

Mining Knowledge in Geographical Data

Krzysztof Koperski

Jiawei Han

Junas Adhikary

Huge amounts of data have been stored in databases, data warehouses, geographic information systems, and other information repositories, and this data is still growing rapidly [4]. Companies are building data warehouses capable of storing hundreds of terabytes of data related to natural resource exploration; astronomical databases are measured in terabytes in size; and it is expected that the NASA Earth Observing System will transmit about 50 gigabytes of image data per hour.

This huge amount of data has posed great challenges to traditional data analysis methods for information and knowledge extraction. *Data mining*, or *knowledge discovery in databases* (KDD), has been emerging as a new research field and a new technology for *discovery of interesting, implicit, and previously unknown knowledge from large databases* [4]. Data mining represents the confluence of several research fields, including machine learning, database systems, data visualization, statistics, and information theory.

Besides many studies of knowledge discovery in relational and transaction databases, spatial data mining, which refers to *the extraction of implicit knowledge, spatial relationships, or other patterns not explicitly stored in spatial databases*, has attracted attention in recent research [2, 6, 7, 8, 10, 12]. Spatial data has many features distinguishing it from relational databases. It carries topological and/or distance information, usually organized by sophisticated, multidimensional spatial indexing structures, accessed by spatial data access methods, and often requiring spatial reasoning, geometric computation, and spatial knowledge representation techniques. Thus, spatial data mining demands an integration of data mining with spatial database technologies. 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.

Spatial data mining can be used for browsing spatial databases, understanding spatial data, discovering spatial relationships and relationships between spatial and nonspatial data, reorganizing spatial databases, constructing spatial knowledge-bases, optimizing spatial queries, etc. It is expected to have wide applications in geographic information systems, remote sensing, image database exploration, medical imaging, navigation, and many other areas where spatial data is used.

In this article, a short overview is provided to summarize recent studies on spatial data mining, including spatial data mining techniques, their strengths and weaknesses, how and when to apply them, and what are the challenges yet to be faced.

Statistical Spatial Analysis

Statistics was used as the most common approach for analyzing spatial data. Statistical analysis is a well studied area and therefore there exist a large number of algorithms including various optimization techniques. It handles numerical data well and usually comes up with realistic models of spatial phenomena. However, this approach is usually based on the assumption of statistical independence among the spatially distributed data which may cause problems as many spatial data are in fact interrelated, i.e., spatial objects are influenced by their neighboring objects. Kriging or regression models with spatially lagged forms of the dependent variables can be used to alleviate this problem to some extent. Unfortunately, it makes the whole modeling process more complicated and can only be done by experts with a fair amount

