International Journal of Computational Intelligence Techniques, ISSN: 0976–0466 & E-ISSN: 0976–0474
Volume 1, Issue 1, 2010, PP-06-13

ACO based spatial data mining for traffic risk analysis

Ravikumar K. and Gnanabaskaran A.
Department of CSE, K. S. Rangasamy College of Technology, Tiruchengode, TamilNadu, 637211
mkravikkumar@gmail.com, gnanabas_2000@yahoo.com

Abstract- Many organizations have collected large amounts of spatially referenced data in various application areas such as geographic information systems (GIS), banking and retailing. These are valuable mines of knowledge vital for strategic decision making and motivate the highly demanding field of spatial data mining i.e., discovery of interesting, implicit knowledge from large amounts of spatial data. Most government local administrations collect and/or use geographical databases on the road accidents, on the road network and sometimes on the vehicle flow and sometimes on the mobility of inhabitants. In addition, other databases provide additional information on the geographical environment - trend layers - like administrative boundaries, buildings, census data, etc. These data contain a mine of useful information for the traffic risk analysis. There was a first study aiming identifying and at predicting the accident risk of the roads. It used a decision tree that learns from the inventoried accident data and the description of the corresponding road sections. However, this method is only based on tabular data and does not exploit geographical location. Using the accident data, combined to trend data relating to the road network, the traffic flow, population, buildings, etc., this project aims at deducing relevant risk models to help in traffic safety task. The existing work provided a pragmatic approach to multi¬layer geo-data mining. The process behind was to prepare input data by joining each layer table using a given spatial criterion, then applying a standard method to build' a decision tree. It allows the end-user to evaluate the results without any assistance by an analyst or statistician. The existing work did no consider multi-relational data mining domain. The efficiency of risk factor evaluation requires automatic filtering of spatial relationships. The quality of a decision tree depends, on the whole, of the quality of the initial data which are incomplete, incorrect or non-relevant data inevitably leads to erroneous results. The proposed model develops an ant colony algorithm for the discovery of spatial trend patterns found in a GIS traffic risk analysis database. The proposed ant colony based spatial data mining algorithm applies the emergent intelligent behavior of ant colonies to handle the huge search space encountered in the discovery of this knowledge. Genetic algorithm is deployed to evaluate the spatial risk pattern rule sets to its optimization on search phase in quick successions. The experimental results on a geographical traffic (trend layer) spatial database show
that our method has higher efficiency in performance of the discovery process and in the quality of trend patterns discovered compared to other existing approaches using non-intelligent decision tree heuristics.
Keywords-Geographic information systems (GIS), Ant Colony Optimization (ACO), Spatial Data Mining (SDM), Genetic Algorithm (GA)

I. Introduction

patial 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 explosive growth of spatial data and widespread use of spatial databases emphasize the need for the automated discovery of spatial knowledge. 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 geospatial data are crucial to organizations which make decisions based on large spatial datasets including NASA, the National Imagery and Mapping Agency (NIMA), and the National Cancer Institute (NCI). These organizations are spread across many application domains including ecology and environmental management, public safety, transportation, Earth science, epidemiology, and climatology. Specific features of geographical data that preclude the use of

general purpose data mining algorithms are rich data types (e.g., extended spatial objects) implicit spatial relationships among the variables, observations that are not independent and spatial autocorrelation among the features. Our focus of this overview is on the methods of spatial data mining, i.e., discovery of interesting knowledge from spatial data. Spatial data are the data related to objects that occupy space. The institutions concerned by the road safety are studying the application of data mining techniques for traffic risk analysis task. The proposed work aims at spatial feature of the accidents and their interaction with the geographical environment. In previous work, the system has implemented some spatial data mining methods such as generalization and characterization. The proposal of this work uses intelligent ant agent to evaluate the search space of the traffic risk analysis along with usage of genetic algorithm for risk pattern (areas, building cause's obstacles) storage and referral to the ant agents.

