Spatial Data Mining Techniques

Dr. J.Dhillipan[1] , K.Dhakshnamurthy[2] and D.B.Shanmugam[3]

[1]Asst.Prof.,(S.G) & Head, MCA Department, SRM University, Ramapuram Campus, Chennai
[1]Jd_pan@yahoo.co.in
[2]Assistant Professor, Department of BCA, King Nandhivarman College of Arts & science, Thellar
[2]kdmurthy_arni@yahho.com
[3]Associate Professor, Department of MCA, Sri Balaji Chockalingam Engineering College , Arni
[3]dbshanmugam@gmail.com

ABSTRACT

Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial datasets. 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. The requirements of mining spatial databases are different from those of mining classical relational databases. The spatial data mining techniques are often derived from spatial statistics, spatial analysis, machine learning, and databases, and are customized to analyze massive data sets. In this report some of the spatial data mining techniques have discussed along with some applications in real world.

Index Term— National Imagery and Mapping Agency (NIMA), the National Cancer Institute (NCI), and the United States Department of Transportation (USDOT).

1. INTRODUCTION

The explosive growth of spatial data and widespread use of spatial databases emphasize the need for the automated discovery of spatial knowledge. Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases. The complexity of spatial data and intrinsic spatial relationships limits the usefulness of conventional data mining techniques for extracting spatial patterns. Efficient tools for extracting information from geo-spatial data are crucial to organizations which make decisions based on large spatial datasets, including NASA, the National Imagery and Mapping Agency (NIMA), the National Cancer Institute (NCI), and the United States Department of Transportation (USDOT). These organizations are spread across many application domains including ecology and environmental management, public safety, transportation, Earth science, epidemiology, and climatology.

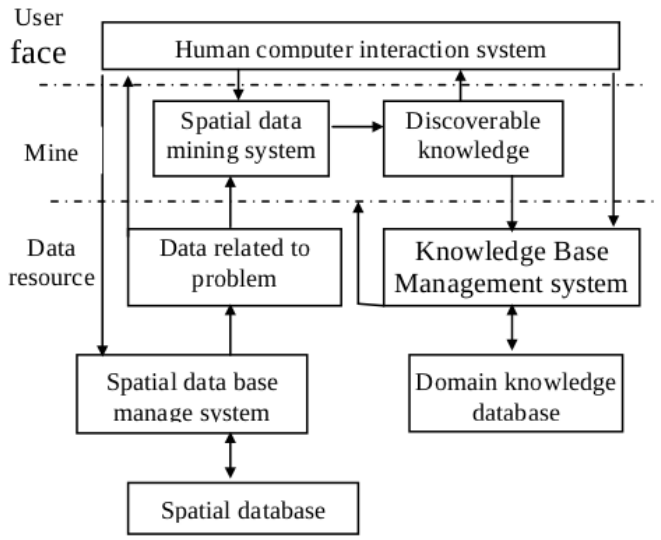
General purpose data mining tools such as Clementine and Enterprise Miner, are designed to analyze large commercial databases. Although these tools were primarily designed to identify customer-buying patterns in market basket data, they have also been used in analyzing scientific and engineering data, astronomical data, multi-media data, genomic data, and web data. Extracting interesting and useful patterns from spatial data sets 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. Specific features of geographical data that preclude the use of general purpose data mining algorithms are: i) rich data types(e.g., extended spatial objects) ii) implicit spatial relationships among the variables, iii) observations that are not independent, and iv) spatial autocorrelation among the features.

The spatial data mining can be used to understand spatial data, discover the relation between pace and the non-space data, set up the spatial knowledge base, excel the query, reorganize spatial database and obtain concise total characteristic etc. The system structure of the spatial data mining can be divided into three layer structures mostly, such as the Fig 1 show .The customer interface layer is mainly used for input and output, the miner layer is mainly used to manage

data, select algorithm and storage the mined knowledge, the data source layer, which mainly includes the spatial database (camalig) and other related data and knowledge bases, is original data of the spatial data mining.