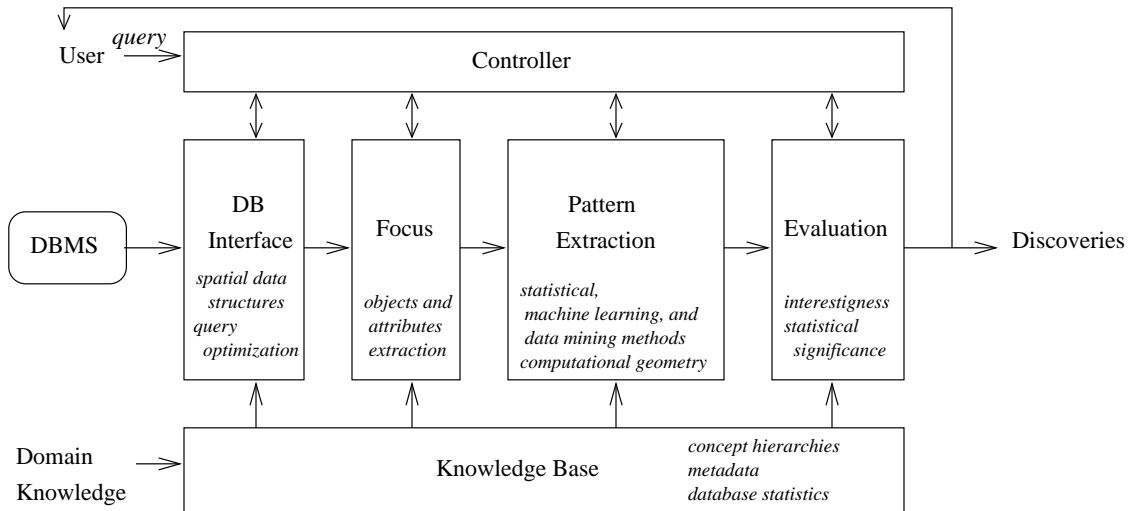


Figure 1: A Spatial Data Mining Process

of domain knowledge and statistical expertise. In other words, it is not the kind of technique which can be used by many end users for the analysis of spatial data. Furthermore, nonlinear rules cannot be modeled well, and symbolic values like names are handled poorly by the statistical approach. Statistical methods also do not work well with incomplete or inconclusive data and it is expensive to compute the results.

Spatial data mining methods allow extension of traditional spatial analysis methods by putting emphasis on efficiency, cooperation with database systems, better interaction with user, and discovery of new types of knowledge.

Using Relational Data Mining Techniques
Major *et al.* [9] used *IXLTM* commercial tool for mining of a tropical storm database. The goal was to predict if hurricanes can reach the U.S. territory. Data describing hurricanes were decomposed to observations at points. These observations were stored in a traditional relational database. Attributes like position of the hurricane, speed, direction, angle to the coast, etc. were used. Since multiple tuples describing the single hurricane in different points were stored, some data were interdependent. The interdependency of data causes problems because the algorithm which was used assumes independence of data. The GIS system was used to support the selection of the best rules. This study shows the necessity of extension of traditional data mining techniques toward spatial

data mining for better analysis of complex spatial phenomena and spatial objects.

The Mining Process

Spatial data mining process can be modeled using architecture¹ presented in Figure 1. During the data mining process, the user may control every step of knowledge discovery. Background knowledge, like spatial and non-spatial concept hierarchies, or information about the database, is stored in a knowledge base. Data is fetched from the storage using the *DB interface* which enables optimization of the queries. Spatial data index structures may be used for efficient processing. The *Focusing Component* decides which parts of data are useful for pattern recognition. For example, it may decide that only some attributes are relevant to the knowledge discovery task, or it may extract objects whose usage may promise good results. Rules and patterns are discovered by the *Pattern Extraction* module. This module may use statistical, machine learning, and data mining techniques in conjunction with computational geometry algorithms to perform the task of finding rules and relations. The interestingness and significance of these patterns is processed by *Evaluation* module to possibly eliminate obvious and redundant knowledge. The components may

¹Adapted from Matheus, C.J., Chan, P.K., and Piatetsky-Shapiro, G. Systems for Knowledge Discovery in Databases. *IEEE Trans. Knowledge and Data Engineering* 5,5 (Oct. 1993), 903-913.

interact between themselves through the *Controller* part, which also provides feedbacks for query refinement. Discoveries are finally passed to the user for verification.

Methods

Knowledge discovered from spatial databases can be of various forms including characteristic rules for general description of spatial data; discriminant rules discriminating or contrasting a class of spatial data from other class(es); association rules associating one or a set of features by another set of features; deviation and evolution rules describing temporal changes; or rules describing prominent structures or clusters. These rules may be presented in a number of forms and they can be used for description of various spatial objects. In the following sections, we categorize and describe a number of spatial data mining algorithms.

Generalization-Based Mining

Data and objects in databases often contain detailed information at primitive concept levels. It is often desirable to summarize a large set of data and present it at a high concept level. For example, one may like to summarize the detailed temperature and precipitation data in a region and present its general weather pattern. This requires *generalization-based data mining*, which first abstracts a large set of relevant data from a low concept level to relatively high ones and then performs knowledge extraction on the generalized data.

