An Empirical Research on Spatial Data Mining

K. Sivakumar, A.S.Prakaash

Abstract*—Spatial data mining is a process of extracting expertise from large volumes of spatial data collected from different applications such as remote sensing, geographic systems and social networks, etc. The collected spatial data are too difficult for the human to analyze. Recent research focuses on data mining to extend the data mining scope from relational storages to spatial databases. A lot of effort put forth to summarize various spatial based knowledge discovery in data mining such as based on generalization, clustering based, spatial associations based, and approximations and aggregations based knowledge discovery are discussed. The discussion shows that spatial data mining is a promising area of information discovery and can lead to extensive research and many challenging issues.

Keywords:Spatial data mining, Clustering, Spatial associations, Knowledge discovery, Aggregation.

I. INTRODUCTION

For large databases, data collection techniques like reading barcode, remote sensing, satellite telemetry, etc. store vast volumes of data.

This vast data collection involves the need to extract discovery of knowledge, leading to a promising field called Knowledge Discovery Databases (KDD)[15]. Database knowledge exploration is the discovery of necessary patterns from large databases and is combined with multiple fields such as machine learning, database systems, data visualization, analytics, and information theory. Most data mining studies are presented in relational databases and have become a big demand in spatial databases, temporal databases, interactive databases, object-oriented databases, etc.[1]. Spatial data is connected to objects that occupy space by the forms of spatial data and the relationships between them.

Spatial data is defined by spatial indexing structures and access methods that pose challenges to extract information from spatial data are used to access spatial data. Space data mining is an analysis of indirect information and other patterns that are not directly contained in space databases [9]. This discovery of knowledge in machine learning[12], database systems[16] and statistics[11] is the basis for the discovery of knowledge in databases. Advanced spatial databases, such as spatial data structures[6], spatial reasoning[4] and so on, are also helped pave the way for spatial data mining.

There is a high complexity in working with spatial information which inhibits the performance of spatial data mining algorithms with access methods and requires an understanding of spatial data to extract the necessary patterns from broad spatial databases. Discovering similarities between geographic and semi-spatial data, query optimization, and information reorganization are the highly challenging tasks. Different forms of spatial information are characteristic laws, filtering rules, collection of influential clusters, spatial data associations, and so on. Therefore, a detailed discussion of spatial data mining[2] can enhance this article. The overall picture of spatial data mining and the analysis with the aspects relevant to spatial exploration of different categories in information discovery were addressed.

The research contribution of this paper is

1. Survey related to spatial data mining is described clearly.

2. Various standard approaches of data extraction in spatial mining are discussed.

3. Future directions of spatial data mining are addressed in detail.

The above points make a clear distinction between this survey and other recent surveys. It gives the description as broad as previous works. The paper's organization as follows: Section II reviews the background of spatial data mining, architecture, and different types of discovery of knowledge. Section III deals with possible directions for research. The current work is outlined in Section IV.

II. BACKGROUND AND RELATED WORK

Statistical spatial analysis is used to analyze spatial data and became a well studied research area. Therefore, a huge amount of optimization methods were developed. Realistic models of numerical data of spatial phenomena were handled well with existing optimization methods. But, the statistical independence among the spatially distributed data became a major disadvantage due to many spatial data are made by neighboring objects. Regression models can be used to solve this drawback but the whole modeling process is more complicated because of not enough amounts of statistical expertise and it is not the good approach for the end users to analyze the spatial data.

Nonlinear rules cannot be modeled by the statistical approach and do not work with incomplete data. It also requires expensive results estimation to discover information from large databases. Many computational methods concentrate on relational databases incorporating fields such as machine learning, databases, and statistics[5]. Before the technique of statistical cluster analysis, machine learning techniques are widely used in spatial data mining. Certain methods of generalization, inference, and

Revised Manuscript Received on September 14, 2019.

K. Sivakumar, Professor, Department of Mathematics,Sathyabama Institute of Science and Technology,Chennai, Tamilnadu, India

A.S.Prakaash, Research Scholar, Department of Mathematics,Sathyabama Institute of Science and Technology,Chennai, Tamilnadu, India.

