ISSN: 2583-195X

Technoarete Transactions on Intelligent Data Mining and Knowledge Discovery
Vol-1, Issue-1, November 2021

Application of Spatial Data Mining in the Process of Discovering Interesting and Previously Unknown, but Potentially Useful, Patterns from Large Spatial Datasets

Abinash Das

Master of Computer Application, Kushagra Institute of Information and Management Science, BPUT, India
Author Email: dasabinash65@gmail.com

Abstract

Spatial data mining has been typically used in the Geographical Information system from physical datasets and locations to real-world events. One of the procedures generally used in capital data mining is vector data representation. Vector data is the most commonly used data across the world. Information in this format consists of tips, angles, and quadrilaterals. It is the simplest method of analysing the data where the vector data consists of tips correlate pairs to indicate a physical location in the world. These points can be joined in a particular way to form closed areas marked as quadrilaterals. Vector data is extremely useful for storing and representing data that has discrete boundaries such as international borders, streets, buildings, and many more. Modern technologies such as Google use geological information and open street maps to represent the data in vector data stricture wise.

Keywords

Data, Information System, Security, Spatial Data Mining, Vector Data.

INTRODUCTION

Spatial data mining is used to discover potentially useful data, interesting and non-tribal patterns from spatial datasets. The aim of this study is to understand and verify the data mining process in relation to the collection and analysis of the real-time data obtained from a particular research field. Computational data management is mostly used for this purpose since it gives the actual real-time data. Spatial data has unique specialisations like spatial autocorrelation which involves the presence of systematic spatial variation in a given set of variables. Another feature of spatial data mining involves spatial heterogeneity which violates independent and similarly distributed presumption of old statistic and data mining procedures. Advancement in this strategy of data mining has led to additional challenges like a modifiable areal unit problem, focusing on the primary spatial pattern facilities the following are considered as one of the efficient ways for collecting the spatial data. These are hotspot detection, collocation, spatial prediction, and spatial outlier detection.

Process of spatial data mining

Spatial data mining relates to the process of discovering interesting and previously unknown datasets. These datasets are potentially very useful obtained from large data sets. This requires specific techniques and resources to get the specific geographical data into the relevant and useful formats.

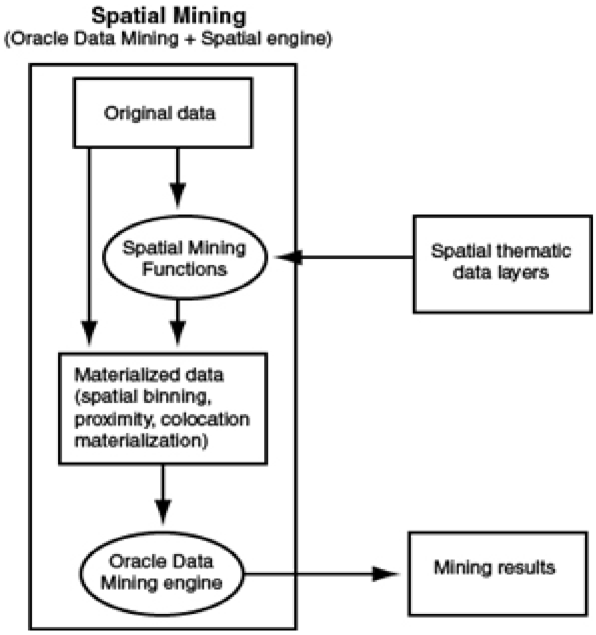

Spatial Mining
(Oracle Data Mining + Spatial engine)
Original data
Spatial Mining Functions
Spatial thematic data layers
Materialized data (spatial binning, proximity, colocation materialization)
Oracle Data Mining engine
Mining results

Figure 1: Factors on which spatial mining of data depends (Source: Zhou *et al.* 2018)

From the above-given diagram it is clear that the process of finding the spatial data depends upon several factors there are:

1. The original data includes the spatial as well nonspatial data which is further processed to form the materialized data. spatial data is any type of information that is directly or indirectly related to a specific geographic location. This kind of data mainly involves point locations or more complicated

objects such as countries, roads and waterbodies.SQL is the most commonly used server in the case of spatial data management (Zhou *et al.* 2018). SQL provides support to two spatial data types which include the geometry data type and geography data type. Spatial data types generally use two models. These are the vector data model and raster data model. Non- spatial data involves the objects which are not involved or related in the perception of a relationship. There are four types of non-spatial data there are integer, float or real, text, and date. The major advantage of materializing the data since it is easily retriable at a faster rate. The data is precomputed and stored at the expense of insert or deleted data. The database will keep the materialized view in synchronization with the real data obtained.