The generalization-based knowledge discovery assumes the existence of background knowledge in the form of concept hierarchies. Concept hierarchies can be explicitly given by the experts, or in some cases, can be generated automatically by data analysis. Two kinds of concept hierarchies, non-spatial and spatial, can be defined for spatial databases. With the ascension of concept hierarchy, information becomes more and more general, but still remains consistent with the lower concept levels. For example, for agriculture concept hierarchy both *jasmine* and *basmati* can be generalized to the concept *rice* which in turn can be generalized to concept *grains*, which also includes *wheat*. A similar hierarchy may exist for spatial data. For example, in a generalization process, regions representing counties can be merged to states and states can be merged to larger regions.

Attribute-oriented induction is an efficient data generalization technique (see Han and Fu

in [4]). It first takes a data mining query expressed in an SQL-like data mining query language and collects the set of relevant data in a database. Then, data generalization is performed by climbing the generalization hierarchies and summarizing the general relationships between spatial and non-spatial data at higher concept levels. It can be done on non-spatial data by (a) climbing the concept hierarchy when attribute values in a tuple are replaced by the generalized values, (b) removing attributes when further generalization is impossible and there are too many distinct values for an attribute, and (c) merging identical tuples. Induction continues until every attribute is generalized to the desired level. During the process of merging of identical tuples, the number of merged tuples is stored as *count* and aggregate values of some quantitative attributes may also be stored to enable quantitative presentation of the acquired knowledge.

The generalized data can be expressed in the form of a generalized relation or data cube on which many other operations can be performed to transform generalized data into different forms of knowledge. For example, drill-down or roll-up operations can be performed to view data at multiple abstraction levels; the generalized relation can be mapped into summarization tables, charts, curves, for presentation and visualization; characteristic and discriminant rules can be extracted; etc.

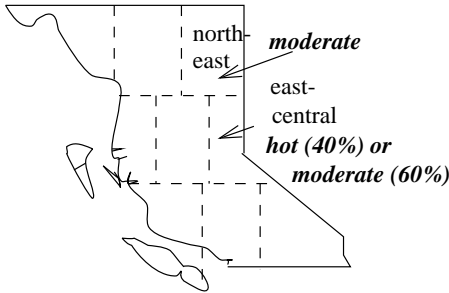
Lu *et al.* [8] extended attribute-oriented induction to spatial databases and presented two algorithms, *spatial-data-dominant* and *non-spatial-data-dominant* generalizations.

Spatial-Data-Dominant Generalization algorithm enables description of spatial regions using high level predicates. First, the algorithm merges spatial regions according to spatial hierarchy, which may lead to a map consisting of a small number of areas. Then, a non-spatial description of each area is produced using the attribute-oriented induction technique as described above. The answer to the query is the description of all regions using a disjunction of a few predicates which characterize each of the generalized regions (Figure 2).

Non-spatial-Data-Dominant Generalization algorithm creates maps which consist of a small number of regions sharing the same high-level non-spatial description. This algorithm starts with attribute-oriented induction on the non-spatial attributes, generalizing them to higher

Spatial-Data-Dominant Generalization

extract characteristic rule
from temperature-map
where province = "B.C." and
period = "summer" and year = 1990
in relevance to region and temperature.



Non-Spatial-Data-Dominant Generalization

extract region
from precipitation-map
where province = "B.C." and
and period = "spring" and year = 1990
in relevance to precipitation and region

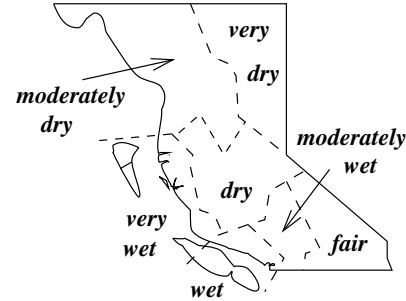


Figure 2: Generalization-Based Spatial Data Mining: Queries and Results

(more general) concept levels. Then, the neighboring areas with the same generalized attribute values are merged together. Approximation can be used to ignore small regions with different non-spatial description (Figure 2).

Data Clustering

Spatial data clustering identifies clusters, or densely populated regions, according to some distance measurement, in a large, multidimensional data set. The data space is usually not uniformly occupied by the data points. Data clustering identifies the sparse and the crowded places, and hence discovers the overall distribution patterns of the data set.

