ANNQUEST DECEMBER 2018
ISSN: 2321-3043
www.stannsannquest.com

Spatial Mining – Issues and Challenges

Aparna Bulusu
Department of Computer Science
St Ann's College for Women
Mehdipatnam

Abstract— Spatial mining can be considered as an offshoot of data mining with certain similarities as well as a number of features that make it very unique and complex. The most important characteristic of spatial mining is that the data looks considerably different from regular data sets used in traditional data mining. The developments in geographical information systems has led to an increase in the amount of spatial data being collected and analyzed and has led to the development of many new applications like predicting disease outbreaks, climate changes, crime hotspots etc. While the fundamental techniques used for mining spatial data have evolved from data mining, they also need to take into account additional dimensions related to time and space which makes spatial mining much tougher to handle. This paper attempts to provide an overview of spatial mining, with a special focuson the challenges faced in spatial pattern recognition.

Keywords— Spatial data mining, Pattern recognition

I. INTRODUCTION

Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases.[1]Spatial mining has become an active research area of late due to the vast technological advancements in areas related to Geographic Information Systems. A Geographic Information Systems (GIS) is a system designed to capture, store, manipulate and analyze various kinds of spatial data. GIS systems use location as the key index variable. Locations or extents in the Earth space–time are recorded either as dates/times of occurrence, and x, y, z co-ordinates representing longitude, latitude, and elevation, respectively. All earth-based spatial–temporal location and extent references must finally map to a real physical location.

Data collection/Representation in GIS:

GIS technologies use digital information, for which various digitized data creation methods are used.[2] The most common method of data creation is digitization, where a physical image is transferred into a digital medium through the use of a CAD program and geo-referencing capabilities. GIS uses spatio-temporal (space-time) location as the key index variable for all other information. Like relational tables use the concept ofprimary keys,GIS systems relate otherwise unrelated information by using location as the key index variable. The key is the location and/or extent in space-time. Any variable that can be located spatially (and also temporally), can be referenced using a GIS. Related by accurate spatial information, an incredible variety of real-world and projected past or future data can be analyzed, interpreted and represented. In developing a digital topographic database for a GIS, topographical maps are the main source, and aerial photography and satellite imagery are extra sources for collecting data and identifying attributes.

Spatial data consists of two distinct types of attributes: Non-spatial attributes and spatial attributes. Non-spatial attributes are used to characterize non-spatial features of objects, such as name, population, house prices, and unemployment rate for a city. They are similar to attributes used in regular data mining. Spatial attributes are used to define the spatial location and extent of spatial objects [3] and include information related to spatial locations, e.g., longitude, latitude and elevation defined in a spatial reference frame, as well as shape.

Spatial datasets are discrete representations of continuous phenomena. GIS systems use two broad methods to store data i.e. raster images and vectors. Satellite images are good examples of raster data. Vector data consists of points, lines, polygons and their aggregate (or multi-) counter parts. A more recent hybrid method of storing data is to use point clouds, which combine three-dimensional points with RGB information at each point, returning a 3D color image. Spatial networks are another important spatial data type

Interpreting spatial data:

Spatial data has positional and topological data that do not exist in general data, and its structure is different according to the kinds of spatial data. Also, the objects on space affect each other and the relationship of objects is also different according to the kinds of objects. Spatial data is usually classified into three basic types:

i. Point referenced data , which is modeled as a fixed collection of spatial locations, S, in a two-dimensional frameworkD (e.g. set of police stations in a metropolitan city);

ii. Areal data, modeled as a finite set of irregular shapedpolygons in a two-dimensional framework D (e.g. set of police districts in a metropolitan city)

iii. Point processdata which is modeled as a random collection of spatial events, collectively referred to as the spatial point patternover a two-dimensional framework D (e.g. home locations of patients infected by a disease).