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.



Thesystematic structure of spatial data mining

Knowledge discovery techniques which ignore spatial autocorrelation typically perform poorly in the presence of spatial data. Often the spatial dependencies arise due to the inherent characteristics of the phenomena under study, but in particular they arise due to the fact that the spatial resolution of imaging sensors are finer than the size of the object being observed. For example, remote sensing satellites have resolutions ranging from 30 meters (e.g., the Enhanced Thematic Mapper of the Landsat 7 satellite of NASA) to one meter (e.g., the IKONOS satellite from SpaceImaging), while the objects under study (e.g., Urban, Forest, Water) are often much larger than 30 meters. As a result, per-pixel-based classifiers, which do not take spatial context into account, often produce classified images with salt and pepper noise. These classifiers also suffer in terms of classification accuracy.

The spatial relationship among locations in a spatial framework is often modeled via a contiguity matrix. A simple contiguity matrix may represent a neighborhood relationship defined using adjacency, Euclidean distance, etc. Example definitions of neighborhood using adjacency include a four-neighborhood and an eight-neighborhood. Given a gridded spatial framework, a four neighborhood assumes that a pair of locations influence each other if they share an edge. An eight neighborhood assumes that a pair of locations influence each other if they share either an edge or a vertex.

A Spatial Framework and Its Four-neighborhood Contiguity Matrix

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. Two spatial data mining techniques for predicting locations, namely the Spatial Autoregressive Model (SAR) and Markov Random Fields (MRF).

2. KNOWLEDGE DISCOVERY

Knowledge discovery in databases (KDD) has been defined as the non-trivial process of discovering valid, novel, and potentially useful, and ultimately understand able patterns from data .The process of KDD is interactive and iterative, involving several steps such as the following ones:

• Selection: selecting a subset of all attributes and a subset of all data from which the knowledge should be discovered.

• Data reduction: using dimensionality reduction or transformation techniques to reduce the effective number of attributes to be considered.

• Data mining: the application of appropriate algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data.

• Evaluation: interpreting and evaluating the discovered patterns with respect to their usefulness in the given application.

Spatial Database Systems (SDBS) 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 .Most existing data mining algorithms run on separate and specially prepared files, but integrating them with a database management system (DBMS) has the following advantages.

Redundant storage and potential inconsistencies can be avoided. Furthermore, commercial database systems offer various index structures to support different types of database queries. This functionality can be used without extra implementation effort to speed-up the execution of data mining algorithms (which, in general, have to perform many database queries). Similar to the relational standard language SQL, the use of standard primitives will speed-up the development of new data mining algorithms and will also make them more portable.

2.1 Spatial Classification

The task of classification is to assign an object to a class from a given set of classes based on the attribute values of this object. In spatial classification the attribute values of neighboring objects are also considered.

The classification algorithm works as follows: The relevant attributes are extracted by comparing the attribute values of the target objects with the attribute values of their nearest neighbors. The determination of relevant attributes is based on the concepts of the nearest hit (the nearest neighbor belonging to the same class) and the nearest miss (the nearest neighbor belonging to a different class). In the construction of the decision tree, the neighbors of target objects are not considered individually. Instead, so-called buffers are created around the target objects and the non spatial attribute values are aggregated over all objects contained in the buffer.

