Knowledge Discovery in Spatial Databases

Jyoti Jadhav
Shah & Anchor Kutchhi College of Engineering,
Chembur, Mumbai
Jyoja_2006@yahoo.com

Jyoti Sawant
Lokmanya Tilak College of Engineering,
Koparkhairane, Navi Mumbai
Jyotis85@rediffmail.com

Abstract: The number and the size of spatial databases, e.g. for geomarketing, traffic control or environmental studies, are rapidly growing which results in an increasing need for spatial data mining. Spatial data sets are at the heart of a variety of scientific and engineering domains, from astrophysics to computational fluid dynamics to robotics. Rapid advances in simulation and experimentation in these domains are yielding an increasing reliance on efficient and effective spatial reasoning algorithms. New applications in other domains, such as aerodynamics, scientific computing, RFID, sensor and actuator networks, and structural bioinformatics, are additionally being cast in terms of mining and reasoning about spatial data. These developments demand effective cross-fertilization and consolidation of computational techniques from qualitative reasoning, data mining, scientific computing, and statistical methodology, in the context of significant applications.

Keywords: Data mining, Spatial Data Mining, Spatial database systems, thematic maps

1 Introduction

The computerization of many business and government transactions and the advances in scientific data collection tools provide us with a huge and continuously increasing amount of data. This explosive growth of databases has far outpaced the human ability to interpret this data, creating an urgent need for new techniques and tools that support the human in transforming the data into useful information and knowledge. *Knowledge discovery in databases* (*KDD*)[1][4][8] has been defined as the non-trivial process of discovering valid, novel, and potentially useful, and ultimately understandable patterns from data . The process of KDD is interactive and iterative, involving several steps. In particular, *data mining* is the step of applying appropriate algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data.

Data mining is the automated process of discovering patterns in data. The purpose is to find correlation among different datasets that are unexpected. Supermarket chains are a prime example of entities that use data mining techniques in an effort to increase sales by trying to find correlations in consumer buying practices. In a hypothetical situation, a data miner might find a pattern that people who purchase high-end cat food also are strong purchasers of floor wax. As a result of this analysis, the supermarket might then place the pet food products in the same aisle as the household cleaners in an attempt to induce higher sales.

On-Line Transaction Processing (OLTP) is the tradional model for enterprise data processing. In OLTP, the emphasis is on transactions involving the input, update, and retrieval of data. On-Line Analytical Processing (OLAP) applications query the database to collate, summarize, and analyze its contents. Data mining augments the OLAP process by applying artificial intelligence and machine learning techniques to find previously unknown or undiscovered relationships in the data. This is different from analytical techniques in which the goal is to prove or disprove an existing hypothesis.

2 Spatial mining

Spatial data are data that have a spatial or location component. Spatial data can be viewed as data about objects that themselves are located in a physical space.

Spatial mining is data mining as applied to spatial databases or spatial data. Some of the applications for spatial data mining are in the areas of GIS systems, geology, environmental science, resource management, agriculture, medicine, and robotics[1][2][3][7].

2.1 Spatial data overview:

Accessing spatial data can be more complicated than accessing nonspatial data. There are specialized operations and data structures used to access spatial data.

Spatial data structures

A common technique used to represent a spatial object is by the smallest rectangle that completely contains that object, minimum bounding rectangle(MBR).

One benefit of the spatial data structures is that they cluster objects based on location. This implies that objects that are close together in the n-dimensional space tend to be stored close together in the data structure and on disk. Thus, these structures could be used to reduce the processing overhead of an algorithm by limiting its search space.

2.1.1 Quad tree

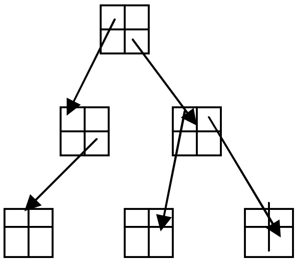

Fig: A Quad tree

A quad tree represents a spatial object by a hierarchical decomposition of the space into quadrants(cells). This process is shown by using the triangle. Triangle is shown as three shaded squares. The area has been divided into two layers of quadrant divisions. The number of layers needed depends on the precision desired.
Each level in the quad tree corresponds to one of the hierarchical layers. Each of the four quadrants at that layer has a related pointer to a node at the next level if any of the lowest level quadrants are shaded[15].

