CHAPTER-24
Mining Spatial Databases

24.Mining Spatial Databases

24.1 Introduction

A spatial database stores a large amount of space-related data, such as maps, preprocessed remote sensing or medical imaging data, and VLSI chip layout data. Spatial databases have many features distinguishing them from relational databases. They carry topological and/or distance information, usually organized by sophisticated, multidimensional spatial indexing structures that are accessed by spatial data access methods and often require spatial reasoning, geometric computation, and spatial knowledge representation techniques.

Spatial data mining refers to the extraction of knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases. Such mining demands an integration of data mining with spatial database technologies. It can be used for understanding spatial data, discovering spatial relationships and relationships between spatial and nonspatial data, constructing spatial knowledge bases, reorganizing spatial databases, and optimizing spatial queries. It is expected to have wide applications in geographic information systems, geomarketing, remote sensing, image database exploration, medical imaging, navigation, traffic control, environmental studies, and many other areas where spatial data are used. A crucial challenge to spatial data mining is the exploration of efficient spatial data mining techniques due to the huge amount of spatial data and the complexity of spatial data types and spatial access methods.

Statistical spatial data analysis has been a popular approach to analyzing spatial data. The approach handles numerical data well and usually proposes realistic models of spatial phenomena. However, it typically assumes statistical independence among the spatially distributed data, although in reality, spatial objects are often inter-related. Moreover, experts having a fair amount of domain knowledge and statistical expertise can only perform most statistical modeling. Furthermore, statistical methods do not work well with symbolic values, or incomplete or inconclusive data, and are computationally expensive in large databases. Spatial data allows the extension of traditional spatial analysis methods by placing minimum emphasis on efficiency, scalability, cooperation with database systems, improved interaction with the user, and the discovery of new types of knowledge.

24.2 Spatial Data Cube Construction and Spatial OLAP

As with relational data, we can integrate spatial data to construct a data warehouse that facilitates spatial data mining. A spatial data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of both spatial and non-spatial data in support of spatial data mining and spatial-data-related decision-making processes. Let's have a look at the following example,

Example There are about 3000 weather probes distributed in British Columbia (BC), each recording daily temperature and precipitation for a designated small area and transmitting signals to a provincial weather station. With a spatial data warehouse that supports spatial OLAP, a user can view weather patterns on a map by month by region, and by different combinations of temperature and precipitation, and can dynamically drill down or roll up along any dimension to explore desire patterns, such as "wet and hot regions in the Fraser Valley in Summer 1999".

There are several challenging issues regarding the construction and utilization of spatial data warehouses. The first challenge is the integration of spatial data from heterogeneous sources and systems. Spatial data are usually stored in different industry firms and government agencies using various data formats. Data formats are not only structure-specific (e.g., raster- vs. vector-based spatial data object-oriented vs. relational models, different spatial storage and indexing structures, etc.), but also vendor-specific (e.g., ESRI, MapInfo, Intergraph, etc.). There has been a great deal of work on the integration and exchange of heterogeneous spatial data, which has paved the way for spatial data integration and spatial data warehouse construction.

The second challenge is the realization of fast and flexible on-line analytic processing in spatial data warehouses. The star schema model is a good choice for modeling spatial data warehouses since it provided a concise and organized warehouse structure and facilitates OLAP operation However, in a spatial warehouse, both dimensions and measures may contain spatial components.

There are three types of dimensions in a spatial data cube:

- A nonspatial dimension contains only nonspatial data .Nonspatial dimensions temperature and precipitation can be constructed for the warehouse since each contains nonspatial data whose generalizations are nonspatial (such as "hot" for temperature. and "wet" for precipitation).
- A spatial-to-nonspatial dimension is a dimension whose primitive-level data are spatial but whose generalization, starting at a certain high level, become nonspatial. For example, the spatial dimension City relays geographic data fit the U.S. map. Suppose that the dimension's spatial representation of, say, Seattle is generalized to the string "pacific-northwest". Although "Pacific-northwest" a spatial concept, its representation is not spatial (since, in our example, it if string).It therefore plays the role of a nonspatial dimension.
- A spatial-to-spatial dimension is a dimension whose primitive level and all its high-level generalized data are spatial. For example, the dimension equ_temperature_region contains spatial data, as do all of its generalizations, such as with regions covering 0-5 degrees (Celsius), 5-10 degrees, and so on.