Retrieval Number: L113610812S219/2019©BEIESP
DOI: 10.35940/ijitee.L1136.10812S219

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

classification are also improved in spatial repositories for information discovery. A review of the basic concept of spatial data mining and its various commonly used methods in spatial data mining is discussed in this section[13].

2.1 Spatial mining architecture

Various designs are proposed for extracting information such as DBLEARN/DBMINER [7], parallel architecture [8], and multicomponent architecture [12]. Existing architectures have been used to manage data mining in space and appear to be very general. Figure 1 presents the basic architecture for spatial data mining. Using the DB interface, data is extracted from the storage to enable the query optimization. For efficient processing, R-trees, a spatial data index structures is used. The focusing element is used to identify useful data parts for pattern recognition to select objects with good results in their use. Pattern extraction module used to discover rules and patterns with the usage of statistical, machine learning, data mining, and computational geometry algorithms. To eliminate redundant knowledge, the evaluation module used to find the meaning of patterns. The last four components communicate with each other through the part of the controller [18].

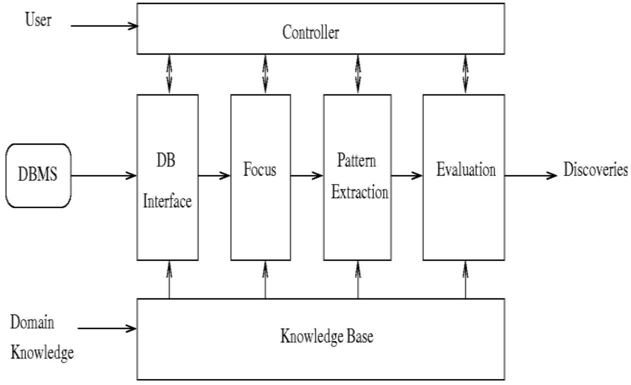

Figure 1. Basic architecture of spatial data mining

2.2 Knowledge Discovery in Database

2.2.1 Generalization based knowledge discovery

The knowledge discovery based on generalization requires concept hierarchies from spatial databases such as non-spatial and spatial and can be automatically generated through data analysis. The concept tree with the lower levels of concept remains consistent and for spatial information a different hierarchy can exist. Areas are incorporated into broader areas in a system of generalization. For the hierarchy of generalization, induction on attribute-oriented is performed to summarize relationships at higher levels of abstraction between spatial and non-spatial information.

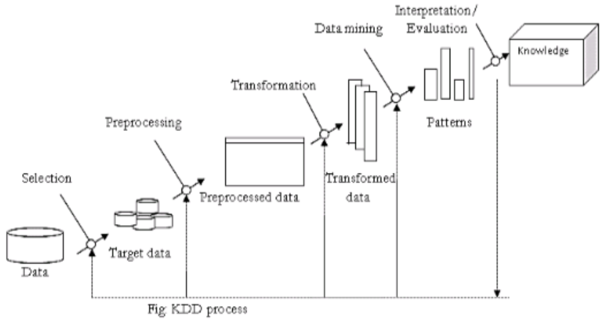

Figure 2. Generalization based knowledge discovery

In non-spatial data, when attribute values in a tuple are removed, attributes are removed when generalization is further impossible and identical tuples are merged, the concept hierarchy is changed to generalized values. Induction to the desired level is continued up to each generalized attribute. When in the generalized table the different values for the attribute are no higher than the generalization threshold, the desired level is reached. The number of merged tuples is stored in additional attribute count in order to enable acquired knowledge. Figure 2 shows the discovery of knowledge based on generalization. Two generalization-dominant and non-spatial-data-dominant algorithms are presented. All algorithms adopted the mined rules and the user started the discovery process with an explicitly SQL-like query[10].