2.1.2 R-tree

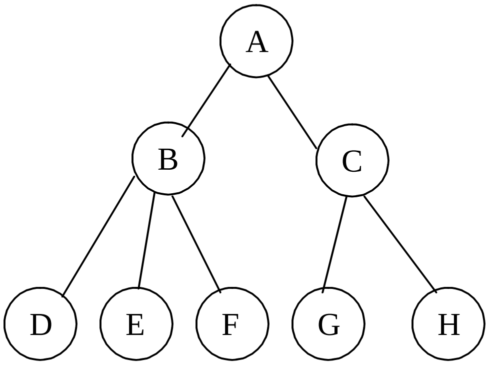

Fig: R-tree

One approach to indexing spatial data represented as MBRs is an R-tree. Each successive layer in the tree identifies smaller rectangles. In an R-tree, cells may actually overlap. An object is represented by an MBR that is located within one cell. A cell is the MBR that contains the related set of objects (or MBRs) at a lower level of decomposition. Each level of decomposition is identified with a layer in the tree. As spatial objects are added to the R-tree, it is created and maintained by algorithms similar to those found for B-trees. The size of the tree is related to the number of objects.

Algorithms to perform spatial operators using an R-tree are relatively straightforward. Suppose we wished to find all objects that intersected with a given object. Representing the query object as an MBR, we can search the upper levels of the R-tree to find only those cells that intersect the MBR query.Those subtrees that do not intersect the query MBR can be discarded[11].

2.1.3 k-D Tree

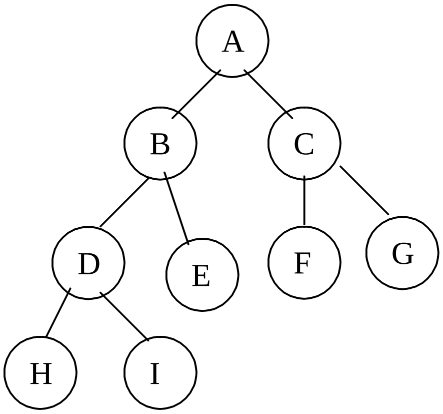

Fig: k-D-tree

A k-D tree was designed to index multiattribute data, not necessarily spatial data. The k-D tree is a variation of a binary search tree where each level in the tree is used to index one of the attributes. We illustrate the use of the k-D tree assuming a two-dimensional space. Each node in the tree represents a division of the space into two subsets based on the division point used.

Each lowest level cell has only one object in it. The division are not made using MBRs. Initially, the entire region is viewed as one cell and thus the root of the k-D tree. The area is divided first along one dimension until each cell has only one object in it[15].

2.2 Data Input

The data inputs of spatial data mining are more complex than the inputs of classical data mining because they include extended objects such as points, lines, and polygons. The data inputs of spatial data mining have two distinct types of attributes: non-spatial attribute and spatial attribute. Non-spatial attributes are used to characterize non-spatial features of objects, such as name, population, and unemployment rate for a city. They are the same as the attributes used in the data inputs of classical data mining. Spatial attributes are used to define the spatial location and extent of spatial objects. The spatial attributes of a spatial object most often include information related to spatial locations, e.g., longitude, latitude and elevation, as well as shape. Relationships among non-spatial objects are explicit in data inputs, e.g., arithmetic relation, ordering, is instance of, subclass of, and membership of. In contrast, relationships among spatial objects are often implicit, such as overlap, intersect, and behind. One possible way to deal with implicit spatial relationships is to materialize the relationships into traditional data input columns and then apply classical data mining techniques. However, the materialization can result in loss of information. Another way to capture implicit spatial relationships is to develop models or techniques to incorporate spatial information into the spatial data mining process.

2.3 Statistical Foundation

Statistical models are often used to represent observations in terms of random variables. These models can then be used for estimation, description, and prediction based on probability theory. Spatial data can be thought of as resulting from observations on the stochastic process Z(s): s 2 D, where s is a spatial location and D is possibly a random set in a spatial framework. Here we present three spatial statistical problems one might encounter: point process, lattice, and geostatistics[16].