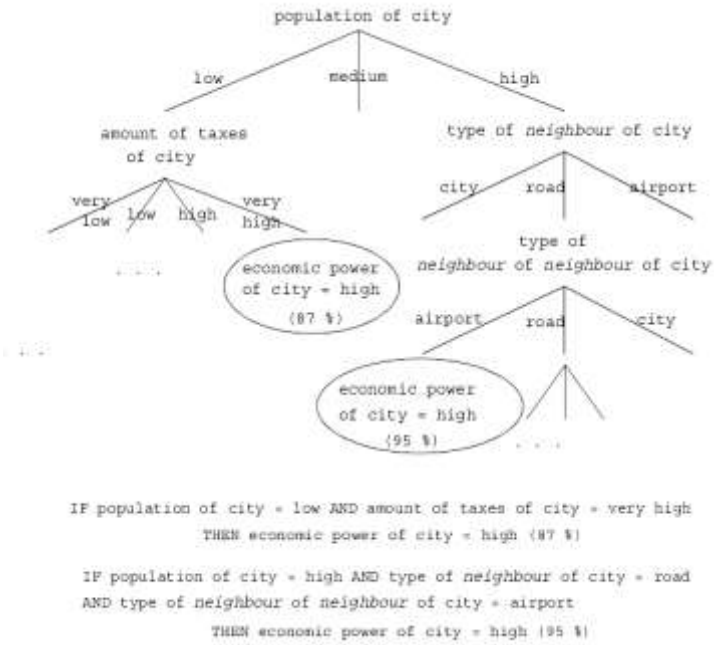
For instance, in the case of shopping malls a buffer may represent the area where its customers live or work. The size of the buffer yielding the maximum information gain is chosen and this size is applied to compute the aggregates for all relevant attributes.

The task of classification is to assign an object to a class from a given set of classes based on the attribute values of the object. In spatial classification the attribute values of neighbouring objects may also be relevant for the membership of objects and therefore have to be considered as well.

The extension to spatial attributes is to consider also the attribute of objects on a neighbourhood path starting from the current object. Thus, we define generalized attributes for a neighbourhood path $p = [o_1, \ldots, o_k]$ as tuples (attribute-name, index) where index is a valid position in p representing the attribute with attribute name of object o_{index}. The generalized attribute (economic-power,2), e.g., represents the attribute economic-power of some (direct) neighbour of object o_1.

Because it is reasonable to assume that the influence of neighbouring objects and their attributes decreases with increasing distance, we can limit the length of the relevant neighbourhood paths by an input parameter max-length.

Furthermore, the classification algorithm allows the input of a predicate to focus the search for classification rules on the objects of the database fulfilling this predicate. It depicts a sample decision tree and two rules derived from it. Economic power has been chosen as the class attribute and the focus is on all objects of type city.



Sample decision tree and rules discovered by the classification algorithm

2.2 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 analyses.

In spatial statistics, the standard against which spatial point patterns are often compared is a completely spatially random point process, and departures indicate that the pattern is not distributed randomly in space. Complete spatial randomness (CSR) [Cressie1993] is synonymous with a homogeneous Poisson process. The patterns of the process are independently and uniformly distributed over space, i.e., the patterns are equally likely to occur anywhere and do not interact with each other. However, patterns generated by a non-random process can be either cluster patterns(aggregated patterns) or decluster patterns(uniformly spaced patterns).