II. Literature Review

Spatial data mining fulfills real needs of many geomantic applications. It allows taking advantage of the growing availability of geographically referenced data and their potential richness. This includes the spatial analysis of risk such as epidemic risk or traffic accident risk in the road network. This work deals with the method of decision tree for spatial data classification. This method differs from conventional decision trees by taking account implicit spatial relationships in addition to other object attributes. Ref [2, 3] aims at taking account of the spatial feature of the accidents and their interaction with the geographical environment. It involves a new field of data mining technology that is spatial data mining. In the previous work, the system has implemented some spatial data mining methods such as generalization and characterization. This work [3] presents the approach to spatial classification and its application to extend TOPASE. Clustering in spatial data mining is to group similar objects based on their distance, connectivity, or their relative density in space. In the real world, there exist many physical obstacles such as rivers, lakes and highways, and their presence may affect the result of clustering substantially. In this project, the system studies the problem of clustering in the presence of obstacles and defines it as a COD (Clustering with Obstructed Distance) problem. As a solution to this problem, the system proposes a scalable clustering algorithm, called COD-CLARANS [5,6]. Spatial Clustering with Obstacles Constraints (SCOC) has been a new topic in Spatial Data Mining (SDM). In [8] the author proposes an Improved Ant Colony Optimization (IACO) and Hybrid Particle Swarm Optimization (HPSO) method for SCOC. In the process of doing so, the system first use IACO to obtain the shortest obstructed distance, which is an effective method for arbitrary shape obstacles, and then the system develop a novel HPKSCOC based on HPSO and K-Medoids to cluster spatial data with obstacles, which can not only give attention to higher local constringency speed and stronger global optimum search, but also get down to the obstacles constraints. Spatial clustering is an important research topic in Spatial Data Mining (SDM). Many methods have been proposed in the literature, but few of them have taken into account constraints that may be present in the data or constraints on the clustering. These constraints have significant influence on the results of the clustering process of large spatial data. In this project, the system discuss the problem of spatial clustering with obstacles constraints and propose a novel spatial clustering method based on Genetic Algorithms (GAs) and KMedoids, called GKSCOC, which aims to cluster spatial data with obstacles constraints.[9] Spatial data mining method is used to enrich Customer Intelligence analytical function in this project. The system first proposes a spatial data classification method which can handle the uncertainty property of customer data. On the basis of spatial classification rules, the system then proposes a detection method of potential customers by map overlapping. Deep spatial analytical function is realized in customer intelligence system which can not be done by traditional data mining method. With the coming of E-business, the enterprises are now faced harder competition than before. So they now focus their attention on customers instead of their production only. In order to win the competition, the enterprises have to provide

ISSN: 0976–0466 & E-ISSN: 0976–0474, Volume 1, Issue 1, 2010

their customers more individualized and more efficient service. Customer intelligence (C1) system appears in recent years to meet the need of above emergence. From the analytical function of the system, customer intelligence is a decision analytical method which includes customer identification, customer selection, customer acquirement, customer improvement and customer maintenance [12]. The spatial co-location rule problem is different from the association rule problem, since there is no natural notion of transactions in spatial data sets which are embedded in continuous geographic space. In this project, the system provides a transaction-free approach to mine co-location patterns by using the concept of proximity neighborhood. A new interest measure, a participation index, is also proposed for spatial co-location patterns. The participation index is used as the measure of prevalence of a co-location for two reasons. Modeling spatial context (e.g., autocorrelation) is a key challenge in classification problems that arise in geospatial domains. In [13] Markov random fields (MRF) are a popular model for incorporating spatial context into image segmentation and land-use classification problems. The spatial auto regression (SAR) model [14], which is an extension of the classical regression model for incorporating spatial dependence, is popular for prediction and classification of spatial data in regional economics, natural resources, and ecological studies.

III. Genetic and ACO Based Spatial Data Mining Model

The proposed spatial data mining model uses ACO integrated with GA for risk pattern storage. The proposed ant colony based spatial data mining algorithm applies the emergent intelligent behavior of ant colonies. The proposed system handle the huge search space encountered in the discovery of spatial data knowledge. It applies an effective greedy heuristic combined with the trail intensity being laid by ants using a spatial path. GA uses searching population (set) to produce a new generation population. It evolves into the optimum state progressively by exerting a series of genetic operators such as selection, crossover and mutation etc on traffic risk patterns. The proposed system develops an ant colony algorithm for the discovery of spatial trends in a GIS traffic risk analysis database. Intelligent ant agents are used to evaluate valuable and comprehensive spatial patterns.