2.2.2 Clustering based knowledge discovery

Cluster analysis has been studied extensively and can be found directly from the data without using concept hierarchies. The spatial data mining process based clustering algorithms like PAM is reported to be inefficient in computational complexity and CLARANS (Clustering Large Applications based upon RANdomized Search) is developed for analyzing cluster. For n objects and k clusters, the PAM algorithm has been built with a representative entity for each cluster that is called a medoid. After selecting k medoids, repeatedly attempts better choice of medoids to evaluate all possible pairs of items and the value of clustering is determined for each such combination. The best choice of points is selected as the medoidsfor each iteration. The price is O(k(nk) ^{2}) but quite inefficient for computationally with large values of n and k. NSD(CLARANS) was built on the basis of CLARANS, spatial dominant approach, SD(CLARANS) and non-spatial dominant approach. All algorithms presume that through a training query, DBLearn, the user defines the form of rule to be mined and relevant data.Data Center Infrastructure Management (DCIM) combines IT and facilities management in order to centralize decision control. The main problem with DCIM tools is the analysis of large volumes of data obtained from thousands of resources in real time [3].

2.2.3 Spatial associations based knowledge discovery

To discover association rules of spatial data, spatial objects that are associated in mining large transaction spatial databases is introduced. A law of spatial association in which predicates of space represent a principle of spatial association. Examples are topological relationships such as intersecting, overlapping, disjointing, etc.; spatial orientations such as left, west, etc.; distance data such as near, far, etc. In the spatial association principle, the discovered rules, minimum support and minimum confidence are used. There are a lot of correlations between objects, but there may be little trust in laws. In addition, associations can explain using the low confidence minimum support levels using the highest confidence level. These thresholds can be different when the number of objects with the same description is smaller at each level of non-spatial description of objects. A strong rule is a rule with a great deal of support and great trust.A progressive search method used to minimize spatial computations for mining strong spatial association rules. Computation begins at the high level of spatial predicates and a couple of objects satisfy that the predicate is in the distance no greater than the threshold. Therefore, algorithms are used based on spatial computations such as R-trees or methods of plane-sweep. But expensive spatial computations are applied at lower levels of concept and it will certainly not be large in detailed spatial relationships. The iteration process is carried out at high levels using minimum support for the discovery of knowledge [18].

2.2.4 Approximation and aggregation based knowledge discovery

The characteristics of the clusters with respect to features is the major problem to measure the aggregate proximity. The aggregate proximity is the proximity of the cluster set of points to a feature against the distance between a cluster and the boundary of features. Using the nearest k neighbor like k-d trees, R-trees turns out to be unable to conduct the search because the distance between the cluster and a function is the distance like centroids between the boundaries. The costs of building the indices are prohibitive due to indices may not be used frequently. Therefore, computational geometry were presented to find out given cluster characteristics in terms of the features close to it and used such concepts to reduce multiple levels of candidate features. A large number of features are collected for aggregation and approximation from multiple maps along with the cluster and knowledge about spatial relationships are discussed[5].

III. SPATIAL DATA MINING CHALLENGES AND RESEARCH DIRECTIONS& RESULTS

In spatial data mining, data access methods are distinct from accessing methods in relational database and handling complex spatial objects using traditional data mining methods are difficult. Spatial data mining algorithms lacks in refining discovery patterns. The error patterns are increasing the search space of algorithm and there is a need to design an effective knowledge discovery algorithm for removing unnecessary data. The development of database query language is needed for efficient spatial data mining. But the knowledge of the domain expert is not used efficiently in the knowledge discovery process. The process of spatial data mining is also unable to control by users because knowledge extraction through spatial data mining is limited. Recently, developed knowledge system is constrained to database field.

Spatial data mining with extended spatial database relational model and is not managed through relational databases. In order to design a spatial database, deductive and active databases are developed as advanced database systems such as Object-Oriented (OO). For the recovery of data, an efficient R-trees are used to make OO server. Therefore, the mining process for image databases examines the use of OO software to generalize complex data objects and mining under uncertainty. An evidential theory, like Bayesian methods, can model uncertainty better than traditional probabilistic models. It is possible to extend fuzzy methods to spatial data mining. An alternate clustering methods can also keep information about each object and fit objects that are the same distance from the medoid.