We distinguish two types of measures in a spatial data cube.

A nonspatial data cube contains only nonspatial dimensions and numerical measures. If a spatial data cube contains spatial dimensions but no spatial measures, its OLAP operations, such as drilling or pivoting, can be implemented in a manner similar to that for nonspatial data cubes.

Of the three measures, area and count are numerical measures that can be computed similarly to that for nonspatial data cubes; regions-map is a spatial measure that represents a collection of spatial pointers to the corresponding regions. Since different spatial OLAP operations result in different collections of spatial objects in region map, it is a major challenge to compute the merges of a large number of regions flexibly and dynamically.

There are at least three possible choices in regard to the computation of spatial measures in spatial data cube construction.

Collect and store the corresponding spatial object pointers but do not perform precomputation of spatial measures in the spatial data cube. This can be implemented by storing, in the corresponding cube cell,, a pointer to a collection of spatial object pointers, and invoking and performing the spatial merge (or other computation) of the corresponding spatial objects, when necessary, on- the-fly. This method is a good choice if only spatial display is required (i.e., no real spatial merge has to be performed), or if there are not many regions to be merged in any pointer collection (so that the on-line merge is not very costly), or if on-line spatial merge computation is fast (recently, some efficient spatial merge methods have been developed for fast spatial OLAP). Since OLAP results are often used for on-line spatial analysis and mining, it is still recommended to precompute some of the spatially connected regions to speed up such analysis.

Precompute and store a rough approximation of the spatial measures in the spatial data cube. These choices good for a rough view or coarse estimation of spatial merge results under the assumption that it requires little storage space. For example, a minimum bounding rectangle (MBR), represented by two points, can be taken as a rough estimate of a merged region. Such a precomputed result is small and can be presented quickly to users. If higher precision is needed for specific cells, the application can either fetch precomputed high-quality results, if available, or compute them on-the-fly.

Selectively precompute some spatial measures in the spatial data cube. This can be a smart choice. The selection can be performed at the cuboid level, that is, either precompute and store each set of mergeable spatial regions for each cell of a selected cuboid, or precompute none if the cuboid is not selected. Since a cuboid usually consists of a large number of spatial objects, it may involve precomputation and storage of a large number of mergeable spatial objects, some of which may be rarely used. Therefore, it is recommended to perform selection at a finer granularity level examining each group of mergeable spatial objects in a cuboid to determine whether such a merge should be precomputed. Decision should be based on the utility (such as access frequency or access priority), sharability of merged regions, and the balanced overall cost of space and on-line computation.

With efficient implementation of spatial data cubes and spatial OLAP, generalization-based descriptive spatial mining, such as spatial characterization and discrimination, can be performed efficiently.

24.3 Spatial Association Analysis

Similar to the mining of association rules in transactional and relational databases, spatial association rules can be mined in spatial databases. A spatial association rule is of the form

A =>B [s%, c%].

Where A and B are sets of spatial or nonspatial predicates, s% is the support of the rule, and c% is the confidence of the rule. For example, the following is a spatial association rule: is a (X, "school") A close to(X, “sports center”) => close to(X, “park”) [0.5%, 80%). This rule states that 80% of schools that are close to sports centers are also close to parks and 0.5% of the data belongs to such a case.

Various kinds of spatial predicates can constitute a spatial association rule. Examples include distance information (such as close to and far-away) topological relations (like intersect, overlap, and disjoint), and spatial orientations (like left of and west of).

Since spatial association mining needs to evaluate multiple spatial relationships among a large number of spatial objects, the process could be quite costly. An interesting mining optimization method called progressive refinement can be adopted in spatial association analysis. The method first mines large data sets roughly using a fast algorithm and then improves the quality of mining in a pruned data set using a more expensive algorithm.

