has been intensely studied and successfully applied to many applications. To the spatial data mining task at hand, the attractiveness of cluster analysis is its ability to find structures or clusters directly from the given data, without relying on any hierarchies. However, cluster analysis has been applied rather unsuccessfully in the past to general data mining and machine learning. The complaints are that cluster analysis algorithms are ineffective and inefficient. Indeed, for cluster analysis to work effectively, there are the following key issues:

- Whether there exists a natural notion of similarities among the “objects” to be clustered. For spatial data mining, our approach here is to apply cluster analysis only on the spatial attributes. If these attributes correspond to point objects, natural notions of similarities exist (e.g., Euclidean or Manhattan distances). However, if the attributes correspond to polygon objects, the situation is more complicated. More specifically, the similarity (or distance) between two polygon objects may be defined in many ways, some better than others. But, more accurate distance measurements may require more effort to compute. The main question then is for the kind of spatial clustering under consideration, which measurement achieves the best balance.
- Whether clustering a large number of objects can be efficiently carried out. Traditional cluster analysis algorithms are not designed for large data sets, with say more than 1,000 objects.

III. RELATED WORK

A Clustering Algorithm Based On Randomized Search (Clarans)

CLARANS is a main-memory clustering technique, while many of the aforementioned techniques are designed for out-of-core clustering applications. We concede that whenever extensive I/O operations are involved, CLARANS is not as efficient as the others. However, we argue that CLARANS still has considerable applicability. Consider the 2D objects to be discussed in this paper. Each object is represented by two real numbers, occupying a total of 16 bytes. Clustering 1,000,000 objects would require slightly more than 16 M bytes of main memory. The algorithm takes as an input the number, *k*, of the desired clusters but such a parameter is often hard to determine in realistic applications.

So a good clustering algorithm should minimize the input parameters. The algorithm first randomly selects *k* points as the centers for the required clusters and assign each data point to its nearest center to form the required clusters. Then the algorithm tries to find better solutions. Better solution means a new set of centers that minimize the sum of the distances that each object has to cover to the center of its clusters.

A. Proposed Algorithm

Author proposed an efficient algorithm for spatial clustering of large spatial databases. The algorithm overcomes the problems of the previous work. The

algorithm divides the spatial area into rectangular cells and labels each cell as *dense* (contains relatively large number of points) or *non-dense*. The algorithm finds all the maximal, connected, and dense regions that form the clusters by a breadth-first search and determine a center for each region.

Algorithm SCLD (Spatial clustering in large database)

***Input*:**

1. A set of *N* objects in a spatial area *S*.
2. A square number, *m*, which represents the number of cells in the spatial area such that *m*<<*N*.
3. A percentage, *h*, used to determine the dense cells according to definition

***Output*:** Clusters with their centers

***Method*:**

1 *w* √*m*;
2 Divide the spatial area into *m* rectangular cells by dividing each of dimension of the spatial area into *w* equal segments;
3 For(*i*=0; *i*<*m*; *i*++)
{
determine for the cell, *ci*, parameters:
- *ci n* : the number of the objects in that cell.
- *ci m* : the mean object of all objects in that cell.
}
4 *d round((m/n)*h)* ;
5 For(*i*=0; *i*<*m*; *i*++)
{
if (n_{ci} d) then
ci is labeled as *dense;*
else
ci is labeled as *non-dense;*
}
6 *j*=0;
7 For(*i*=0; *i*<*m*; *i*++)
{
if (*ci* is dense) then,
{
if (*ci* is not processed yet) then,
{
- Construct a new cluster, *rj*, and mark the cell *ci* as an element of *rj;*
- Put the *dense* neighboring cells of *ci* in a list, *Q;*
- While (*Q* is not empty)
{
- Take the first element, *c'*, from *Q*, mark *c'* as an element of the cluster *rj* and add to *Q* the *dense* neighbouring cells of *c'*, that have not been processed yet;
}
- *j*++;
}
}
8 For(*i*=0; *i*<*j*; *i*++)
Compute the mean object of the cluster *rj;*
9 For(*i*=0; *i*<*j*; *i*++)
{
- List all neighboring non-empty cells of *rj* in a list , *QN;*
- Sort, *QN* in a descending order, according to the number of objects in the cells;
}
10 While (*QN* isn't empty)
{
- Take the first element, *e*, of *QN*;
- Find all neighboring clusters of *e;*
- Determine which of those clusters can be extended to contain *e* without becoming non-dense;
- Assign *e* to the nearest cluster obtained in previous step;
}
11 Re-compute a center for each extended cluster;
12 Output the clusters with their centers;

