1 Abstract

Widespread use of spatial databases is leading to an increasing interest in mining interesting, useful but implicit spatial patterns[11, 6, 14, 13] just as the widespread use of relational database triggered interest in classical data mining [3, 2, 30]. Spatial patterns of interest include characterization of the locations of a feature (e.g. crime) and its association with other spatial features (e.g. population density, distance to transportation network, etc.).

One of the major challenge in spatial data mining arises from the very large sizes of spatial databases. A high performance compute server with a large capacity storage server is essential for experimental work in evaluating spatial data mining techniques. The focus of our proposed work is to design and evaluate scalable algorithms for mining spatial patterns in large spatial databases in context of critical applications. Our team is capable of addressing these problems as we have been working on developing spatial data models[23], spatial indexing[26], spatial query processing[25, 27, 24] and data mining[25, 16, 21] for different application domains, including transportation[23, 28] and terrain visualization[27].

2 Spatial Data Mining : Introduction

Spatial databases[1, 10, 22, 17] has been an active area of research for over two decades. Its results, e.g. spatial multi-dimensional indexes [20] and OGIS [9] spatial data model, are being used in a number of applications of geographical info. systems[4, 19, 12, 29] ranging from crime mapping to environmental and ecological studies.

Widespread use of spatial databases is leading to an increasing interest in mining interesting, useful but implicit spatial patterns[11, 6, 14, 13] just as the widespread use of relational database triggered interest in classical data mining [3, 2, 30]. Spatial patterns of interest include characterization of the locations of a feature (e.g. crime) and its association with other spatial features (e.g. population density, distance to transportation network, etc.).

One of the major challenge in spatial data mining arises from the very large sizes of spatial databases. A high performance compute server with a large capacity storage server is essential for experimental work in evaluating spatial data mining techniques. The focus of our proposed work is to design and evaluate scalable algorithms for mining spatial patterns in large spatial databases in context of critical applications. Our team is capable of addressing these problems as we have been working on developing spatial data models[23], spatial indexing[26], spatial query processing[25, 27, 24] and data mining[25, 16, 21] for different application domains, including transportation[23, 28] and terrain visualization[27].

3 Background Information

Foundations of spatial data mining include spatial statistics, and data mining.

Spatial Statistics: The purposes of spatial statistical models can be divided into three categories: descriptive, explanatory, and predictive. Descriptive models characterize the distribution of the spatial phenomenon. Often description is based on a set of spatial statistics and indices. For example, a spatial distribution may be classified into random or clustered using spatial autocorrelation (e.g Moran's I coefficient), nearest neighbor index or quardat analysis[5, 8].

Explanatory model deals with spatial associations, i.e. relationships between a phenomenon and the factors affecting its spatial distribution. For example, in order to explain why crime clusters occur in a certain area, roles of population density, density of vacant houses, poverty rate etc. may be examined. More detailed analysis may explore how each factor may influence the crime locations. Example techniques are based on chi-square tests and spatial correlation coefficients using appropriate geographic units.

Predictive models may be used subsequently for prediction or simulation of alternative management strategies. For example, near future crime rate may be predicted given the current conditions and growth factor of significant factors (e.g. population density, poverty rates) under certain assumptions. Alternatively, these models may explore what may happen if certain conditions are changed via new management strategies. Example techniques include regression using appropriate geographic units, structural factors (e.g. local features of the geographic unit) as well as spatial factors (e.g. absolute location, distance to certain features and neighborhood effects such as spatial autocorrelation).

Spatial modeling may involve all feature types (points, lines, and polygons). Choice of geographic unit is a key decision for polygonal features. Polygonal geographic units may be arbitrary (e.g. a grid), based on existing

boundaries (e.g. administrative or political) or derived from data distribution (e.g. areas homogeneous with respect to significant factors).