2.3.1 Point process:

A point process is a model for the spatial distribution of the points in a point pattern. Several natural processes can be modeled as spatial point patterns, e.g., positions of trees in a forest and locations of bird habitats in a wetland. Spatial point patterns can be broadly grouped into random or non-random processes.

2.3.2 Lattice:

A lattice is a model for a gridded space in a spatial framework. Here the lattice refers to a countable collection of regular or irregular spatial sites related to each other via a neighborhood relationship. Several spatial statistical analyses, e.g., the spatial autoregressive model and Markov random fields, can be applied on lattice data.

2.3.3 Geostatistics:

Geostatistics deals with the analysis of spatial continuity and weak stationarity, which is an inherent characteristics of spatial data sets. Geostatistics provides a set of statistics tools, such as kriging to the interpolation of attributes at unsampled locations.

3. Output Patterns

There are four important output patterns for spatial data mining: predictive models, spatial outliers, spatial co-location rules, and spatial clustering[16].

3.1 Predictive Models

The prediction of events occurring at particular geographic locations is very important in several application domains. Examples of problems which require location prediction include crime analysis, cellular networking, and natural disasters such as fires, floods, droughts, vegetation diseases, and earthquakes. In this section we provide two spatial data mining techniques for predicting locations, namely the Spatial Autoregressive Model (SAR) and Markov Random Fields (MRF).

3.2 Spatial Outliers

Outliers have been informally defined as observations in a dataset which appear to be inconsistent with the remainder of that set of data or which deviate much from other observations. The identification of global outliers can lead to the discovery of unexpected knowledge and has a number of practical applications in areas such as credit card fraud, athlete performance analysis, voting irregularity, and severe weather prediction.

Detecting spatial outliers is useful inmany applications of geographic information systems and spatial databases, including transportation, ecology, public safety, public health, climatology, and location-based services.

A spatial outlier is a spatially referenced object whose nonspatial attribute values differ significantly from those of other spatially referenced objects in its spatial neighborhood. Informally, a spatial outlier is a local instability or a spatially referenced object whose non-spatial attributes are extreme relative to its neighbors, even though the attributes may not be significantly different from the entire population[16].

3.3 Spatial Co-location Rules

Boolean spatial features are geographic object types which are either present or absent at different locations in a two dimensional or three dimensional metric space, e.g., the surface of the Earth. Examples of boolean spatial features include plant species, animal species, road types, cancers, crime, and business types.

Co-location patterns represent the subsets of the boolean spatial features whose instances are often located in close geographic proximity. Examples include symbiotic species, e.g., Nile crocodile and Egyptian plover in ecology, and frontage roads and highways in metropolitan road maps.

3.4 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. For example, clustering is used to determine the hot spots in crime analysis and disease tracking. Hot spot analysis is the process of finding unusually dense event clusters across time and space. Many criminal justice agencies are exploring the benefits provided by computer technologies to identify crime hot spots in order to take preventive strategies such as deploying saturation patrols in hot spot areas.

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. The test is critical before proceeding with any serious clustering analysis.

4. Algorithms in Spatial Data mining

4.1 Spatial Classification Algorithms

Spatial classification problems are used to partition sets of spatial objects. Spatial objects could be classified using nonspatial attributes, spatial predicates(spatial attributes), or spatial and nonspatial attributes. Concept hierarchies may be used, as may sampling. Generalization and progressive refinement techniques may be used to improve efficiency[1].

4.1.1 ID3 Extension

The concept of neighborhood graphs has been applied to perform classification of spatial objects using ID3 extension[3].

4.1.2 Spatial Decision Tree

one spatial classification technique builds decision trees using a two-step process similar to that used for association rules. The basis of the approach is that spatial objects can be described based on objects close to them. A description of the classes is then assumed to be based on an aggregation of the most relevant predicates for objects nearby[14].

4.2. Spatial Clustering Algorithms

Spatial clustering algorithms must be able to work efficiently with large multidimensional databases. In addition , they should be able to detect clusters of different shapes. Other desirable features for spatial clustering are that the clusters found should be independent of the order in which the points in the space are examined and that the clusters should not be impacted by outliers.