A. Geospatial Data Mining

Geospatial data repositories tend to be very large. Moreover, existing GIS datasets are often splintered into feature and attribute components that are conventionally archived in hybrid data management systems. Algorithmic requirements differ substantially for relational (attribute) data management and for topological (feature) data management. Related to this is the range and diversity of geographic data formats that also presents unique challenges. The digital geographic data revolution is creating new types of data formats beyond the traditional "vector" and "raster" formats. Geographic data repositories increasingly include ill-structured data such as imagery and geo-referenced multimedia. The strength of GIS is in providing a rich data infrastructure for combining disparate data in meaningful ways by using spatial proximity. Also, through the use of trend map coloring, geographic visualization of individual variables can be very effective for identifying competitive hot zones, merchandising opportunities, etc. The next logical step to take GIS analysis beyond demographic reporting to true market intelligence is to incorporate the ability to analyze and condense a large number of variables into a single forecast or score. This is the strength of predictive data mining technology and the reason why there is such a hand-in-glove fit between GIS & data mining. Depending upon the specific application, GIS can combine historical customer or retail store sales data with syndicated demographic, business, traffic, and market research data as shown Fig 1. This dataset is then ideal for building predictive models to score new locations or customers for sales potential, cross-selling, targeted marketing, customer churn, and other similar applications. Geospatial data repositories tend to be very large. Moreover, existing GIS datasets are often splintered into feature and attribute components that are conventionally archived in hybrid data management systems. Algorithmic requirements differ substantially for relational (attribute) data management

and for topological (feature) data management. Related to this is the range and diversity of geographic data formats that also presents unique challenges. The digital geographic data revolution is creating new types of data formats beyond the traditional "vector" and "raster" formats. Geographic data repositories increasingly include ill-structured data such as imagery and geo-referenced multimedia. The strength of GIS is in providing a rich data infrastructure for combining disparate data in meaningful ways by using spatial proximity. Also, through the use of trend map coloring, geographic visualization of individual variables can be very effective for identifying competitive hot zones, merchandising opportunities, etc. The next logical step to take GIS analysis beyond demographic reporting to true market intelligence is to incorporate the ability to analyze and condense a large number of variables into a single forecast or score. This is the strength of predictive data mining technology and the reason why there is such a hand-in-glove fit between GIS & data mining. Depending upon the specific application, GIS can combine historical customer or retail store sales data with syndicated demographic, business, traffic, and market research data as shown Fig 1. This dataset is then ideal for building predictive models to score new locations or customers for sales potential, cross-selling, targeted marketing, customer churn, and other similar applications.

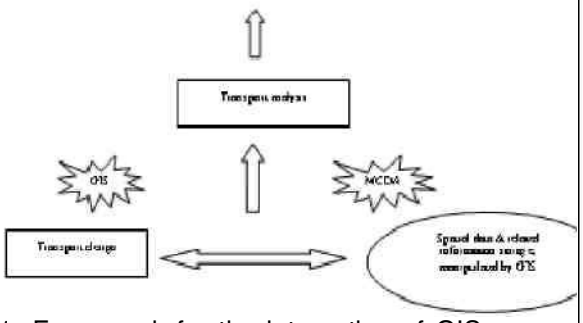

Fig.1- Framework for the integration of GIS and multi-criteria decision analysis

B. Ant Colony Optimization

Ant Colony Optimization (ACO) is a meta-heuristic inspired by real ant colonies in nature. Ant colonies intelligently solve complex discrete problems like finding shortest path although its individuals are very simple and not intelligent enough to solve such problems on their own. The main underlying idea is to use a multi-agent parallel search on the different possible combinations of solution components. The decision to choose a component is based on a local problem data and a dynamic shared global memory of the colony that contains a history on the quality of previously obtained results. Considering the challenges faced in the problem of spatial trend detection we can see that ACO can suggest efficient properties in these aspects. Firstly, the ant agents can search for the trend starting from their own start point in a completely distributed manner. This omits the need to get the start point node from the user. Secondly to guide the stochastic search of the ants, the pheromone trails can help the ants to exploit the trend detection experience of the colony. This guides the search process to converge to a better subspace potentially containing more and better trend patterns. Finally some measures of attractiveness can be defined for selecting a feasible spatial object from the neighborhood graph which can effectively guide the trend detection process of an ant (Fig 2).

