International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 9, September- 2014

A Review on Various Nearest Neighborhood Algorithm for Spatial Data Mining Applications

D. Gandhimathi, MCA., M.Phil.,*
Assistant Professor,
Dept. of CA and SS,
Sri Krishna Arts and Science College,
Coimbatore, Tamil Nadu.

C. Gomathi, M.Sc., M.Phil.,
Assistant Professor,
Dept. of CA and SS,
Sri Krishna Arts and Science College,
Coimbatore, Tamil Nadu,

K. Devika Rani Dhivya, M.Sc., M.Phil., MBA.,
Assistant Professor,
Dept. of CA and SS, Sri Krishna Arts and Science College,
Coimbatore, Tamil Nadu.

***Abstract*— The extraction of hidden predicated information from large databases is defined as Data mining. It is also the process of identifying meaningful, new correlation pattern and trend by sifting through a large amount of data stored in repositories, using pattern recognition techniques. Data sources for data mining can be text mining, web mining, sequence mining, spatial datamining.The Process of identifying interesting and unknown potentially use full pattern from large spatial data set is known as Spatial Data mining. Spatial data mining seeks to perform similar generic function as conventional data mining tools, but take the spatial (location) features of spatial information. Spatial Data mining techniques are widely applied to location identification, road network, flow of vehicles etc. This paper outlines the general purpose of spatial mining, Neighborhood nodes to identify the spatial data's, few algorithms to identify the neighborhood nodes**.

Keywords*— *Data mining, Spatial Data mining, Nearest Neighborhood Algorithm

I. INTRODUCTION

The process of analyzing data from various perspectives and summarized information are termed as KDD. Data mining analysis tends to work up from the data and the best techniques which use large volume of data to arrive at reliable conclusion and decisions. Spatial data mining is the applications of dataminig to spatial data's. The wide spread use of GIS by the public and the company necessitate the development of adequate mining tools for georeferenced location(Spatial) data .Using GIS ,the user can query spatial data and perform simple analytical task using program or queries.GIS methods are crucial for data access, Spatial join and graphical map display.

Spatial Data Mining (SPDM) is the process of extracting knowledge, spatial data and frame a relation between spatial and non spatial data's [3]. To extract the spatial data's if we receive only one data set patterns are related to the concept of clusters regularities. For more than one data set patterns are related to co-locations in space [2]. SPDM has deep roots in traditional spatial analysis field-spatial analysis, cartography etc. and also in data mining field to do the process of Clustering, Classification, Association rule etc. [1].

Various applications of SPDM to identify the nearest shop through the search of google map-Geo marketing, environmental studies, prediction of vehicle movement ,online position aware, Location based services and so on. Data analysis in geographic data is to locate near to one another in space which can share similar attribute values [3]. To extract useful knowledge from spatial datamining is to identify the neighbors and object. Because significant influence on the object is consider through attributes of the neighbors of some object.

II. GIS IN SPATIAL DATA MINING

Through the technology of internet access through wireless technologies-Bluetooth, WAP will bring the required connectivity based on satellite technologies which is identified as Global Positioning System(GPS).Some of the applications include position enabled tourist services, safety and security services, traffic services etc. These are the huge area of applications of "Geographic", as well as database management technologies. GIS is the process of extracting features about a particular object or location from the images and data generated using the remote sensing and GIS technologies. It caters to requirements in various sectors ranging from industrial to non-industrial, defense to local security, and public to private. GIS is much more flexible than traditional cartography. Geographical information about Maps which contain different locations, utilities, routes, etc if stored in a common storage location, the information can be easily retrieved. GIS storage needs to be designed and then modeled to form a complete geo database. A key index variable for GIS uses spatio-temporal (space-time) location for different information. With key index variables various tables can be related similar to a relational database. Keys are identified as location in space-time. Security intelligence uses techniques of GIS spatial analysis. [13] Through GIS various process such as hypotheses development, visualization in geographical space, searching of patterns in multi-dimensional abstract space.

Combining with data mining process different task of pre-selecting and presenting to the analyst for the given demand's are extracted with GIS.