4.2.1 CLARANS Extensions

The main memory assumption of CLARANS is totally unacceptable for large spatial databases. Two approaches to improve the performance of CLARANS by taking advantage of spatial indexing structures have been proposed[13].

The first approach uses a type of sampling based on the structure of an R*-tree. To ensure the quality of the sampling, the R*-tree is used to guarantee that objects from all areas of the space are examined. The second technique improves on the manner in which the cost for a prominent change is calculated. Instead of examining the entire databases, only the objects in the two affected clusters must be examined. A region query can be used to retrieve the needed objects.

4.2.2 SD(CLARANS)

Spatial dominant CLARANS[SD(CLARANS)] assumes that items to be clustered contain both spatial and nonspatial components. It first clusters the spatial components using CLARANS and then examines the nonspatial attributes within each cluster to derive a description of that cluster**.**

4.2.3 DBCLASD

DBCLASD (Distribution Based Clustering of Large Spatial Databases) assumes that the items within a clusters are uniformly distributed and that points outside the cluster probably do not satisfy this restriction. Based on this assumption, the algorithm attempts to identify the distribution satisfied by the distances between nearest neighbors.

4.2.4 BANG

The BANG approach uses a grid structure similar to k-D tree. The structure adapts to the distribution of the items so that more dense areas have a larger number of smaller grids, while less dense areas have a few large ones.

4.2.5 WaveCluster

WaveCluster can be find arbitrarily shaped clusters and does not need to know the desired number of clusters. A wavelet transform is used as a filter to determine the frequency content of the signal.

A wavelet transform of a spatial object decomposes it into a hierarchy of spatial images. They can be used to scale an image to different sizes.

4.2.6 Approximation

Approximation can be used to identify the characteristics of clusters. This is done by determining the features that are close to the clusters. Clusters can be distinguished based on features unique to them or that are common across several clusters.

5. Thematic Maps

The thematic maps illustrate spatial objects by showing the distribution of attributes or themes. Each map shows one (or more) of the thematic attributes. These attributes describe the important nonspatial features of the associated spatial object. For example, one thematic map may show elevation, average rainfall, and average temperature. Raster based thematic maps represent the spatial data by relating pixels to attribute values of the data. For example, in a map showing elevation, the color of the pixel can be associated with the elevation of that location. A vector based themativ map represents objects by a geometric structure. In addition, the object then has the thematic attribute values[15][16].

6. Spatial Database Systems (SDBS)

These are database systems for the management of spatial data. To find implicit regularities, rules or patterns hidden in large spatial databases, e.g. for geo-marketing, traffic control or environmental studies, spatial data mining algorithms are very important. Attribute-oriented induction can be performed by using (spatial) concept hierarchies to discover relationships between spatial and non-spatial attributes.

In the clustering algorithm CLARANS, which groups neighboring objects automatically without a spatial concept hierarchy, is combined with attribute-oriented induction on non-spatial attributes introduces spatial association rules which describe associations between objects based on different spatial neighborhood relations. There exist algorithms to detect properties of clusters using reference maps and thematic maps. For instance, a cluster may be explained by the existence of certain neighboring objects which may “cause” the existence of the cluster. For spatial classification it is important that class membership of a database object is not only determined by its non-spatial attributes but also by the attributes of objects in its neighborhood.

In spatial trend analysis, patterns of change of some non-spatial attribute(s) in the neighborhood of some database object are determined. We argue that data mining algorithms should be integrated with existing DBMS, i.e. they should not run on separate files but they should run directly on a database. Thus, redundant storage and potential inconsistencies can be avoided. Furthermore, the query operations provided by a DBMS may be used, for example, to select subsets relevant for data mining or to support the user in evaluating the discovered patterns. In this paper, we introduce a set of database primitives for mining in spatial databases. These primitives are sufficient to express most of the algorithms for spatial data mining from the literature; in particular they can express the algorithms reviewed above. We present techniques for efficiently supporting these primitives by a DBMS[1][14][16].

7. Future research

A further research topic includes the Comparison of classical data mining techniques with spatial data mining techniques. It is possible to materialize the implicit relationships into traditional data input columns and then apply classical data mining techniques.