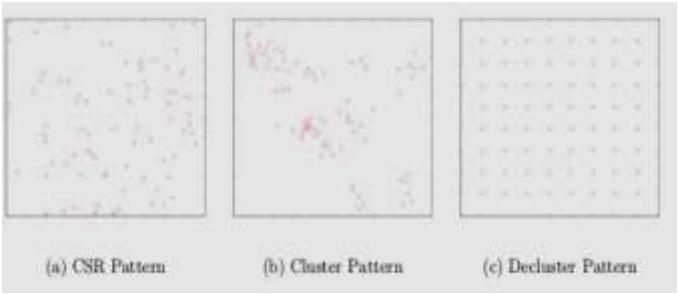


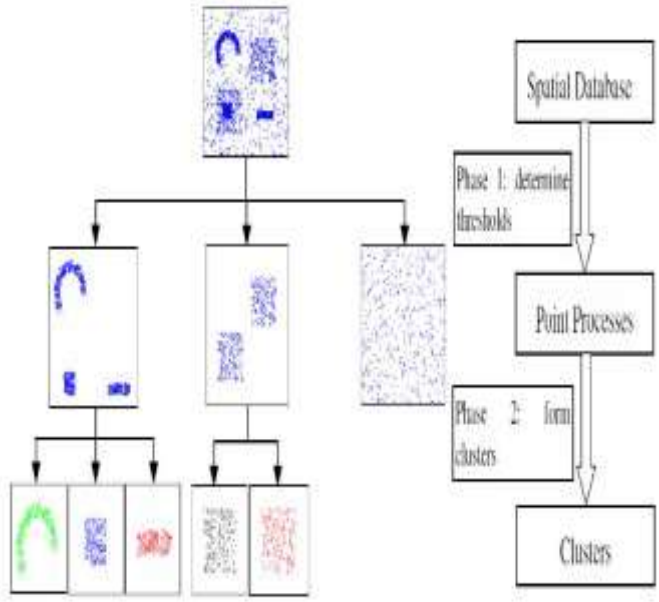
Illustration of CSR, Cluster, and Decluster Patterns

Several statistical methods can be applied to quantify deviations of patterns from a complete spatial randomness point pattern. One type of descriptive statistics is based on quadrats (i.e., well defined area, often rectangle in shape).Usually quadrats of random location and orientations in the quadrats are counted, and statistics derived from the counters are computed. Another type of statistics is based on distances between patterns; one such type is Ripley's K-function . After the verification of the statistical significance of the spatial clustering, classical clustering algorithms can be used to discover interesting clusters.

2.2.1 DECODE (DiscovEring Clusters Of Different dEnsities)

Discovering clusters in complex spatial data, in which clusters of different densities are superposed, severely challenges existing data mining methods. For instance, in seismic research, foreshocks (which indicate forthcoming strong earthquakes) or aftershocks (which may help to elucidate the mechanism of major earthquakes) are often interfered by background earthquakes. In this context, data are presumed to consist of various spatial point processes in each of which points are distributed at a constant, but different intensity. DECODE is a new density-based cluster method(DECODE) to discover clusters of different densities in spatial data. The novelties of DECODE are 2-fold:

(1) It can identify the number of point processes with little prior knowledge.It can automatically estimate the thresholds for separating point processes and clusters.



Flowchart of the method for discovering clusters of different densities in spatial data Two strategies have been adopted for finding density homogeneous clusters in density-based methods:

(1) Grid-based clustering method

Map data into a mesh grid and identify dense regions according to the density in cells. The main advantage : detection capability for finding arbitrary shaped clusters and their high efficiency in dealing with complex data sets which are characterized by large amounts of data, high dimensionality and multiple densities.

(2) Distance-based clustering method

Often based on the mth nearest neighbor distance. When clusters with different densities and noise coexist in a data set, few existing methods can determine the number of processes and precisely estimate the parameters.

Therefore, DECODE is a solution for it. DECODE is based upon a reversible jump Markov Chain Monte Carlo(MCMC) strategy and divided into three steps:

(1) Map each point in the data to its m^{th} nearest distance.

(2)Classification thresholds are determined via a reversible jump MCMC strategy.

(3) Clusters are formed by spatially connecting the points whose m^{th} nearest distances fall into a particular bin defined by the thresholds.

3. ASSOCIATION RULES

Spatial association rule is a rule indicating certain association relationship among a set of spatial and possibly some non spatial predicates. A strong rule indicates that the patterns in the rule have relatively frequent occurrences in the database and strong implication relationships. Additional data organization and retrieval tools can only handle the storage and retrieval of explicitly stored data. The extraction and comprehension of the knowledge implied by the huge amount of spatial data, though highly desirable, pose great challenges to currently available spatial database technologies.