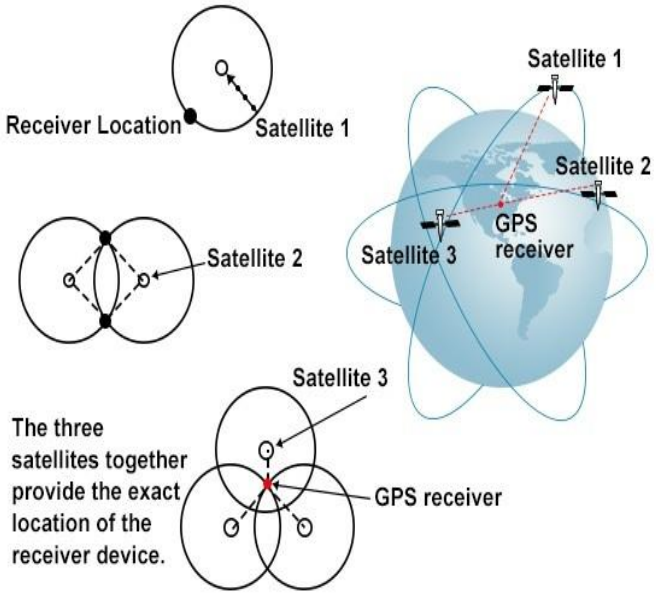


Fig: 1 GPS Tracking

III. SPATIAL DATA MINING

Extraction of knowledge and spatial relationship and other properties which can be stored in spatial data base are defined as spatial mining. Discovery of interesting relationship and characteristics through in spatial database are identified as Spatial mining [1].Through spatial data mining we can understand the spatial data-capturing intrinsic relationship between spatial and non spatial data, presenting data regularity at higher conceptual level. Reorganizing spatial database to accommodate data semantics to achieve better performance.

Digital information is used in Modern GIS technologies, for digitizing the data. Key index variable as used as spatial data's for GIS usage. Extracted Data's are restructured into different formats to use in GIS. For example: A satellite image map can be converted to vector structure using GIS. Spatial data mining system are identified as logical progression for spatial data analysis technology using GIS and Data mining techniques. The data inputs for spatial data mining are complex than the inputs of regular data mining because which supports points, lines and polygons for various objects. Two distinct types of attributes use in spatial data mining inputs are: non-spatial attribute and spatial attribute. Spatial locations are defined as spatial attributes for spatial objects. Where spatial locations are denoted as- longitude, latitude, elevation etc.

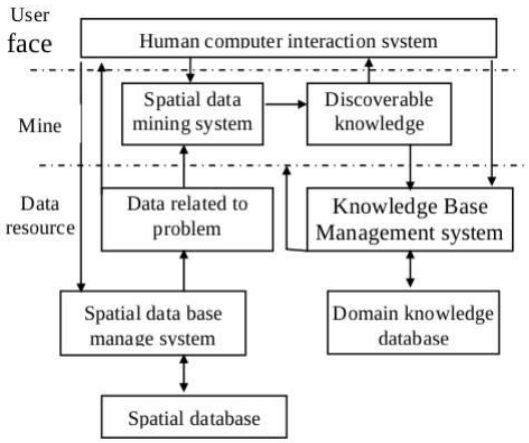


Fig: 2 Spatial Data Mining Architecture

❖ *Spatial Data Mining Task*

The main tasks of spatial data mining are:

(i) Data summarization,

(ii) Identify classification rules,

(iii) Create clusters for similar objects,

(iv) Characterize data using, predicted associations and dependencies

(v) For general trends detect deviations.

The above tasks are carried out using different methods, which are derived through statistics and machine learning process. In General the extension of data mining tasks using spatial data's with its criteria are identified as spatial data mining

(i) Summarization of Data

The goal to describe data involves extending statistical methods of variance or factorial analysis to spatial structures. Then apply generalization method to spatial data.

(ii) Classification Rule

Frame a set of rules to determine the class for the classified object according the attributes in its database. Through Association rules patterns are described, which are available in the database.

(iii) Clustering of Similar objects

Clusters are framed for groups the object from its database which should support object in one cluster are similar than the objects from different clusters are dissimilar

(iv) Characteristic rules

To describe dependencies and association for some part of database e.g. "Vehicle is an object in the place which crosses various routes." Differences between two parts of database are described as e. g. to identify the differences between districts with high and low educated rate.