For example, in medical imaging, a spatially discriminating evolution rule is used to find out how certain features evolve over time. Many applications may require spatial data mining to be performed using multiple thematic maps. To extend the generalization-based, interleaving spatial and nonspatial generalizations are considered. Generalization using temporal spatial data involves generalization with sequences of maps being collected at different intervals of time. Parallel data mining using parallel machines can accelerate significantly the parallel knowledge discovery in spatial data mining. The enhancement of data mining methods with mature statistical methods produces interesting new methods and key issues are the design of the user interface in popularizing knowledge discovery techniques. Using 3D devices, the user interface is extended to select objects of interest. The vocabulary must be versatile enough to cover the amount of algorithms that are contained in spatial databases with a wide variety of data types. Knowledge discovery is not enough but also it has to be presented in understanding manner. Humans understanding level of visual data and scenes are also exploited in the data mining. Nevertheless, visualization of multidimensional information is an inexperienced field. Spatial data mining can use computer graphics modeling techniques in this case [14][17].

IV. CONCLUSION

Spatial data mining is a growing research area with a wide range of applications in geo information systems, medical imaging, robot operation, etc. While various methods have been proposed to uncover hidden information from spatial data, this paper analyzed current spatial data mining techniques along with strengths and weaknesses. The future directions for spatial data mining are presented in spatial databases with unexplored information exploration topics that make spatial data mining an attractive and challenging field of research.

REFERENCES

1 R. Agrawal and R. Srikant 1994, “Fast algorithms for mining association rules”, International Conference on VLDB, pp. 487-499, Santiago, Chile, Sept.

2 AsmitaBist and Mainazfaridi 2017, “A survey: On spatial data mining”, International Journal of Engineering Trends and Technology (IJETT) – Volume 46 Number 6, April.

3 Diego GarcıaSaiz, Marta Zorrilla and Jose Luis Bosque 2017, “A Clustering-based Knowledge Discovery Process for Data Center Infrastructure Management”, The journal of supercomputing, Volume 73, Issue 1, January.

4 M. J. Egenhofer 1991, “Reasoning about Binary Topological Relations”, Proc. 2nd Symp. SSD'91, pp. 143-160, Zurich, Switzerland, August.

5 U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy 1996, “Advances in Knowledge Discovery and Data Mining”,AAAI/MIT Press, Menlo Park, CA.

6 R. H. Guting 1994, “An introduction to spatial database systems”, VLDB Journal, 3(4):357-400, October.

7 Han, Jiawei and Fu, Yongjian 1999, “Exploration of the Power of Attribute-Oriented Induction in Data Mining’.

8 M. Holsheimer and M. Kersten 1994, “Architectural Support for Data Mining”, In CWI Technical Report CS-R9429, Amsterdam, The Netherlands.

9 W. Lu, J. Han, and B. C. Ooi 1993, “Discovery of General Knowledge in Large Spatial Databases”,In Proc. Far East Workshop on Geographic Information Systems pp. 275-289, Singapore, June.

10 Dr.M.Hemalatha and N. Naga Saranya 2011, “A Recent Survey on Knowledge Discovery in Spatial Data Mining”, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May.

11 C. J. Matheus, P. K. Chan, and G. Piatetsky-Shapiro 1993, “Systems for Knowledge Discovery in Databases”, IEEE Trans. Knowledge and Data Engineering, 5:903-913.

12 R. S. Michalski, J. M. Carbonnel, and T. M. Mitchell 1983, “Machine Learning: An Articial Intelligence Approach”,Morgan Kaufmann, Los Altos, CA.

13 R. Ng and J. Han 1994, “Efficient and effective clustering method for spatial data mining”, International Conference on Very Large Data Bases, pp. 144-155, Santiago, Chile, September.

14 [G.Parthiban 2018, “A survey report on spatial data mining”, International Journal of Engineering Science Invention .

15 G. Piatetsky-Shapiro and W. J. Frawley 1991, “Knowledge Discovery in Databases”, AAAI/MIT Press, Menlo Park, CA.

16 M. Stonebraker 1993, “Readings in Database Systems”, 2ed.. Morgan Kaufmann.

17 N. Sumathi, R. Geetha and Dr. S. SathiyaBama 2008, “Spatial Data Mining Techniques Trends and its Applications”, Journal of Computer Applications, Vol.1, No.4, Oct-Dec.

18 Thirunavukkarasu K and Dr. ManojWadhwa 2016, “Spatial Data System: Architecture and Applications”, International Journal of Computer Science Trends and Technology (IJCST), Volume 4 Issue 5, Sep - Oct.