Data clustering has been studied in statistics, machine learning and data mining with different methods and emphases. Clustering analysis techniques proposed in data mining take large spatial databases into consideration.

CLARANS (Clustering Large Applications based upon RANdomized Search) algorithm was developed by Ng and Han [10] for cluster analysis to improve quality and efficiency of clustering. The clustering process in CLARANS can be presented as searching a graph where every node is a potential solution, i.e., a set of k medoids which are the most centrally located points in clusters. The clustering obtained after replacing a single medoid is called the *neighbor* of the current clustering.

The number of neighbors to be randomly tried is restricted by the parameter *maxneighbor*. If a better neighbor is found, CLARANS moves to the neighbor's node and the process is started again, otherwise the current clustering produces a local optimum. If a local optimum is found, CLARANS starts with new randomly selected node in search for a new local optimum. The number of local optima to be searched is also bounded by the parameter *numlocal*. As one can see CLARANS does not look through all solution space, but it does not confine itself to any specific sample. The computational complexity of every iteration in CLARANS is basically linearly proportional to the number of objects. CLARANS can be also used to find the most natural number of clusters k_{nat} and to detect outliers, e.g., points that do not belong to any cluster.

Based upon CLARANS, two spatial data mining algorithms were developed in a fashion similar to the approach discussed earlier: *spatial-dominant algorithm*, and *non-spatial-dominant one*. Both algorithms assume that the user specifies the type of the rule to be mined and relevant data through a learning request in a similar way as in an experimental database mining system, DBMiner (see Han and Fu in [4]). These algorithms use CLARANS for clustering and find high-level non-spatial description of objects in every cluster using attribute-

oriented induction. For example, using these algorithms one can find that in Vancouver expensive housing units are clustered in 3 locations. In the downtown cluster there are mainly expensive condominiums; in the waterfront cluster mansions and single houses are located; and the third cluster consists mainly of single houses.

CLARANS in Large Spatial Databases - Focusing Methods.

The efficiency of clustering algorithms may be significantly improved by using spatial data structures. One can improve efficiency of clustering algorithm using some sampling methods. However, bad sampling may lead to poor quality of clustering. Ester *et al.* [2] proposed the algorithm to improve the quality of sampling. The algorithm clusters only the most central objects of the leaf nodes of the R^* -tree. Because at the leaf nodes only neighboring points are stored, the loss of clustering quality is small and during the experiments it was reported to be from 1.5% to 3.2% whereas the speed of the clustering increased by a factor of 50, which was the number of points stored in a leaf node. Other proposed techniques use R^* -tree structure to perform computation only on pairs of objects which improves the efficiency of clustering comparing with checking all pairs of objects as done in the CLARANS algorithm.

Clustering Features and CF trees. R -trees are not always available and their construction may be time consuming. Zhang *et al.* [12] presented another algorithm, BIRCH (Balanced Iterative Reducing and Clustering), to deal with this problem. The presented method is the incremental one with the possibility of adjusting the memory requirements to the size of the memory that is available. The authors used two concepts called *Clustering Feature* and *CF tree*.

A *Clustering Feature CF* is a triple which summarizes information about subclusters of points. Given a set of points in a subcluster, $\{X_i\}$, CF is defined as $CF = (N, \vec{LS}, SS)$ where N is the number of points in the subcluster, \vec{LS} is the linear sum on N points, i.e., $\sum_{i=1}^N \vec{X}_i$, and SS is the square sum of data points, i.e., $\sum_{i=1}^N \vec{X}_i^2$. The *Clustering Features* are sufficient for computing clusters and they constitute an efficient storage information method as they summarize information about the subclusters of points instead of storing all points.

A *CF tree* is a balanced tree which stores Clustering Features. A point which is inserted to CF tree is added to the closest subcluster.

After the construction of the CF tree the Clustering Features from the leaf nodes can be used by any clustering algorithm. Using the algorithm about 100,000 points were clustered in 50 seconds in one experiment. The reported experiments showed the linear scalability of the algorithm with respect to the number of points, its insensibility to the input order, and the good quality of clustering.