(v) Trend detection

Trends in database are temporal pattern for series of data. Basically it is defined for neighborhood of a spatial object as a pattern of change of a non-spatial attribute [10]

In this paper the work is focused on Neighborhood relations for spatial attribute. How spatial data's are defined as neighborhood graph, neighborhood path,& neighborhood index. Some discussion is done on various algorithms to identify the nearest neighborhood node .Finally few applications of spatial dataminig based on nearest neighborhood node.

IV. NEIGHBORHOOD RELATION IN SPATIAL DATA MINING

The neighborhood relations for Spatial data mining process are based on neighborhood graphs and neighborhood paths for an spatial objects. The logical operator are used to combine the various types of spatial relations such as: topological, distance and direction relations for complex neighborhood relation. Points, lines, polygons or polyhedrons are all are identified as Spatial objects..For example, a polygon can be represented by its edges .Topological relations are based on the boundaries, interiors and complements of the two related objects and are invariant under transformations which are continuous, one-one, onto and whose inverse is continuous [2].

The concepts of neighborhood graphs and neighborhood paths are relate the neighborhood relations, which uses basic operations to manipulation the spatial objects. [5]

Definition: ***Neighborhood graphs-*** Let *neighbor* be a neighborhood relation and *DB* be a database of spatial objects. A *neighborhood graph* $G^{DB}_{neighbor}$ =(N,E)□ is a graph where the set of nodes *N=DB* which corresponds to the set of edges *E*□□ *N* X *N* , where the pair of nodes are identified as (*n1*, *n2*) iff *neighbor*(*n1*,*n2*) holds..

Defnition : Martin Ester, Alexander Frommelt, Hans-Peter Kriegel, Jörg Sander [6],had defined ***Neighborhood Path*** -A *neighborhood path* is a sequence of nodes [*n1, n2, . . ., nk*], where *neighbor*(*ni*, *ni*+1) holds for all *ni* □□*N*□□1 □*i*□*k* The number *k* of nodes is called the *length* of the neighborhood path. A neighborhood path [*n1, n2, . . ., nk*] is *valid* iff □□*i* □□*k, j < k: i*□*j*□*ni* □□*nj*

Spatial Indexing Structure through R-trees [8] is used in Spatial Data Base Management System to speed up the processing of Spatial Queries to retrieve its nearest neighborhood node.

Definition: Neighborhood index -Let *DB* be a set of spatial objects and let *max* and *dist* be real numbers. Let *D* be a direction relation and *T* be a topological relation. Then the *neighborhood index* for *DB* with maximum distance *max*, denoted by

$I^{DB}max$ = {(O_1,O_2,*dist*,*D*,*T*) | O_1, O_2□□□*DB*, O_1 *distance=dist* O_2 □□*dist* □□*max* □□O_2 *D* O_1 □□O_1 *T* O_2}.

Neighborhood graphs are supported by Neighborhood Indices. Neighborhood indices for neighborhood graph are calculated using Critical distance. Where the critical distance of a neighborhood relation r is identified as the maximum possible distance between the pair of object O_1 and O_2 which satisfies O_1 r O_2.

A. Nearest Neighborhood Algorithm

Martin Ester ,Hans-Peter Kriegel Jorg Sander [4], Defines the nearest neighborhood algorithm which process the neighborhood operation with the use of neighborhood index .According to the algorithm Index selection are use to select neighborhood index which is proceeded to identify the spatial index structure. Candidate objects are retrieved using Filter step. In the refinement step ,for all the candidate neighborhood relation and additional predicate pred are evaluated for all objects which return the resulting neighbors.

Martin Ester, Stefan Gundlach, Hans-Peter Kriegel, Jörg Sander [8], Discussed that in a neighborhood graphG^{DB}if *c-distance*(*r*) because all neighbors then neighborhood index Nc^{DB} is *applicable* to the G^{DB} ,where *r* can be denoted in the neighborhood index. In two indices $Nc1^{DB}$ and $Nc2^{DB}$, $Nc1^{DB}$ is more efficient as it will be smaller than the index $Nc2^{DB.}$ Neighborhood algorithm are extended using depth-first search. A buffer size *max-length* is use to store the intermediate results. The nodes are retrieved for the purpose of potential extensions of a candidate path, then neighbors operations are use to indicating that the efficiency of the operation.