A spatial characteristic rule is a general description of a set of spatial-related data. For example, the description of the general weather patterns in a set of geographic regions is a spatial characteristic rule. A spatial discriminant rule is the general description of the contrasting or discriminating features of a class of spatial-related data from other class(es). For example, the comparison of the weather patterns in two geographic regions is a spatial discriminant rule. A spatial association rule is a rule which describes the implication of one or a set of features by another set of features in spatial databases. For example, a rule like most big cities in Canada are close to the Canada-U.S. border" is a spatial association rule. A strong rule indicates that the patterns in the rule have relatively frequent occurrences in the database and strong implication relationships. A spatial association rule is a rule in the form of

P1 Ù … Ù Pm ® Q1 Ù … ÙQn (c%) ,

where at least one of the predicates P1 ,…, Pm , Q1 ,…, Qn is a spatial predicate, and c% is theconfidence of the rule.

A rule “P □ Q/S” is strong if predicate “P Ù Q” is large in set S and the confidence of“P ® Q/S” is high.

Steps for extracting the association rules:

STEP 1: Task_relevant_DB := extract task relevant objects(SDB ,RDB); (Relevant objects are collected into one database)

STEP 2: Coarse_predicate_DB := coarse spatial computation(Task relevant DB); (Spatial algorithm is executed at the coarse resolution level)

STEP 3: Large_Coarse_predicate_DB := filtering_with_mininmum_support(Coarse_predicate_DB); (computes the support for each predicate in Coarse_predicate_DB,and filters out those entries whose support is below the minimum support threshold at the top level.)

STEP 4: Fine_predicate_DB := refined_spatial_computation(Large_Coarse_predicate_DB); (Algorithm is executed at fine resolution level.)

STEP 5: Find_large_predicates_and_mine_rules(Fine_predicate_DB);

4 CHALLENGES AND APPLICATIONS

4.1 Challenges

A noteworthy trend is the increasing size of data sets in common use, such as records of business transactions, environmental data and census demographics. These data sets often contain millions of records, or even far more. This situation creates new challenges in coping with scale. The challenges and impacts can be classified into three main areas, namely, geographic information in knowledge discovery, geographic knowledge discovery in geographic information science and geographic knowledge discovery in geographic research.

The huge volume of spatial data . So retrieval and storage is difficult. The spatial data types/structures are complex .Expensive spatial processing operations.

4.2 Application In Land Use Dynamic Monitoring

The mass data storaged in spatial database includes spatial topological, nospatial properties and objects appearing variety on the time .The main knowledge types that can be discovered in the spatial database are: general geometric knowledge, spatial distribution rules, spatial association rules, spatial clustering rules, spatial characteristic rules, spatial discriminate rules, spatial evolution rules etc.. For land use dynamic monitoring, according to the knowledge mined in the spatial database, there are following several applications.

a. Make prediction of land variety According to geographic location, soil characters, geologic circumstance, prevent or control flood information, transportation circumstance etc. of the land, Making use of the spatial distribution rules, spatial clustering rules, spatial characteristic rules, spatial discriminate rules to analyze can get the distributing and the future development of the land .

b. Provide decision support for the city planning Spatial data mining technique makes use of general geometric knowledge, spatial distribution rules, spatial association rules, spatial evolution rules to get many factors about terrain, prevent or control flood, preventive pollution during the city planning for providing good data environment in city construction.

c. Valid management and analysis of remote sensing monitoring result According to algorithm of spatial data mining, based on knowledge discovery, it can validly output various statistical charts, images, query result and analysis result for decision.

4.3 Application In Visual Data Mining

For data mining of large data sets to be effective, it is also important to include humans in the data exploration process and combine their flexibility, creativity, and general knowledge with the enormous storage capacity and computational power of today's computers. Visual data mining applies human visual perception to the exploration of large data sets. Presenting data in an interactive, graphical form often fosters new insights, encouraging the formation and validation of new hypotheses to the end of better problem-solving and gaining deeper domain knowledge. Visual data mining often follows a three step process: Overview first, zoom and filter, and then details-on demand.