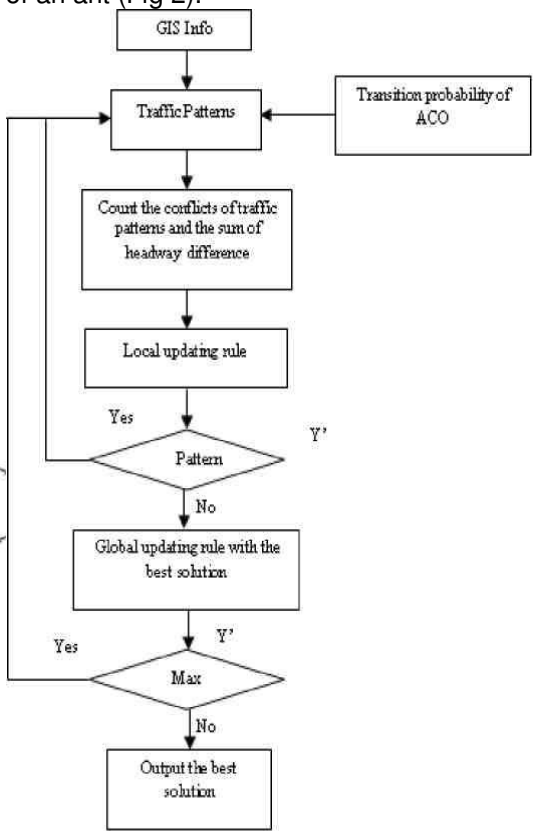

Fig. 2- Conflicts of Traffic Patterns by Ants

C. Genetic Algorithm

The proposed algorithm of spatial clustering based on GAs can be described as follows. Divide an individual risk pattern of the traffic generating objects (chromosome) into n part and each part is corresponding to the classification of a datum element. The optimization criterion is defined by a Euclidean distance among the data frequently, and the initial population is produced at random. Its genetic operators are similar to standard GA's. This method can find the global optimum solution and not influenced by an outlier, but it only fits for the situation of small traffic risk pattern data sets and classification number.

IV. Experimental Evaluation

ACO with GA integration SPDM model is proposed to be tested in the framework of traffic risk analysis. The analysis is done on a spatial database provided in the framework of an industrial collaboration. Contains data on the road accidents and others on the geographical environment (buildings, roads etc.). The objective is to construct a predictive model. The system model looks correspondences between the accidents and the other trend layers as the road network, building… etc. It applies classification by decision tree while integrating the accidents spatial character and their interaction with the geographical environment. The experimental evaluation is made on a geographical traffic (trend layer) spatial database to depict higher efficiency in performance of the discovery process. It proves that better quality of trend patterns discovered compared to other existing approaches using non-intelligent decision tree heuristics. Reliable data constitute the key to success of a decision tree. An efficient parallel and near global optimum search for traffic risk patterns are evaluated using genetic algorithm. It combines the concept of survival of the fittest with a structured interchange. GAs imitate natural selection of the biological evolution. Improvements in the identification of high or low risk areas can assist the emergency preparedness planning and resource evaluation. To enhance the determination of a risk area we visualize the spatial distribution phenomena like population distribution, building types, workspace distribution etc and its relevance to the emergency services. It examines the relation between population density distribution and incident density.

A. Spatial Data Mining on Traffic Risk Patterns

Spatial data mining on traffic risk pattern focuses on the human vulnerability in built environments. It considers issues like differences between common and rare accidents, commuting of people, and relations between accidents and networks. Visualization and interaction helps to understand the dependencies within and between data sets. Visualization supports formulating hypotheses and answering questions about correlations between certain variables and accidents. Explorative visualization may reveal new variables relevant to the model and relevance of already used variables. It is highly required to analyze the correlations combine spatial data analysis methods with visualization. Risk model development is an interactive and explorative process.

B. ACO on SPDM

ACO has been recently used in some data mining tasks, e.g. classification rule discovery. Considering the challenges faced in the problem of spatial trend detection, ACO suggest efficient properties in these aspects. Ant agents search for the trend starting from their own start point in a completely distributed manner. This omits the need to get the start point node from the user. Guide the stochastic search of the ants, the pheromone trails can help the ants to exploit the trend detection experience of the colony. This guides the search process to converge to a better subspace potentially containing more and better trend patterns. Finally some measures of attractiveness can be defined for selecting a feasible spatial object from the neighborhood graph. ACO on SPDM Effectively guide the trend detection process of an ant ACO has been recently used in some data mining tasks, e.g., classification rule discovery. Considering the challenges faced in the problem of spatial trend detection, ACO suggest efficient properties in these aspects. Ant agents search for the trend starting from their own start point in a completely distributed manner. This omits the need to get the start point node from the user. Guide

the stochastic search of the ants, the pheromone trails can help the ants to exploit the trend detection experience of the colony. This guides the search process to converge to a better subspace potentially containing more and better trend patterns. Finally some measures of attractiveness can be defined for selecting a feasible spatial object from the neighborhood graph. ACO on SPDM Effectively guide the trend detection process of an ant.