Spatial Data Mining: Spatial data mining, a subfield of data mining, is concerned with discovery of interesting and useful but implicit knowledge in spatial databases. Common patterns [2] discovered by data mining algorithms include descriptive patterns (e.g. clustering[13]), explanatory patterns (e.g. association rules[15]) and predictive patterns (e.g. classification rules and decision trees). The foundations of data mining algorithms are in statistics and machine learning. One of the goals of data mining algorithms is to scale up to analyze very large datasets which may not fit in the main memory.

Challenges in spatial data mining [14, 13, 15] arise from following issues. *First,* classical data mining[2, 3] deals with numbers and categories. In contrast, spatial data is more *complex* and includes extended objects such as points, lines, and polygons. *Second,* classical data mining works with explicit inputs, whereas spatial predicates (e.g. overlap) and attributes (e.g. distance, spatial auto-correlation) are often *implicit. Third,* classical data mining treats each input to be independent of other inputs, whereas spatial patterns often must satisfy the constraints of continuity and *high autocorrelation among nearby features.* For example, population density of nearby locations are often related.

4 Application Domain

Crime mapping [7] is a study of the geographic profile of a serial criminal or a category of crime. Crimes are a human phenomena, and their geographical distribution is not random due to factors such as simple geographic convenience for an offender. While Maps offer an easy to understand graphic representations of crime-related issues, discovery and understanding of spatial patterns in locations of crimes can be even more valuable in attempts to fight crime.

Spatial analysis and data mining of urban crime data can find non-trivial patterns, which are beneficial in managing crime. For example, consider the decision about the assignment of police petrols to different areas of a large city. If the distribution of crime locations shows a clustered pattern, law enforcement agencies would be able to target areas of concentration for special preventive measures. Spatial hunting patterns of serial criminal can be used to hypothesize where these offenders might live. Policy makers in police departments might use more complex maps and spatial analysis to observe trends in criminal activity. Similarly, researchers working for politicians, the press, and the general public would be able to develop spatial models relating density of crime incidents to socioeconomic and demographic characteristics.

Descriptive models are being used widely in crime mapping and analysis to identify crime hot spots and allocate police units. One of the current challenges is to develop explanatory spatial models, e.g. to identify association between crime and other structural features (e.g. local population density, police allocation) and spatial features (e.g. distance to bars or police stations) etc. We focus on explanatory models, i.e. spatial associations, in the proposed work.

5 Proposed Research

Spatial databases organize geographic information as a collection of features. Features represent geographic phenomena such as crime locations, population density etc. Data about each feature is represented in spatial units such as two-dimensional points, lines and polygons. For example, the crime feature may be represented as points, i.e. locations (e.g. street address) of crime. Vacant houses, a feature affecting crime, may also be represented as points. Transportation networks, an important feature affecting crime, may be represented a collection of lines representing the center lines of various roads, railroads etc. Demographic features, e.g. population density, poverty rate, are often associated with administrative polygonal units such as census blocks.

The proposed research will focus on finding spatial association among features. This problem can be defined informally as follows. Given a dependent spatial feature (e.g. crime), and a set of other spatial features (e.g. population density, poverty rate, vacant houses, etc.), identify the spatial features with positive or negative associations with the dependent feature.

There are two families of techniques for finding spatial associations, namely association rule[2, 15] from data mining area and spatial statistical methods. Spatial association rule[2, 15] is common representation of spatial association within data mining area. A spatial association rule is often of the form "X implies Y : (c percent)" where X and Y are sets of spatial or non-spatial predicates and c percent is the confidence of the rule. An example of a

spatial association rule is $crime_location(X) => (distance(X, bar) < 1mile) : (70percent)$. This rule may represent the statistics that 70 percents of crime locations are within 1 mile of a bar. Algorithms for finding association rules often use explicit materialization of spatial predicates, e.g. distance to other features. This requires a fair amount of a priori knowledge about the spatial associations being explored. These algorithms are designed to be able to process very large spatial databases, that may not fit in the main memory. A spatial association rule is often used for finding positive association rather than negative associations.