Another way to deal with implicit relationships is to use specialized spatial data mining techniques, e.g., the spatial auto regression and co-location mining. New research is needed to compare the two sets of approaches in effectiveness and computational efficiency. Modeling semantically rich spatial properties, such as topology Statistical interpretation models for spatial patterns.

Spatial connectivity and other complex spatial topological relationships in spatial networks are difficult to model using the continuity matrix. Research is needed to evaluate the value of enriching the continuity matrix beyond the neighborhood relationship[16].

Effective visualization of spatial relationships improving computational efficiency. To facilitate the visualization of spatial relationships, research is needed on ways to represent both spatial and non-spatial features.

8. Conclusions

Extracting interesting and useful patterns from spatial datasets is more difficult than extracting the corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation. This chapter focuses on the unique features that distinguish spatial data mining from classical data mining.

Future research includes optimized implementation of database primitives in the server of a spatial DBMS and a comparison with current implementation. Filters can be used for restricting the search to neighborhood paths "leading *away*" from a starting object.

References

[1] Margaret H. Dunham, Data Mining

[2] Agrawal R., Imielinski T., Swami A.: "Database Mining: A Performance Perspective", *IEEE Transactions on Knowledge and Data Engineering, Vol. 5, No. 6, 1993.*

[3] Ester M., Kriegel H.-P., Sander J.: "Spatial Data Mining: A Database Approach*", Proc. 5th Int. Symp. on Large Spatial Databases, Berlin, Germany, 1997*

[4] Fayyad U. M., .J., Piatetsky-Shapiro G., Smyth P.: "From Data Mining to Knowledge Discovery: An Overview*", In: Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, 1996*

[5] Koperski K., Adhikary J., Han J.: "Knowledge Discovery in Spatial Databases: Progress and Challenges", *Proc. SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, Technical Report 96-08, University of British Columbia, Vancouver, Canada, 1996.*

[6] Chen M.-S., Han J., and Yu P. S. 1996 "Data Mining: An Overview from a Database Perspective", *IEEE Trans. on Knowledge and Data Engineering, Vol. 8, No. 6,*

[7] Ester M., Frommelt A., Kriegel H.-P., and Sander J. 2000 "Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support", *Data Mining and Knowledge Discovery, an International Journal, Kluwer Academic Publishers, Vol. 4, No. 2/3.*

[8]Fayyad U. M., Piatetsky-Shapiro G., and Smyth P. 1996 "Knowledge Discovery and Data Mining: Towards a Unifying Framework", *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. Portland, Oregon, AAAI Press, Menlo Park, California,*

[9]Gueting R. H. 1994 "An Introduction to Spatial Database Systems", *Special Issue on Spatial Database Systems of the VLDB Journal, Vol. 3, No. 4.*

[10] J.L. Bentley, "Multidimensional binary search trees used for associative searching", *Communications of the ACM, 1975.*

[11] A.Guttman, "R-trees:A dynamic index structure for spatial searching", *Proceedings of the ACM International Conference on management of data, June 1984.*

[12] Beng Chin Ooi, Ron Sacks-Davis, and Jiawei han, " Indexing in spatial databases", *www.comp.nus.edu.sg/ooibc/papers.html, 1993.*

[13]Martin Ester, Alexander Frommelt, Hans-Peter Kriegel and Xiaowei Xu. "Knowledge discovery in large spatial databases:Focusing techniques for efficient class identification". *Proceedings of the Fourth International Symposium on Large spatial Databases(SSD), 1995*.

[14] Krzysztof Koperski and Jiawei Han and Nebojsa Stefanovic. An Efficient two-step method for classification of spatial data", *proceedings of the International Symposium on spatial data handling, Sept.1998.*

[15] R.A.Finkel and J.L.Bentley. "Quad trees:Adata structure for retrieval on compsite keys." , *Acta Inforatica,4(1), 1974.*

[16] Trends in Spatial Data Mining Shashi Shekhar Pusheng Zhang, Yan Huang, Ranga Raju Vatsavai ,Department of Computer Science and Engineering, University of Minnesota.