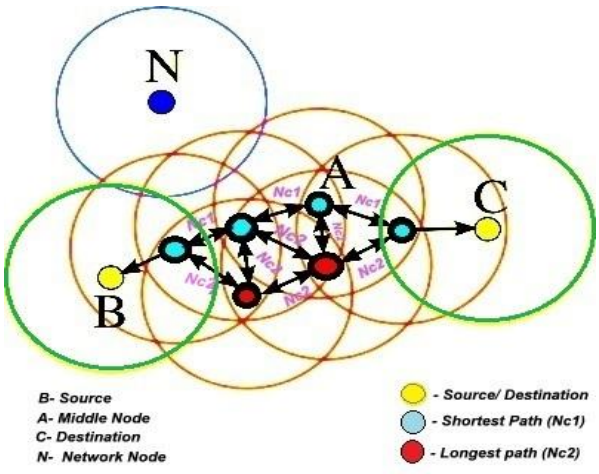
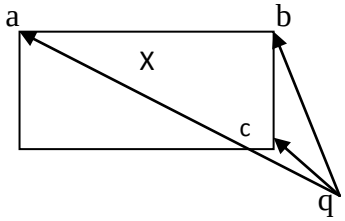


Fig: 3 NNA

The above fig:3 explains the in road network from the source B to the destination C two possible routes are available.The shortest path Nc1 Is better than longest path Nc2 by identifying the nearest node which are closest to each other.

Dr.Ch.GV Prasa,V.Manoj Kumar,R.Pavitra [2] Discussed k Nearest Neighbor (kNN) query which is used to find nearest object to a given query point. kNN Algorithm can use to process in DF–kNN and BF-kNN. Nodes in Depth First manner maintains the k nearest object as candidates when every nodes is either visited or discarded, the k objects remaining in the candidate set are the resultant k nearest neighbors. BF-NN visit data object in tree nodes in the order of their distance to the query point. BF-NN are proceeded by calculating MINDIST.



MAXDIST(q,X)=||q-a||,MINMAXDIST(q,X)=||q-b||
MINDIST(q,X)=||q-c||

With a priority queue PQ as initial values, BF-NN algorithm process is started. Which also has empty set A that will contain the resultant k NNs .If the entry retrieved from PQ is an object, the object is the next NN otherwise retrieved child nodes are stored in the node. For each child node new entry is created, the MINDIST is calculated and the entry is then inserted into PQ. This process is repeated when kNN are identified or PQ is exhausted. Finally set A contain the resultant of kNNs.

B. Applications of Neighborhood Algorithms

Miyoung Jang, Min Yoon, and Jae-Woo Chang [10] discussed K-NN query processing which is used in road network. In location based applications user, moves along with the road network. In [10] proposed a spatial database encryption scheme which produces a transformed database from an original database by using network distance. To generate index, encrypt both distance and anchor information by using order preserving encryption scheme. Along with the above process k-NN query processing algorithm performs on transformed data in road network. Using k-NN query, efficient query processing and spatial data privacy are maintained.

Venkata Ratnam Ganji, Siva Naga Prasad Mannem [11] Use algorithm SODRNN-Stream Outlier Detection which is based on Reverse kNearest Neighbors. Credit card fraud detection is identified using Reverse k-NN. In the process Supervised, Unsupervised learners are identified. Using supervised it need historical database which is used to detect fraud of a type which had occurred earlier. But unsupervised learners use undiscovered types of fraud detection. Unsupervised learners are used in credit card detection, which detect the changes in behavior or unusual transaction.

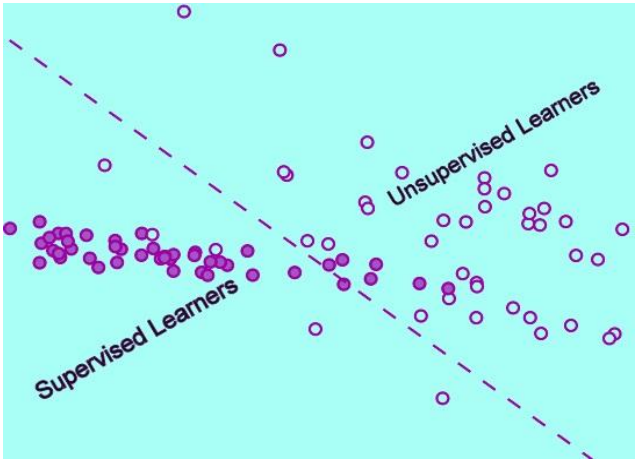


Fig : 4 Supervised & Unsupervised learners

The above figure 4 supervised learners are identified as dark circles where unsupervised learners are identified as hollow circles. Using the supervised details for credit card users identify similar data's of usage of card to detect the fraud undertaken. Each similar data's are grouped as clusters which can be identified as nearest node to utilize the Reverse k-NN process.