Techniques from Spatial Statistics are based on chi-square tests and spatial correlation coefficients using appropriate geographic units. For point features, additional methods based on multi-variate point processes are available. Methods from spatial statistics can be used to find both positive and negative associations. In addition, they do not require explicit materialization of spatial predicate. However, most algorithms for spatial statistics are not designed to be able to process very large spatial databases, that may not fit in the main memory.

The goal of proposed work is to design and evaluate scalable algorithms for the spatial statistical techniques for spatial association. We plan to determine the dominance-zone for various techniques by using algebraic cost models and experiments with implementations.

6 Equipment Justification

The requested computer cluster is essential for the success of the research in this project. Spatial data mining algorithms require a large amount of computation. For example, a spatial join, a possible step in spatial data mining algorithms, to compute a map overlay of two polygonal maps may have to compute millions of polygon-polygon intersections. Computing intersection between a given pair of polygon itself may take tens of thousands of instructions as average number of edges in typical polygons ranges from hundreds to thousands. The tightly-coupled high-performance cluster is essential for developing scalable spatial data mining algorithms. At the same time, due to the large size of spatial data, it is necessary to have high storage capacity and data transfer rate such as those offered by requested storage system.

- Due to the large size of spatial data, it is necessary to have high storage capacity and data transfer rate such as those offered by proposed storage system. For example, TeraServer [18] storing 1-2meter resolution imagery on a small part of earth contains over a terabyte of compressed data. Similarly, the aerial photographs at centimeter level resolution for Minnesota Highways have about a terabyte of image and vector data.

- Due to the large size of individual datasets, it is important to have high bandwidth communication networks. For example, 400 dots per inch digitization of a 9inch by 9inch aerial photograph at 16-bit color per pixel leads to 25 Megabyte data per photograph. Almost 10,000 such images are collected for Minnesota highways every year yielding 250 Gigabytes of data. Loading even a small fraction of such a image dataset (with compression) to our lab. needs high bandwidth internet connection at both ends. Fetching these images from RAID storage to a geographics workstation for wrapping those over say vector elevation models at interactive speeds requires high bandwidth local area network such as ATM.

- Spatial data mining algorithms requires a large amount of computation. For example, a spatial join to compute a map overlay of two polygonal maps may have to compute millions of polygon-polygon intersections. Computing intersection between a given pair of polygon itself may take tens of thousands of instructions as average number of edges in typical polygons ranges from hundreds to thousands. High performance compute servers are essential for spatial data mining algorithms.

References

[1] N. Adam and A. Gangopadhyay. *Database Issues in GIS.* Kluwer Academic Publishers, 1997.

[2] R. Agrawal, T. Imielinski, and A. Swami. Database mining: A performance perspective. *IEEE Trans. on Knowledge and Data Eng.*, 5(6), December 1993.

[3] M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. *IEEE Trans. on Knowledge and Data Eng.*, 8(6), 1996.

[4] N. Chrisman. *Exploring Geographic Information Systems.* John Wiley and Sons, 1997.

[5] N. Cressie. *Statistics for Spatial Data.* Wiley Intersceice, 1993.

[6] D. Mark et al. Workshop on Geographic Information Science and Geospatial Activities at NSF . http://www.geog.buffalo.edu/ncgia/workshopreport.html.

[7] Dan Sadler. Exploring Crime Mapping. National Inst. of Justice Crime Mapping Research Center web-site http://www.ojp.usdoj.gov/cmrc/briefingbook/welcome.html.

[8] P. J. Diggle. *Statistical Analysis of Spatial Point Patterns*. Academic Press, 1983.

[9] K. Gardels. Open gis and on-line environmental libraries. *ACM SIGMOD Record*, 26(1):32–38, March 1997.