Exploration of Spatial Associations

The methods discussed previously find only characteristic rules that describe spatial objects according to their non-spatial attributes. In many situations it is desirable to discover spatial association rules, the rules that associate one or more spatial objects with other spatial objects. The concept of *association rules* was introduced for mining large transaction databases (see Agrawal *et al.* in [4]). Koperski and Han [7] extended this concept to spatial databases. A spatial association rule is of the form $X \rightarrow Y (c\%)$, where X and Y are sets of spatial or non-spatial predicates and $c\%$ is the confidence of the rule. For example, the following rule is a spatial association rule: $is_a(x, school) \wedge close_to(x, sport_center) \rightarrow close_to(x, park) (80\%)$. This rule states that 80% of schools which are close to sport centers are also close to parks. There are various kinds of spatial predicates that could constitute a spatial association rule. Examples include topological relations like *intersects*, *overlap*, *disjoint*, *etc.*; spatial orientations like *left_of*, *west_of*, *etc.*; distance information, such as *close_to*, *far_away*, *etc.*.

In large databases, there may exist a large number of associations between objects but most of them will be applicable to only a small number of objects, or the confidence of rules may be low. For example, a user may not be interested in the relation associating 5% of houses and a single school, but may likely be interested in the rules that apply to at least 50% of houses. Two thresholds, *minimum support* and *minimum confidence*, control filtering out of associations describing small percentage of objects and rules with low confidence. The thresholds can be different at each level of non-spatial description of objects since using the same thresholds one may not find interesting associations at a low concept level because the number of objects satisfying the same predicates could be rather small.

The mining process is initiated by a query

which is to describe a class of objects S using other task-relevant classes of objects, and a set of relevant relations. For example, a user may want to describe urban parks by presenting the description of relations between parks and other objects like: railways, restaurants, zoos, roads, etc. Furthermore, the user can state that he/she is interested only in objects within one kilometer distance from a park. To minimize the number of costly spatial computations the algorithm first uses various approximations, like *maximum bounding rectangles*, and applies more detailed and finer, but more expensive spatial computations only to the patterns having large support at the approximation level. In one experiment, mining 22MB database using this algorithm took about 80 seconds.

Using Approximation and Aggregation

The clustering methods may answer questions *where* the clusters in the spatial database are. Another problem is to find out *why* the clusters are there. This question can be rephrased as “what are the characteristics of the clusters in terms of the features (objects) that are close to them”. The problem is how to measure the aggregate proximity, because statements like *90% of the houses in a cluster are close to feature F* are more informative and interesting than statements like *one house is close to a certain feature F*. The aggregate proximity is the measure of closeness of the set of points in the cluster to a feature as opposed to the distance between a cluster boundary and the boundary of a feature. Knorr and Ng [6] presented a method which enables fast finding of the features that are close to clusters. They used multiple level approximations in the algorithm CRH (where C is for encompassing circle, R for isothetic rectangle, and H for convex hull) to gradually reduce the candidate features. Approximation by circles and then by rectangles is used to eliminate features that have large aggregate distance to the cluster. Then, the algorithm calculates the aggregate proximity of points in the cluster to the convex boundary of each feature that passed through the previous filters. Finally, the algorithm reports the features with the best aggregate proximities showing minimum and maximum distances of points in the cluster to the feature, average distance, and percentages of points located in the distance less than specified thresholds. The algorithm CRH is experimentally reported to have the response time of less than two seconds for processing 50,000 features.

Mining Raster Databases

Exploration of raster databases like image data can be viewed as another case of spatial data mining. Below we present some studies on this topic.

Second Palomar Observatory Sky Survey (POSS-II) (see Fayyad *et al.* in [4]) used decision tree methods for the classification of galaxies, stars and other stellar objects from about 3 terabytes of sky images. Data images were pre-processed by low-level image processing system FOCAS, which selected objects and produced basic attributes like: magnitudes, areas, intensity, image moments, orientation, etc. Based on training data, which was classified by astronomers, sets of rules were constructed for the classification process. Normalization of the image plates using resolution, background level or average intensity, etc. increased the classification accuracy from about 75% when plates were not normalized to about 94% when they were normalized. The performance of decision tree methods was compared with neural networks. Neural networks appeared to be fairly unstable with accuracy varying from 30% to 95%. About 5×10^8 objects were classified. Obtained resolution was one magnitude better than in the previous astronomical studies and it was possible to classify objects with images too faint to be classified by astronomers.