Ravikumar K. , Gnanabaskaran A**.,** [12] Discussed the implicit knowledge collected from spatial data through GIS provide database related to road accidents,road network ,flow of vehicles etc.Identified data can be use for traffic risk analysis. Using decision tree process for accident data's corresponding to road section, we could identify the data's in tabular format but not exploit geographical location. When these challenges combined with ACO, which suggest efficient properties in spatial trend detection. Finally It provide spatial decision tree structure with optimized route structure with ant agents for spatial modeling of traffic risk patterns.

V. CONCLUSION

Using spatial data base interesting relationship and characteristics are discovered for Spatial data mining. Where spatial data's are the data related to objects which occupy space. Neighborhood graphs and neighborhood paths are main aspects of spatial data mining, which is also defined as neighborhood relations between spatial objects. This paper concludes, that the above algorithms where used in various method to find the nearest node used in spatial data's .Thus these algorithm's are more efficient in finding of nearest neighborhood to its related applications.

VI. REFERENCES

[1] *Algorithms for Charaterization and Trend Detection in spatial databases*,Martin Ester,Alexander Frommelt,Hans-Peter kriegel,Jorg Sander,KDD-98,pp.44-50.

[2] *Spatial Data Mining Evaluvation of Visible Nearest Neighbor Query* ,Dr.Ch.GV Prasa,V.Manoj Kumar,R.Pavitra,Publ.IJCST Vol.2,Issue 2,June 2011.

[3] *An Effective Analysis of Spatial Data mining methods using Range Queries*,*Gangireddy Ravikumar, Mallireddy Sivareddy*,JGRCS,Vol3,No1,Jan 2012.

[4] *Algorithms and Application for Spatial DataMining* ,Martin Ester ,Hans-Peter Kriegel Jorg Sander,Publ. Geographic Data Mining and Knowkedge Discovery ,2001.

[5] *Knowledge Discovery in Spatial Databases*,Martin Ester, Hans-Peter Kriegel, Jörg Sander, Invited Paper at 23rd German Conf. on Artificial Intelligence (KI '99),Bonn, Germany, 1999.

[6] *Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support*,Martin Ester, Alexander Frommelt, Hans-Peter Kriegel, Jörg Sander, Issue on: "Integration of Data Mining with Database Technology", Data Mining and Knowledge Discovery, an International Journal, Kluwer Academic Publishers, 1999.

[7] Guttman A.: *"R-trees: A Dynamic Index Structure for Spatial Searching",* Proc. ACM SIGMOD Int. Conf. on Management of Data, 1984, pp. 47-54.

[8] *Database Primitives for Spatial Data Mining*,Martin Ester, Stefan Gundlach, Hans-Peter Kriegel, Jörg Sander, Proc. Int. Conf. on Databases in Office, Engineering and Science, (BTW '99), Freiburg,(1999).

[9] *An Effective analysis of Spatial Data Mining Methods Using Range Queries*,Gangireddy Ravikumar ,Malireddy Sivareddy,Pub.JGRCS ,Vol.3 ,Jan 2012.

[10] *A k-Nearest Neighbor Search Algorithm for Privacy Preservation in Outsourced Spatial Databases* **,** Miyoung Jang, Min Yoon, and Jae-Woo Chang, ISA 2013, ASTL Vol. 21, pp. 223 - 226, 2013 © SERSC 2013.

[11] *Credit card fraud detection using anti-k nearest neighboralgorithm*, Venkata Ratnam Ganji, Siva Naga Prasad Mannem, ISSN : 0975-3397 Vol. 4 No. 06 ,pp. 1035-1039June 2012.

[12] *ACO based spatial data mining for traffic risk analysis*,Ravikumar K. and Gnanabaskaran A**.,** International Journal of Computational Intelligence, ISSN: 0976–0466 & E-ISSN: 0976–0474, Volume 1, Issue 1, 2010, PP-06-13.

[13] www.sblgis.com/feature-extraction.aspx