B. System Design

Proposed system flow will be as follows:

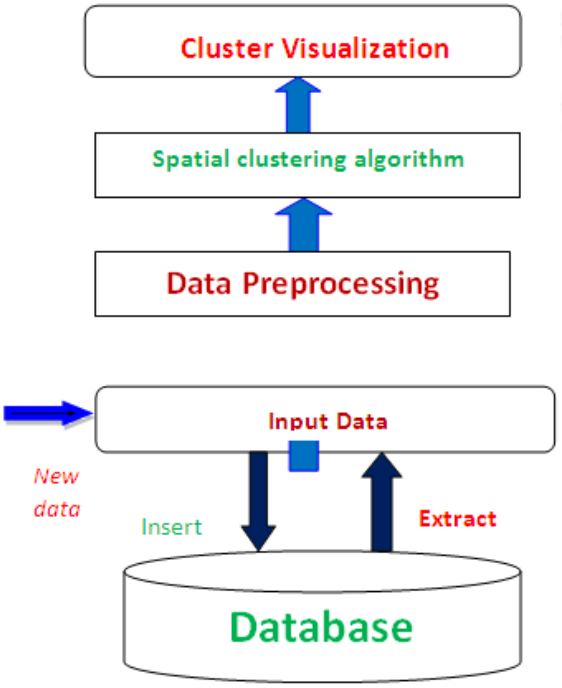

Fig.1. System flow diagram

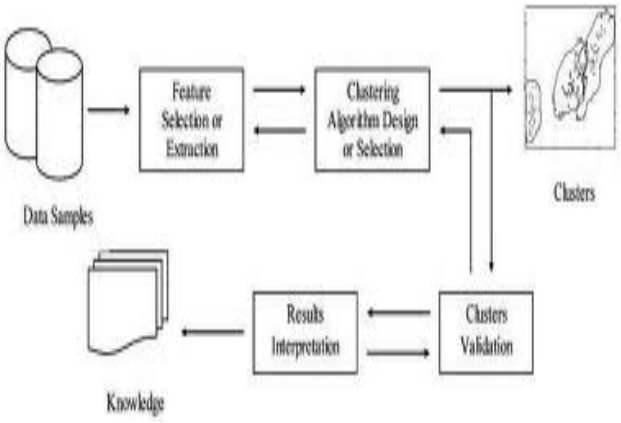

Fig.2. Block diagram for spatial data mining

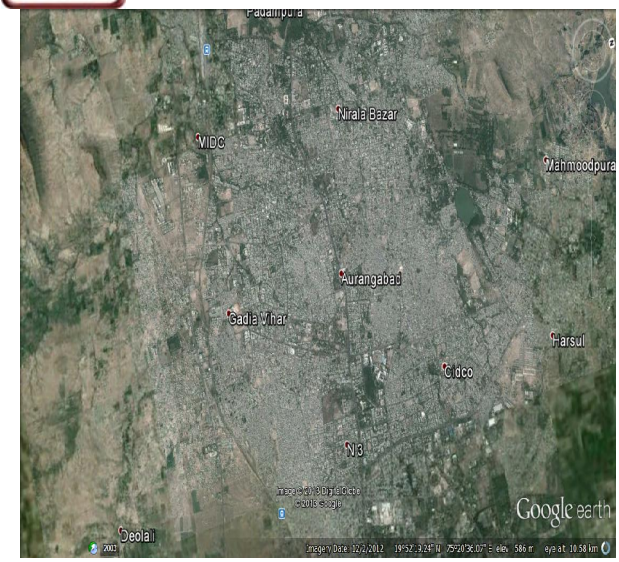

Fig.3. Aurangabad City Area

IV. CONCLUSION

In this paper we proposed a system to develop an application of spatial data mining to find interested spatial data i.e. water region in Aurangabad City in Maharashtra of India using an efficient algorithm for large spatial database called spatial clustering in large database (SCLD).

REFERENCES

[1] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications," Proc. 1998 ACM-SIGMOD, pp. 94–105, 1998.

[2] R. Agrawal, S. Ghosh, T. Imielinski, B. Iyer, and A. Swami, "An Interval Classifier for Database Mining Applications," Proc. 18th Conf. Very Large Databases, pp. 560–573, 1992.