Magellan Study [3] analyzed about 30,000 high resolution radar images of the surface of Venus. The goal was to identify volcanos, a task which would take about 10 man-years if performed manually. The system is composed of three basic components: *data focusing*, *feature extraction*, and *classification learning*. Like all other data focusing techniques, the first component increases the overall efficiency of the system by first identifying the portion of the image being analyzed that is likely to contain a volcano. This is achieved by comparing the intensity of the central pixel of a region to the estimated mean background intensity of its neighborhood pixels. The second component of the system extracts interesting features from the data. Standard methods used in pattern recognition like edge detection or Hough transformation, deal poorly with the variability and noise presented in the case of natural data. Since it is difficult to find attributes describing volcanos exactly, matrices containing images of volcanos were decomposed to eigenvectors. Eigenvalues were treated as attributes describing volcanos.

Then the final task, which is performed using decision trees, is to discriminate between volcanos and other objects looking like volcanos. The incidence angle of the synthetic aperture radar to the planet instrument strongly influenced images of volcanos.

The above studies showed the problems related to differences between images. The necessity of “normalization” of plates was shown to improve intra- and inter- plate classification.

Dealing with Uncertainty. In general, it is difficult for experts to provide classifications with 100% certainty and false classifications can produce large errors during classification because they are treated as negative examples. Smyth *et al.* (in [4]) used Magellan Study [3] to analyze such issues. The work is on the modeling and treatment of subjective label information given by the experts using probabilistic models. It shows that it is possible for the knowledge discovery methods to be modified to handle the lack of absolute ground truths.

Scalable Mining of Geoscientific Raster Data. Shek *et al.*'s [11] described a distributed parallel querying and analysis environment called CONQUEST (CONtent-based QUERying in Space and Time). CONQUEST can be distinguished from other raster database mining tools as it takes into account also temporal components of the datasets and it is designed to take advantage of parallel and distributed processing. The system enables extraction of complex spatio-temporal objects from massive datasets. CONQUEST defines a number of operations to be applied to geoscientific queries for description and extraction of objects. It also takes advantage of distributed processing by pushing some tasks to computers where data is stored. CONQUEST was tested on large climate datasets to detect cyclones and other weather features. For example, mining of 2 GB of data from 10 years history (resolution $2.5^\circ \times 2.5^\circ \times 12\text{hours}$) took 2430 seconds on 4 SS-20 SparcStations, and yielded 7304 instances of upward energy wave propagation events which can be used for example, to explain creation of ozone holes in the atmosphere.

Conclusions and Challenges

The initial success of mining knowledge from the huge amount of geographic data demonstrates the necessity and the potential of this young and promising research field. However,

there are many research issues which need extensive studies to make mining knowledge in large geographical databases a reality, as illustrated below.

- Mechanisms for generalization-based spatial data mining need to be further developed to handle multiple thematic maps and multiple-level interactive mining, and be integrated with spatial indexing, spatial access method, and data warehousing techniques.
- Additional data mining tasks and methods [4], such as data classification, pattern-based or similarity-based mining, and meta-rule guided data mining, should be studied for mining geographic data.
- Visualization of data mining results using spatial database techniques, interactive data mining by visual feedbacks [5], and the possible design of a spatial data mining language, should be analyzed.
- The measurements of interestingness of discovered patterns and the handling of uncertainty and incomplete information in spatial data mining using statistical, fuzzy logic, and rough sets approaches are also important research issues.
- Large amount of remotely sensed images asks for more data mining methods for exploration of such data including detection of anomalies, finding similar pictures and discovery of relationships between various phenomena [1].
- Mining from sources which are distributed over Internet/Intranet and stored in different formats is very important issue for practical applications. It should also include data cleaning and data integration.
- Methods for mining spatial data should be combined with advanced spatial databases, such as object-oriented spatial databases and spatio-temporal databases, as well as statistical analysis, spatial reasoning, and expert system technology to create *Intelligent GIS Systems*.