[10] R.H. Guting. An Introduction to Spatial Database Systems. *VLDB Journal*, 3:357–399, December 1994.

[11] H. Miller and J. Han and M. Egenhofer. NCGIA Specialist Meeting on Discovering Geographic Knowledge in Data-Rich Environments. National Science Foudation, http://www.spatial.maine.edu/ max/varenius/KD.html.

[12] Environmental Systems Research Institute. What is Geographic Information Systems? http://www.esri.com/base/gis/.

[13] E. M. Knorr and R. Ng. Finding Aggregate Proximity Relationship and Commonalities in Spatial Data Mining. *IEEE Trans. on Knowledge and Data Eng.*, 8(6):884–898, December 1996.

[14] K. Koperski, Junas Adhikary, and J. Han. Spatial data mining: Progress and challenges survey paper. In *Proc. Workshop on Research Issues on Data Mining and Knowledge Discovery*, Montreal, QB, Canada, June 1996.

[15] K. Koperski and J. Han. Discovery of Spatial Association Rules in Geographic Information Databases. In *Proc. 4th Intl. Symp. on Large Spatial Databases (SSD 95)*, pages 47–66, 1995.

[16] V. Kumar, S. Shekhar, and B. Amin. A Scalable, Highly Parallel Formulation of the Backpropagation Algorithm for Hypercubes and Related Architectures. *IEEE Trans. on Parallel and Distr. Systems*, 5(10), 1994.

[17] M.F. Worboys. *GIS: A Computing Perspective*. Taylor and Francis, 1995.

[18] Microsoft Terraserver Group. Welcome to the TerraServerTM User's Guide. http://www.terraserver.com.

[19] University Consortium on Geographic Info. Science. Congressional breakfast report on gis reseach. http://http://osu.orst.edu/dept/geosciences/congress/breakfast.html, 1998.

[20] H. Samet. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, 1990.

[21] S. Shekhar and B. Amin. Generalization by Neural Networks. *IEEE Trans. on Knowledge and Data Eng. (April)*, 4(2), 1992.

[22] S. Shekhar, S. Chawla, S. Ravada, A. Fetterer, X. Liu, and C. T. Liu. Spatial databases: Accomplishment and research directions, January 1999.

[23] S. Shekhar, M. Coyle, D-R. Liu, B. Goyal, and S. Sarkar. Data Models in Geographic Information Systems. *Communication of the ACM*, 40(4), 1997.

[24] S. Shekhar, Andrew Fetterer, and Brajesh Goyal. Materialization Trade-Offs in Hierarchical Shortest Path Algorithms. In *Proc. Symposium on Large Spatial Database*, 1997.

[25] S. Shekhar and B. Hamidzadeh. Learning Transformation Rules for Semantic Query Optimization: A Data-Driven Approach. *IEEE Trans. on Knowledge and Data Eng. (Spl. Issue on Discovery in Databases)*, 5(6), 1993.

[26] S. Shekhar and D-R. Liu. CCAM: A Connectivity-Clustered Access Method for Aggregate Queries on Transportation Networks-A Summary of Results. *IEEE Transactions on Knowledge and Data Engineering*, 9(1), January 1997.

[27] S. Shekhar, S. Ravada, V. Kumar, D. Chubb, and G. Turn er. Parallelizing a GIS on a Shared Address Space Architecture. *IEEE Computer (Special Issue on Multiprocessors)*, 29(12), December 1996.

[28] S. Shekhar, T. A. Yang, and P. Hancock. An Intelligent Vehicle Highway Information Management System. *Intl. Jr. on Microcomputers in Civil Engineering (ISSN 0885-9507)*, 8(3), 1993.

[29] U.S. Geological Survey. Geographic information systems. http://www.usgs.gov/research/gis/title.html.

[30] T. Imielinski and Aashu Virmani and Amin Abdulghani. DataMine: Application Programming Interface and Query Language for Database Mining. In *Proc. KDD96*, 1996.