To ensure that the pruned data set covers the complete set of answers when applying the high-quality data mining algorithms at a later stage, an important requirement for the rough mining algorithm applied in the early stage is the superset coverage property: that is, it preserves all of the potential answers. In other words, it should allow a false positive test, which might include some data sets that do not belong to the answer sets, but it should not allow a false negative test, which might exclude some potential answers.

For mining spatial associations related to the spatial predicate close to, we can first collect the candidates that pass the minimum support threshold by applying certain rough spatial evaluation algorithm for example, using a minimum bounding rectangle structure (which registers only two spatial points rather than a set of complex polygons). Then evaluating the relaxed spatial predicate, g close to, which is a generalized c close to covering a broader context that includes close to, touch, and intersect.

If two spatial objects are closely located, their enclosing minimum bounding rectangles must be closely located, matching g close to. However, the reverse is not always true: it the enclosing minimum bounding rectangles are closely located, the two spatial objects may or nay not be located so closely. Thus, the minimum bounding rectangle pruning is a false positive testing tool for closeness: only those that pass the rough test need to be further examined using more expensive spatial computation algorithms. With this preprocessing, only the patterns that are frequent at the approximation level will need to be examined by more detailed and finer, yet more expensive, spatial computation.

24.4 Spatial Clustering Methods

Spatial data clustering identifies clusters, or densely populated regions, according to some distance measurement in a large, multidimensional data set.

24.5 Spatial Classification and Spatial Trend Analysis

Spatial classification analyzes spatial objects to derive classification schemes in relevance to certain spatial properties, such as the neighborhood of a district, highway, or river.

Spatial classification:

Suppose that you would like to classify regions in a province into rich versus poor according to the average family income. In doing so, you would like to identify the important spatial-related factors that determine a region's classification. There are many properties associated with spatial objects, such as hosting a university, containing interstate highways, being near a lake or ocean, and so on. These properties can be used for relevance analysis and to find interesting classification schemes. Such classification schemes may be represented in the form of decision trees or rules.

Spatial trend analysis deals with another issue: the detection of changes and trends along a spatial dimension. Typically, trend analysis detects changes with time, such as the changes of temporal patterns in time-series data. Spatial trend analysis replaces time with space and studies the trend of non spatial or spatial data changing with space. For example, we may observe the trend of changes in economic situation when moving away from the center of a city, or the trend of changes of the climate or vegetation with the increasing distance from an ocean. For such analyses, regression and correlation analysis methods are often applied by utilization of spatial data structures and spatial access methods.

There are also many applications where patterns are changing with both space and time. For example, traffic flows on highways and in cities are both time and space related. Weather patterns are also closely related to both time and space. Although there have been a few interesting studies on spatial classification and spatial trend analysis, the investigation of spatiotemporal data mining is still in its infancy. More methods and applications of spatial classification and trend analysis, especially those associated with time, need to be explored in the future.

24.6 Mining Raster Databases

Spatial database systems usually handle vector data that consist of points, lines, polygons (regions), and their compositions, such as networks or partitions. Typical examples of such data include maps, design graphs, and 3-1.) representations of the arrangement of the chains of protein molecules. However, a huge amount of space-related data are in digital raster (image) forms, such as satellite images, remote sensing data, and computer tomography It is important to explore data mining in raster image databases. Methods for mining raster and image data are examined in the following section regarding the mining of multimedia data.

24.7 Review Questions

1 Explain about Spatial Data Cube Construction and Spatial OLAP

2 Explain about Spatial Association Analysis

3 Explain about Spatial Clustering Methods

4 Explain about Spatial Classification and Spatial Trend Analysis

24.8 References

[1]. Data Mining Techniques, Arun k pujari 1st Edition

[2] .Data warehousung,Data Mining and OLAP, Alex Berson ,smith.j. Stephen

[3].Data Mining Concepts and Techniques ,Jiawei Han and MichelineKamber

[4]Data Mining Introductory and Advanced topics, Margaret H Dunham PEA

[5] The Data Warehouse lifecycle toolkit , Ralph Kimball Wiley student Edition