Some of they key advantages of visual data exploration over automatic data mining techniques alone are:

- yields results more quickly, with a higher degree of user satisfaction and confidence in findings
- are especially useful when little is known about the data and exploration goals

- are vague, because the analyst guides the search and can shift or adjust goals on the fly .
- can deal with highly non-homogeneous and noisy data .
- can be intuitive and require less understanding of complex mathematical or statistical algorithms or parameters .
- can provide a qualitative overview of the data, allowing unexpected phenomena .

5. RESEARCH NEEDS

In this section, we discuss some areas where further research is needed in spatial data mining.

• Comparison of classical data mining techniques with spatial data mining techniques

Existing literature does not provide guidance regarding the choice between classical data mining techniques and spatial data mining techniques to mine spatial data. Therefore new research is needed to compare the two sets of approaches in effectiveness and computational efficiency.

• Modeling semantically rich spatial properties,

such as topology 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

• Improving computational efficiency

Mining spatial patterns is often computationally expensive. For example, the estimation of the parameters for the spatial autoregressive model is an order of magnitude more expensive than that for the linear regression in classical data mining. Research is needed to reduce the computational costs of spatial data mining algorithms by a variety of approaches including the classical data mining algorithms as potential filters or components.

•Preprocessing spatial data T

here is a need for preprocessing techniques for spatial data to deal with problems such as treatment of missing location information and imprecise location specifications, cleaning of spatial data, feature selection and data transformation

6. CONCLUSION

Spatial Data Mining extends relational data mining with respect to special features of spatial data, like mutual influence of neighboring objects by certain factors (topology, distance, direction). It is based on techniques like generalization, clustering and mining association rules. Some algorithms require further expert knowledge that can not be mined from the data, like concept hierarchies. Spatial data mining is a niche area within data mining for the rapid analysis of spatial data.

Spatial data can potentially influence major scientific challenges, including the study of global climate change and genomics. The distinguishing characteristics of spatial data mining can be netaly summarized by the first law of geography:All things are related, but nearby things are more related than distant things. Spatial data mining is being used in various fields like remote sensing sattelite, Visyal data mining to mine data.

BIBLIOGRAPHY

[1] Martin Ester, Alexander Frommelt, Hans-Peter Kriegel, Jörg Sander . Spatial Data Mining:Database Primitives, Algorithms and Efficient DBMS Support . "Integration of Data Mining with Database Technology", Data Mining and Knowledge Discovery, an International -Journal, Kluwer Academic Publishers, 1999.

[2] Shashi Shekhar , Pusheng Zhang , Yan Huang , Ranga Raju Vatsavai. Trends in Spatial Data Mining,Department of Computer Science and Engineering, University of Minnesota 4-192, 200 Union ST SE, Minneapolis, MN 55455

[3] Sashi Sekhar ,Chang-Tien Lu and Pusheng Zhang. A Unified Approach To Detecting Spatial Outlier. Published in:Journal Geoinformatica Volume 7 Issue 2, June 2003

[4] Zhong yong a, Zhang jixian a, Yan qin. Research on spatial data mining technique applied in land use dynamic monitoring.Proceedings of International Symposium on Spatio-temporal Modeling, Spatial Reasoning, Analysis, Data Mining and Data Fusion. Volume XXXVI-2/W25, 2005

[5] Daniel A. Keim Christian Panse Mike Sips . Pixel Based Visual Mining of Geo-Spatial Data . Published in:Journal IEEE Computer Graphics and Applications Volume 24 Issue 5, September 2004

[6] Krzysztof Koperski and Jiawei Han . Discovery of Spatial Association Rules in Geographic Information database.Proceeding SSD '95 Proceedings of the 4th International Symposium on Advances in Spatial Databases Springer-Verlag London, UK ©1995.