C. Spatial Clustering GA

Genetic algorithms are an efficient parallel and near global optimum search method based on nature genetic and selection. GA combines the concept of survival of the fittest with a structured interchange. GAs imitate natural selection of the biological evolution. It uses searching population (set) to produce a new generation population. It evolves into the optimum state progressively by exerting a series of genetic operators such as selection, crossover and mutation etc. GAs automatically achieve and accumulate the knowledge about the search space. GA adaptively controls the search process to approach a global optimal solution. GA performs well in highly constrained problems, where the number of "good" solutions is very small relative to the size of the search space. GAs provide better solution in a shorter time, including complex problems to solve by traditional methods.

V. Results and Discussion

The proposed results provide spatial decision trees for traffic risk patterns with optimized route structure with the ant agents. The proposed model classifies objects according to spatial information (using the ant agent and the distance pheromone). Spatial classification provided by the proposed scheme is simple and efficient. It allows adapting to different decision tree algorithm for the spatial modeling of traffic risk patterns. It uses the structure of geo-data in multiple trend layers which is characteristic of geographical databases. Finally, the quality of this analysis is improved by enriching the spatial database by multiple geographical trends, and by a close collaboration with a domain-specialist in traffic risk analysis. The advantage of proposed technique allows the end-user to evaluate the results without any assistance by an analyst or statistician. GAs automatically achieve and accumulate the knowledge about the search space of the ACO. GA adaptively controls the traffic risk pattern search process to approach a global optimal solution. Perform well in highly constrained traffic risk pattern, where the number of "good" solutions is very small relative to the size of the search space. GAs provide better solution in a shorter time. The below table 1 shows the comparative figures of our experimentation of SPDM with decision tree heuristics and combined for the traffic patterns. Data cleaning by first, eliminating the detail second, by eliminating information (attributes) useless for the analysis (the combination of projection and selection of the relational database model make this). Retrieving spatial relationships: a spatial join using the spatial join index makes this. Our proposal uses ant agents along with a distance based join between the trend layers "accidents" and "constructions". Building the gene based decision tree with ant colony optimization techniques. The proposed results are interpreted to extract the relevant rules generated by genetic algorithm.

Table 1- Comparison Of SPDM-Decision Tree With SPDM-ACO-GA

Schemes	Min. Degree	Max Degree	Avg Degree	Min. Distance	Max Distance	Avg Distance
SPDM-ACO-GA	:4	40	15.45	60	578	234
SPDM-Decision Tree:	6	60	25.34	76	654	346

The current application results show a use case of spatial decision trees. The contribution of this approach to spatial classification lies in its simplicity and its efficiency. It makes it possible to classify objects according to spatial information (using the distance). It allows adapting any decision tree algorithm or tool for a spatial modeling problem. Furthermore, this method considers the structure of geo-data in multiple trends (patterns) which is characteristic of geographical databases. The graph below indicates the number of trends found and paths examined using SPDM Decision Tree and SPDM-ACO-GA models for traffic risk pattern analysis.

It raised some new questions to which the system will focus our future research. The first problem is related to the limitation of most data mining methods. In effect, they only operate on a single-table and single

ISSN: 0976–0466 & E-ISSN: 0976–0474, Volume 1, Issue 1, 2010

row example format. If the data are stored over multiple tables, or a table contains examples that are described by several rows, it should be pre-processed to fit in the expected single-table format. This conversion may lose potentially valuable information. Furthermore, it involves a pre-processing overhead. In spatial data mining, it is especially important to provide the multi-tables / multi-rows-examples, in reason precisely of the spatial relationships. In order to overcome this limitation, our method transforms multi-tables in a single-table thanks to the spatial join processed. But, each example (an accident and its type) is duplicated, as many times as there exits neighboring constructions. This could affect the classification process and the rightness of its resulting rules. Recently, this problem has been addressed by what is called multi-relational data mining. This will be valuable especially for spatial data mining purpose. At the moment, the user should enter spatial criteria. A domain-specialist has probably a precise idea of the relevant criteria but this becomes difficult when the number of trends (patterns) increases.