2. Spatial data in original data is prepared by vector data and raster data model to prepare the materialized data. These procedures of vector & raster data models involve the following steps. These are operations such as spatial binning, proximity, and colocation materialization.

3. The oracle data mining engine processes the materialized data which includes the spatial and nonspatial data to generate the specific data mining results.

Benefits and challenges of spatial data mining for discovering unknown dataset 250

The benefits of the data mining process can be stated as below:

1. To create different business models, marketing companies use data mining processes. This can be done by analysing the response from the customers sent via advertisement. This means that the marketers can sell profitable products to specific customers [6].
2. Helping the financial institutions with information on loans, credit reports as well as by creating a model related to the history of dealing with existing customers. Thus, it is helping the bank to detect fraudulent transactions.

Create different business models, marketing companies
Helping the financial institutions
Motivating the researchers
Increase brand loyalty
Increase the company revenue

Figure 2: Benefits of spatial data mining
(Source: Zhou *et al.* 2018)

3. Data mining is motivating the researchers to accelerate the method of analysing the data according to the preferred mindset of the customers and specific geographic location regarding the demand and supply rate of any particular product launched in the market [7].
4. Data mining can be used to increase brand loyalty in business by understanding customers' needs and habits regarding the choice for any specific data.
5. It can be used to increase the company revenue which reduces the cost and services of the product by collecting the information needed regarding the processing of the raw materials. Thus, increasing the profitability of a company to around 80% [11].

Challenges faced spatial data mining for discovering unknown data:

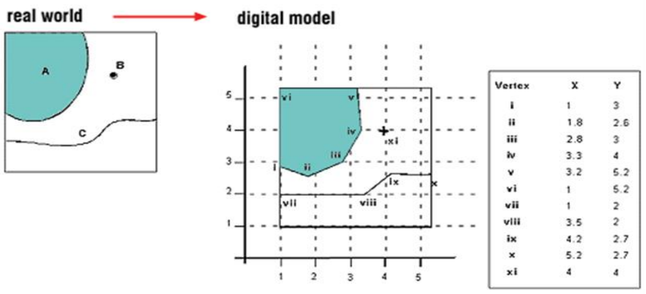
1. The concern about personal privacy is one of the major challenges faced by data mining. This is because people are afraid that their personal information is collected in an unethical way. For example, the business relationship between the organization and the customers may not last forever and some hackers take the advantage of this situation because the data of the customers are stored along with the information of their monetary transaction which involves the bank details of every customer. Hackers can use this information and can badly affect the financial conditions of customers [8].
2. Security issues have been another major problem of data mining since the information of the employees as well the customers including the social security number, date of birth, and much more information are provided during the survey of data mining. This information goes to the hands of the hackers and fraudsters when they hack specific e-commerce sites which becomes a beneficial factor for them to utilize [9].

METHODS AND TECHNIQUES

This research is based on ***secondary qualitative data*** which has been collected from several reliable sources. For this research, a lot of ***scholarly articles*** were collected, and based on that the results are described in this study. According to the opinion of [1], directional data consists of lines joined with the endpoints which meet at the junction. The vectorial system consists of two factors: the one which merges with the spatial data and the one which manages the thematic data. The directional system acts as a hybrid model since it bridges the relationship between the databases for the characteristics with a geological dataset for the spatial data. The identifier acts as a major element in such cases since it provides different results for each specific item which helps the machine to connect to both the databases. In a directional-based system, the calculation of the geological data is represented in the form of correlation and the building blocks of geographical information are given as border, angle, and quadrilateral shape. Therefore, more details regarding this are explained below.

RESULTS AND DISCUSSION

The directional data system is generally used to store non-linear data. Directional storage involves the storage of clearly expressed topological information that elevates on a high scale. Thus, it only collects the information which defines the purposes and can be stated as non-existent data. There are different ways of the double database, involving the geographic and schematic data. A vector system involves two components: the one that operates the geographical data and the one that manages the schematic data. This results in the formation of a dual system since it bridges links the relationship between the database for the characteristics with one geographical data with another geographical data. The identifier acts as the key element for such cases since the identifier is unique and different for each object which allows the system to get connected to both the databases [5].