Typical operations performed on spatial databases:

Spatial databases support a wide variety of spatial operations as specified by the Open Geospatial Consortium standard[4]:

- Spatial Measurements: Computes line length, polygon area, the distance between geometries, etc.
- Spatial Functions: Modify existing features to create new ones, for example by providing a buffer around them, intersecting features, etc.
- Spatial Predicates: Allows true/false queries about spatial relationships between geometries. Examples include "do two polygons overlap" or 'is there a residence located within a mile of the area we are planning to build the landfill?
- Geometry Constructors: Creates new geometries, usually by specifying the vertices (points or nodes) which define the shape.
- Observer Functions: Queries which return specific information about a feature such as the location of the center of a circle

Some special features of geographical/spatial data that make spatial mining different from data mining [5] are as follows: i) The spatial relationships among the variables, ii) The spatial structure of errors, iii) The presence of mixed distributions as opposed to commonly assumed normal distributions, iv) Observations that are not independent and identically distributed, v) Spatial autocorrelation among the features, and vi) non-linear interactions in feature space. While conventional data mining algorithms can be applied to spatial data, they usually perform poorly.

Spatial Mining techniques:

Some of the most common mining tasks performed on spatial data include figuring out co-location patterns, finding classification and regression models to fit spatial data , spatial clustering and detecting spatial outliers

- **Spatial classification***:* Spatial classification methods extend the general-purpose classificationmethods to consider not only attributes of the object to be classified but also the attributes of neighboring objects and their spatial relations [6] .In spatial classification, the objects are classified according to both spatial and non-spatial attributes . Spatial classification makes extensive use of decision trees. A big difference between the spatial classification and regular classification is that the aggregation value of the spatial objects in a nearby region is used in spatial classification. Artificial neural networks (ANN) are also being used for a broadvariety of problems in spatial analysis. Remote sensing is one of the major areasthat commonly use classification methods to classify image pixels into labeled categories.

- **Spatial association rule mining**: Association rule mining was originally intended to discover interesting associations or rules between items in large transaction databases[7] Similar to the mining of association rules in transactional or relational databases, spatial association rules also can be mined in spatial databases by considering spatial properties and predicates [8]. A spatial association rule is expressed in the form A) B [s%, c%], where A and B are sets of spatial or non-spatial predicates, s% is the support of the rule, and c% is the confidence of the rule. Various spatial predicates (e.g. close to, far away, intersect, overlap, etc.) can be used in spatial association rules. However spatial association mining is computationally expensive as huge number of spatial predicates need to be considered in deriving association rules from a large spatial datasets. Also most of the generated rules maybe trivial or obvious. Domain knowledge is needed to filter out trivial rules and focus only on new and interesting findings.

- **Spatial pattern recognition / co location pattern**: A prime example of spatial patterns is co-occurrence patterns, which represent subsets of spatial features whose instances are often located in close geographic proximity.). Colocation rule discovery is a process to identify colocation patterns from large spatial datasets with a large number of Boolean features. Spatial co-location pattern mining is similar to, but technically very different from, association rule mining[9] . Given a dataset of spatial features and their locations, a co-location pattern represents subsets of features frequently located together, such as a certain species of animals that tend to habitat within a certain climate zone. A user-specified neighborhood is needed as a container to check which features co-locate in the same neighborhood as spatial data cannot be as easily understood as transactional data. Colocation

mining algorithm on spatial data is more expensive than the apriori algorithm for classical association rule mining.

- **Spatial Clustering:** Spatial clustering is a process of grouping a set of spatial objects into clusters so that objects within a cluster have high similarity in comparison to one another ,but are dissimilar to objects in other clusters. Cluster analysis is carried out on spatial attributes, by making use of similarities between spatial data points like Euclidean or Manhattan distances.[10]Spatial clustering can be applied to group similar spatial objects together; the implicit assumption is that patterns in space tend to be grouped rather than randomly located. However, the statistical significance of spatial clusters should be measured by testing the assumption in the data. For eg, NASA earth observation systems generate a large sequence of global snapshots of the earth, that include various atmospheric, land, and ocean measurements such as sea surface temperature ,pressure, precipitation, and net primary production, Attributes that have high correlation are used to retrieve interesting relationships among spatial objects of earth science data, like identifying the land locations whose climate is severely affected by El Nino. Three types of clustering analysis are usually employed, spatial clustering(i.e., clustering of spatial points), regionalization (i.e., clustering with geographic contiguity constraints), and point pattern analysis(i.e., hot spot detection with spatial scan statistics.