[3] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," Proc. 1993 ACM Special Interest Group on Management of Data, pp. 207–216, 1993.

[4] M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering Points To Identify the Clustering Structure," Proc. 1999 ACM Special Interest Group on Management of Data, pp. 49–60, 1999.

[5] W.G. Aref and H. Samet, "Optimization Strategies for Spatial Query Processing," Proc. 17th Conf. Very Large Databases, pp. 81-90, 1991.

[6] A. Borgida and R. J. Brachman, "Loading Data into Description Reasoners," Proc. 1993 ACM Special Interest Group on Management of Data, pp. 217–226, 1993.

[7] P. Bradley, U. Fayyad, and C. Reina, "Scaling Clustering Algorithms to Large Databases," Proc. Fourth Int'l Conf. Knowledge Discovery and Data Mining, pp. 9–15, 1998.

[8] T. Brinkhoff and H.-P. Kriegel, B. Seeger, "Efficient Processing of Spatial Joins Using R-Trees," Proc. 1993 ACM Special Interest Group on Management of Data, pp. 237-246, 1993.

[9] D. Dobkin and D. Kirkpatrick, "A Linear Algorithm for Determining the Separation of Convex Polyhedra," J. Algorithms, vol. 6, no. 3, pp. 381–392, 1985.

[10] M. Ester, H. Kriegel, and X. Xu, "Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification," Proc. Fourth Int'l Symp. Large Spatial Databases. (SSD '95), pp. 67–82, 1995. [11] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Large Clusters in Large Spatial Databases with Noise," Proc: Second Int'l Conf. Knowledge Discovery and Data Mining, 1996.

[12] S. Guha, R. Rastogi, and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases," Proc. 1998 ACM Special Interest Group on Management of Data, pp. 73–84, 1998.

[13] O. Gu¨ nther, "Efficient Computation of Spatial Joins," Proc. Ninth Conf. Data Eng., pp. 50-60, 1993.

[14] J. Han, Y. Cai, and N. Cercone, "Knowledge Discovery in Databases: an Attribute-Oriented Approach," Proc. 18th Conf. Very Large Databases, pp. 547–559, 1992.

[15] A. Hinneburg and D. A. Keim, "An Efficient Approach to Clustering in Large Multimedia Databases with Noise," Proc. 1998 Int'l Conf. Knowledge Discovery and Data Mining, pp. 58–65, 1998.

[16] Y. Ioannidis and Y. Kang, "Randomized Algorithms for Optimizing Large Join Queries," Proc. 1990 ACM Special Interest Group on Management of Data, pp. 312–321, 1990.

[17] Y. Ioannidis and E. Wong, "Query Optimization by Simulated Annealing," Proc. 1987 ACM Special Interest Group on Management of Data, pp. 9–22, 1987.

[18] G. Karypis, E.-H. Han, and V. Kumar, "CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling," Computer, vol. 32, no. 8, pp. 68–75, Aug. 1999.

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

[20] D. Keim, H. Kriegel, and T. Seidl, "Supporting Data Mining of Large Databases by Visual Feedback Queries," Proc. 10th Conf. Data Eng., 1994.

[21] D. Kirkpatrick and J. Snoeyink, "Tentative Prune-and-Search for Computing Fixed-Points with Applications to Geometric Computation," Proc. Ninth ACM Symp. Computational Geometry, pp. 133–142, 1993.

[22] R. Laurini and D. Thompson, Fundamentals of Spatial Information Systems. Academic Press, 1992.

[23] W. Lu, J. Han, and B. Ooi, "Discovery of General Knowledge in Large Spatial Databases," Proc. Far East Workshop Geographic Information Systems, pp. 275–289, 1993.

[24] G. Milligan and M. Cooper, "An Examination of Procedures for Determining the Number of Clusters in a Data Set," Psychometrika, vol. 50, pp. 159–179, 1985.

AUTHOR'S PROFILE

Ms. Meghana Nagori
passed her bachelor degree in Computer Science and Engineering from Government Engineering College of Aurangabad of BAMU University Aurangabad. She is completing her research work on computer science for doctorate degree. Currently she is working as Associate professor in an autonomous Government Engineering College of Aurangabad.

Ms. Varsha Kundlikar
passed her B.E. in information technology from Government Engineering College of Aurangabad of BAMU University Aurangabad in 2006. She is working as lecture in Government Polytechnic Aurangabad and appeared to complete her Master degree in Computer Science.