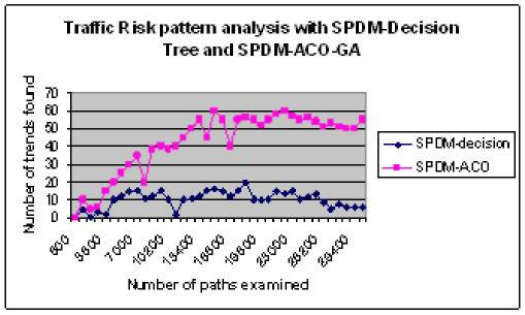

Fig. 3-Traffic Risk pattern analysis with SPDM-Decision Tree and SPDM-ACO-GA

VI. Conclusion

The Spatial data mining system of ACO with GA have shown that traffic risk patterns are discovered efficiently and recorded in the genetic property for avoiding the accident risk in traffic dense spatial regions. The proposal of our system analyzes existing methods for spatial data mining and mentioned their strengths and weaknesses. The variety of yet unexplored topics and problems makes knowledge discovery in spatial databases an attractive and challenging research field. The system believes that some of the suggestions mentioned have already been though about by researchers and work may have already started on them. This work gives a pragmatic approach to multi-layer geo-data mining. The main idea is to prepare input data by joining each layer table using a given spatial criterion, then applying a standard method to build' a decision tree. The most advantage is to demonstrate the feasibility and the interest of integrating neighborhood properties when analyzing spatial objects. Our future work will focus on adapting recent work in multi-relational data mining domain, in particular on the extension of the spatial decision trees based on neural network. Another extension will concern automatic filtering of spatial relationships. The system will study of its functional behavior and its performances for concrete cases, which has never been done before. Finally, the quality of this analysis could be improved by enriching the spatial database by other geographical trends, and by a close collaboration with a domain specialist in traffic risk analysis. Indeed, the quality of a decision tree depends, on the whole, of the quality of the initial data.

References

[1] Agrawal R., and Srikant R. 1994. Fast algorithms for Mining Association Rules. In Proc. of Very Large Databases.

[2] Anselin, L. 1988. Spatial Econometrics: Methods and Models. Dordrecht, Netherlands: Kluwer. Anselin, L. 1994. Exploratory Spatial Data Analysis and Geographic Information Systems. In Painho, M., ed., New Tools for Spatial Analysis, 45{54.

[3] Anselin, L. 1995. Local Indicators of Spatial Association: LISA. Geographical Analysis 27(2):93{115.

[4] Barnett, V., and Lewis, T. 1994. Outliers in statistical Data. John Wiley, 3rd edition edition. [Besag1974] Besag, J. 1974. Spatial Interaction and Statistical Analysis of Lattice Systems. Journal of Royal Statistical Society: Series B 36:192{236.

[5] Bolstad, P. 2002. GIS Foundamentals: A Fisrt Text on GIS.Eider Press.

[6] Cressie, N. 1993. Statistics for Spatial Data (Revised Edition). New York: Wiley.

[7] Han, J.; Kamber, M.; and Tung, A. 2001. Spatial Clustering Methods in Data Mining: A Survey. In Miller, H., and Han, J., eds., Geographic Data Mining and Knowledge Discovery. Taylor and Francis.

[8] Hawkins, D. 1980. Identification of Outliers. Chapman and Hall. [Jain & Dubes1988] Jain, A., and Dubes, R. 1988. Algorithms for Clustering Data. Prentice Hall.

[9] Jhung, Y., and Swain, P. H. 1996. Bayesian Contextual Classification Based on Modified M-Estimates and Markov Random Fields. IEEE Transaction on Pattern Analysis and Machine Intelligence 34(1):67{75.

[10] Koperski, K., and Han, J. 1995. Discovery of Spatial Association Rules in Geographic Information Databases. In Proc. Fourth International Symposium on Large Spatial Databases, Maine. 47-66.

[11] Shekhar, S.; Schrater, P. R.; Vatsavai, R. R.; Wu, W.; and Chawla, S. 2002. Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. IEEE Transaction on Multimedia 4(2).

[12] Zhang, P.; Huang, Y.; Shekhar, S.; and Kumar, V. 2003. Exploiting Spatial Autocorrelation to Efficiently Process Correlation-Based Similarity Queries. In Proc. of the 8th Intl. Symp. on Spatial and Temporal Databases.

ISSN: 0976–0466 & E-ISSN: 0976–0474, Volume 1, Issue 1, 2010