- **Spatial Outlier detection**:

Outliers are defined as observations in a data set which appear to be inconsistent with the remainder of that set of data [11], or which deviate so much from other observations as to arouse suspicions that they were generated by a different mechanism. A spatial outlier [12] is a spatially referenced object whose non-spatial attribute values differ significantly from those of other spatially referenced objects in its spatial neighborhood. Spatial outlier detection is based on comparing the non- spatial attributes of spatially related objects. Attributes like location, neighborhood and distance are used along with the principle of spatial co relation to determine spatially related objects and then the non-spatial attributes are compared to detect outliers. Certain measures used for the purpose of spatial outlier detection include Variogram clouds and Moran scatter plots.

Applications

Spatial mining has widespread applications over many areas. The application domains include transportation, ecology, homeland security, public health, climatology, and location-based services. Analysis of spatial data through use of pattern recognition and machine learning algorithms is helping in identifying poverty maps across the world. Understanding and predicting climate changes is also an extremely important application. Climate modeling is now being carried out on climate data to predict and assess climate changes as well as help with policy making. Spatial analysis is being used for monitoring changes in crop types and yields and also to predict dependency on fossil fuels and energy emissions. A massive amount of video surveillance data is now available which is being mined for detection of crime hotspots. . GIS and location intelligence applications can be the foundation for many location-enabled services that rely on analysis and visualization. Ambient geographic information processed from social media feeds is another rich source of information having a multitude of location based applications. Mobility services including routing and navigational services also make use of spatial data. Spatial clustering and outlier detection is helping track diseases and crimes.

Conclusion

Spatial data mining plays an important role in extracting interesting spatial patterns and features, capturing relationships between spatial and non-spatial data and presenting data regularity concisely at various levels of granularity. GIS data sourced from high-resolution digital terrain and aerial imagery, coupled with powerful computers and Web technology is changing the quality, utility, and expectations of GIS systems. However, extracting interesting and useful patterns from spatial datasets is more difficult than extracting corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation which increases computational complexity. Because of the huge amounts of spatial data being obtained from satellite images, medical equipment, 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 and help in extracting interesting spatial patterns and features. Mining spatial data has given rise to new applications but also poses several challenges. New approaches are required to overcome the computational complexity and process very large amount of data. Also new models are required that can handle the spatial and temporal constraints efficiently.

REFERENCES

[1] P. Stolorz, H. Nakamura, E. Mesrobian, R. Muntz, E. Shek, J. Santos, J. Yi, K. Ng, S. Chien, R. Mechoso, and J. Farrara. Fast Spatio-Temporal Data Mining of Large Geophysical Datasets. In

Proceedings of the First International Conference on Knowledge Discovery and Data Mining, AAAI Press, 300-305, 1995.

[2] https://en.wikipedia.org/wiki/Geographic_information_system

[3] P. Bolstad. GIS Fundamentals: A First Text on GIS. Eider Press, 2002

[4] Botts, Mike, et al. "OGC® sensor web enablement: Overview and high level architecture." International conference on GeoSensor Networks. Springer Berlin Heidelberg, 2006.

[5] Shekhar, Shashi, et al. "Trends in spatial data mining." Data mining: Next generation challenges and future directions (2003): 357-380

[6] Ester, Martin, Hans-Peter Kriegel, and Jörg Sander. "Knowledge discovery in spatial databases." Mustererkennung 1999. Springer Berlin Heidelberg, 1999. 1-14.

[7] Agrawal, R., Imielinski, T., Swami, A. (1993). Mining association rules between setsof items in large databases. In ACM SIGMOD international conference onmanagement of data (pp. 207–216).

[8] Appice, A., Ceci, M., Lanza, A., Lisi, F. A., &Malerba, D. (2003). Discovery of spatialassociation rules in geo-referenced census data: A relational mining approach.

[9] Shekhar, Shashi, and Yan Huang. "Discovering spatial co-location patterns: A summary of results." International Symposium on Spatial and Temporal Databases. Springer Berlin Heidelberg, 2001.

[10] Shekhar, Shashi, Pusheng Zhang, and Yan Huang. "Spatial data mining."Data mining and knowledge discovery handbook. Springer US, 2009. 837-854.

[11] Barnett and T. Lewis. Outliers in Statistical Data. John Wiley, 3rd edition edition, 1994.

[12] S. Shekhar, C. Lu, and P. Zhang. Graph-based Outlier Detection : Algorithms and Applications (A Summaryof Results). In Proc. of the Seventh ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining,2001.