In summary, we believe that spatial data mining is an attractive and challenging research theme which may bring knowledge discovery power to geographic information systems and boost new GIS applications in the years to come.

References

- [1] Crompt, R.F. and Cambell, W.J. Data Mining of Multidimensional Remotely Sensed Images. In *Proceedings of 2nd International Conference on Information and Knowledge Management* (Nov. 1-5, Arlington, VA), ACM, New York, USA, 1993, pp. 471-480.
- [2] Ester, M., Kriegel, H.-P., and Xu, X. Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification. In *Advances in Spatial Databases, Proceedings of 4th Symposium, SSD'95* (Aug. 6-9, Portland, Maine). Springer-Verlag, Berlin, 1995, pp. 67-82.
- [3] Fayyad, U. and Smyth, P. Image Database Exploration: Progress and Challenges. In *Proceedings of 1993 Knowledge Discovery in Databases Workshop*, (July 11-12, Washington, DC). AAAI Press, Menlo Park, CA, 1993, pp. 14-27.
- [4] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., Eds. *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, Menlo Park, CA, 1996.
- [5] Keim, D. and Kriegel, H.-P. Visualization Techniques for Mining Large Databases: A Comparison. *IEEE Trans. Knowledge and Data Engineering* 8,6 (Dec. 1996), 923-938.
- [6] Knorr, E. M. and Ng R.T. Finding Aggregate Proximity Relationships and Commonalities in Spatial Data Mining. *IEEE Trans. Knowledge and Data Engineering* 8,6 (Dec. 1996), 884-897.
- [7] Koperski, K. and Han J. Discovery of Spatial Association Rules in Geographic Information Databases. In *Advances in Spatial Databases, Proceedings of 4th Symposium, SSD'95*. (Aug. 6-9, Portland, Maine). Springer-Verlag, Berlin, 1995, pp. 47-66.
- [8] Lu, W., Han, J., and Ooi, B. C. Discovery of General Knowledge in Large Spatial Databases. In *Proceedings of Far East Workshop on Geographic Information Systems* (June 21-22, Singapore). World Scientific, Singapore, 1993, pp. 275-289.
- [9] Major, J. and Mangano J. Selecting among Rules Induced from a Hurricane Database. In *Proceedings 1993 Knowledge Discovery in Databases Workshop*, (July 11-12, Washington, DC). AAAI Press, Menlo Park, CA, 1993, pp. 28-47.
- [10] Ng, R. and Han J. Efficient and effective clustering method for spatial data mining. In *Proceedings of 1994 International Conference Very Large Data Bases* (Sept. 12-15, Santiago, Chile). Morgan Kaufmann, San Francisco, CA, 1994, pp. 144-155.
- [11] Shek, E.C., Muntra, R.R., Mesrobian, E., and Ng, K. Scalable Exploratory Data Mining of Distributed Geoscientific Data. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (Aug. 2-4, Portland, Oregon). AAAI Press, Menlo Park, CA, 1996, pp. 32-37.
- [12] Zhang, T., Ramakrishnan, R., and Livny, M. BIRCH: an Efficient Data Clustering Method for Very Large Databases. In *Proceedings of ACM-SIGMOD International Conference on Management of Data* (June 4-6, Montreal, Canada). ACM, New York, 1996, pp. 103-114.

Additional references and figures can be found at <http://db.cs.sfu.ca/GeoMiner/survey>.

KRZYSZTOF KOPERSKI (koperski@cs.sfu.ca) is a Ph.D. candidate at School of Computing Science, Simon Fraser University, Canada.

JIAWEI HAN (han@cs.sfu.ca) is a professor of computing science at Simon Fraser University, Canada.

JUNAS ADHIKARY (adhikary@cs.sfu.ca) is a M.Sc. candidate at School of Computing Science, Simon Fraser University, Canada.

Current mailing address of the authors: School of Computing Science, Simon Fraser University, Burnaby, BC, Canada, V5A 1S6.