Vertex	X	Y
i	1	3
ii	1.8	2.6
iii	2.8	3
iv	3.3	4
v	3.2	5.2
vi	1	5.2
vii	1	2
viii	3.5	2
ix	4.2	2.7
x	5.2	2.7
xi	4	4

Figure 3: Graph representing the different elements of the vector data system [2]

A directional data system defines the distinctive substances. Some examples of the distinctive substances are fire hydrants, roads, and water bodies. A directional data system is broken down into three basic types. These are the points, lines, and polygons and all these three types of vector data are composing the coordinates and attributes. A tip is used as a single correlates pair to define a specific geographic location. The points don't have any specific dimensions. They can have actual real-world dimensions. In the case of a geographic information system, no dimensions are assumed. Each point has associated tribute information and the information is attached to the centre of the point. A line vector type is defined by an ordered set of coordinates. According to [3] Each line and curve is composed of multiple line segments and the curved lines are always represented mathematically. The line is composed of two components that are the nodes and the vertex. A vertex is defined where the line changes its direction. A node is defined as where a line begins or ends. The smallest border will have two nodes a start junction and an end node whereas the longer borders will have two junctions and many vertices in between where the line changes its direction. Examples of geological phenomena that are represented well by junctions are roads, pipelines, outlines of different objects, and power lines. For example, if a line represents a road, each road segment between two intersections may have its address information. The intersection can represent a stop sign or a spotlight [10].

The last directional information type is the quadrilateral that is formed by a set of connected lines where both ends have the same coordinates. Since the start and the endpoints have the same coordinates, the quadrilateral will get enclosed and will have a curved space. Attribute information is attached to the centre of the quadrilateral irrespective of the complexity of the quadrilateral. Some examples of geographical phenomena designed with quadrilaterals are lakes, cities, tree stands, and different political boundaries [4].

Vector type data are usually involved in locating the specific geographical infrastructure with the help of specific signs and symbols.

CONCLUSION

Spatial data mining has an important contribution to Geographic Information systems that provides the ability to capture and analyse the spatial and topographic data of any specific location around the world. It acts as a mathematical model for representing geographical objects and landmarks as data types. Vector type data are usually involved in locating the specific geographical infrastructure with the help of specific signs and symbols. Thus, helping the geomorphologist to survey their particular research regarding any geographic location by only optimizing the vector type data in their research of interest. Data mining is motivating the researchers to accelerate the method of analysing the data according to the preferred mindset of the customers and specific geographic location regarding the demand and supply rate of any particular product launched in the market.

REFERENCES

[1] Arabameri, A., Pradhan, B., Rezaei, K., Sohrabi, M. and Kalantari, Z., 2019. GIS-based landslide susceptibility mapping using numerical risk factor bivariate model and its ensemble with linear multivariate regression and boosted regression tree algorithms. *Journal of Mountain Science*, *16*(3), pp.595-618.

[2] Atluri, G., Karpatne, A. and Kumar, V., 2018. Spatio-temporal data mining: A survey of problems and Chen, W., Zhang, S., Li, R. and Shahabi, H., 2018. Performance evaluation of the GIS-based data mining techniques of the best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling. *Science of the total environment*, *644*, pp.1006-1018.methods. *ACM Computing Surveys (CSUR)*, *51*(4), pp.1-41.

[3] Bakhshinategh, B., Zaiane, O.R., ElAtia, S. and Ipperciel, D., 2018. Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, *23*(1), pp.537-553.

[4] Dutt, A., Ismail, M.A. and Herawan, T., 2017. A systematic review on educational data mining. *Ieee Access*, *5*, pp.15991-16005.

[5] Joseph, S.I.T. and Thanakumar, I., 2019. Survey of data mining algorithms for intelligent computing system. *Journal of trends in Computer Science and Smart technology (TCSST)*, *1*(01), pp.14-24.

[6] Lee, S., Lee, M.J. and Jung, H.S., 2017. Data mining approaches for landslide susceptibility mapping in Umyeonsan, Seoul, South Korea. *Applied Sciences*, *7*(7), p.683.

[7] Maneiro, R., Amatria, M. and Anguera, M.T., 2019. Dynamics of Xavi Hernández's game: a vectorial study through polar coordinate analysis. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, *233*(3), pp.389-401.

[8] Naghibi, S.A., Moghaddam, D.D., Kalantar, B., Pradhan, B. and Kisi, O., 2017. A comparative assessment of GIS-based data mining models and a novel ensemble model in groundwater well potential mapping. *Journal of Hydrology*, *548*, pp.471-483.

[9] Prabakaran, S. and Mitra, S., 2018, April. Survey of analysis of crime detection techniques using data mining and machine learning. In *Journal of Physics: Conference Series* (Vol. 1000, No. 1, p. 012046). IOP Publishing.

[10] Tang, C., He, Y., Zhou, G., Zeng, S. and Xiao, L., 2018. Optimizing the spatial organization of rural settlements based on life quality. *Journal of Geographical Sciences*, *28*(5), pp.685-704.

[11] Zhao, X. and Chen, W., 2020. GIS-based evaluation of landslide susceptibility models using certainty factors and functional trees-based ensemble techniques. *Applied Sciences*, *10*(1), p.16.