



GeoMiner: A System Prototype for Spatial Data Mining *

Jiawei Han Krzysztof Koperski Nebojsa Stefanovic

GeoMiner Research Group, Database Systems Research Laboratory

School of Computing Science

Simon Fraser University

Burnaby, BC, Canada V5A 1S6

E-mail: {han, koperski, nstefano}@cs.sfu.ca

Abstract

Spatial data mining is to mine high-level spatial information and knowledge from large spatial databases. A spatial data mining system prototype, GeoMiner, has been designed and developed based on our years of experience in the research and development of relational data mining system, DBMiner, and our research into spatial data mining. The data mining power of GeoMiner includes mining three kinds of rules: *characteristic rules*, *comparison rules*, and *association rules*, in geo-spatial databases, with a planned extension to include mining *classification rules* and *clustering rules*. The SAND (*Spatial And Nonspatial Data*) architecture is applied in the modeling of spatial databases, whereas GeoMiner includes the *spatial data cube construction module*, *spatial on-line analytical processing (OLAP) module*, and *spatial data mining modules*. A spatial data mining language, GMQL (*Geo-Mining Query Language*), is designed and implemented as an extension to *Spatial SQL* [3], for spatial data mining. Moreover, an interactive, user-friendly data mining interface is constructed and tools are implemented for visualization of discovered spatial knowledge.

1 Introduction

With the rapid progress of research into data mining and data warehousing in recent years [5], many data mining and data warehousing systems have been developed for mining knowledge in relational databases and data warehouses. Spatial data mining is a subfield of data mining that deals with the extraction of implicit knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases. With a huge amount of spatial data collected by satellite telemetry systems, remote sensing systems, regional sales systems and other data collection tools, it is crucial to develop tools for discovery of interesting

*Research is partially supported by the Natural Sciences and Engineering Research Council of Canada under the grant OGP0037230, by the Networks of Centres of Excellence Program (with the participation of PRECARN association) under the grants IRIS:IC-2 and HMI-5, and by research grants from BC Science Council, MPR Teltech Ltd. and the Hughes Research Laboratories.

Permission to make digital/hard copy of part or all this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication and its date appear, and notice is given that copying is by permission of ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee. SIGMOD '97 AZ, USA

© 1997 ACM 0-89791-911-4/97/0005...\$3.50

knowledge from large spatial databases. In addition, many relational databases contain spatial information as well, such as the residence of a customer or the location of a store. Therefore, it is important to mine knowledge related to both spatial and nonspatial objects in large databases. Unfortunately, there have not been many spatial data mining systems reported in previous studies.

Recent advances in the research on spatial data structures and spatial databases enable the creation of large spatial databases which can be queried in an effective way [3, 6]. These advances in combination with the researches into spatial reasoning [3] and the advances in data mining in relational databases [5] promote the research into spatial data mining [4, 10, 11, 12, 13, 14].

The GeoMiner research group in our *Database Systems Research Laboratory* has been working on data mining and especially spatial data mining for several years. Based on our previous research into spatial data mining [10, 12, 13] and our experience in the research and development of a relational data mining system DBMiner [7, 8, 9], a spatial data mining system prototype, GeoMiner, has been designed and developed. The current GeoMiner system includes the *spatial data cube construction module*, the *spatial on-line analytical processing (OLAP) module*, and *spatial data mining modules* which consist of data mining modules for mining *characteristic rules*, *comparison rules*, and *association rules*.

The SAND (*Spatial And Nonspatial Data*) architecture is applied in the modeling of spatial databases, whereas spatial data mining modules include mining both spatial knowledge and the relationships between spatial and nonspatial components. A spatial data mining language, GMQL (*Geo-Mining Query Language*), is designed and implemented as an extension to *Spatial SQL* [3]. Moreover, an interactive, user-friendly data mining interface is constructed and tools are also implemented for visualization of spatial data and the discovered spatial knowledge.

A more detailed description of the GeoMiner system is presented in Section 2. A summary and a discussion of our on-going research are in Section 3.

2 GeoMiner: A database mining system prototype

The GeoMiner system is an extension and evolution from a relational data mining system, DBMiner, researched and developed in our laboratory [9]. The DBMiner system (see references and the system via <http://db.cs.sfu.ca/DBMiner>) currently contains the following five data mining functional modules: *characterizer*, *comparator*, *associator*, *predictor*,

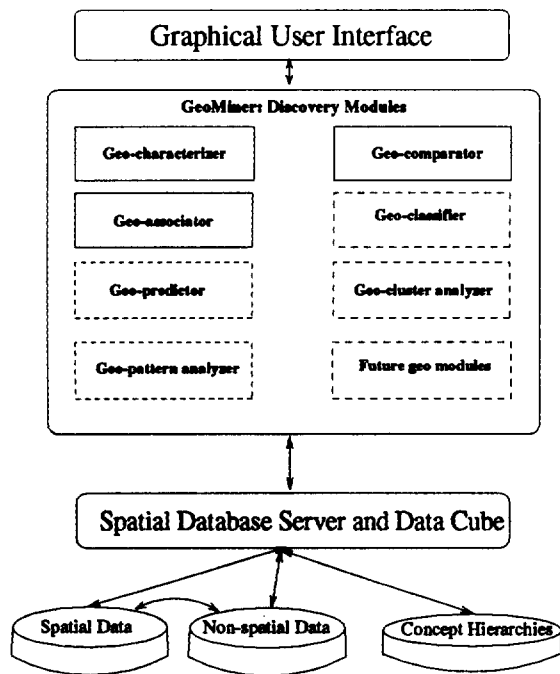


Figure 1: General architecture of GeoMiner

and classifier. Several additional data mining modules, including mining time-related data, are at the research and development stage. DBMiner is implemented by integration of data mining and data warehousing techniques, including data cube construction and manipulation [9], attribute-oriented induction [7], multi-level association analysis [8], statistical data analysis, machine learning, etc., for mining relational data.

GeoMiner is constructed on top of the DBMiner system. The functions for mining nonspatial data are directed to the DBMiner system; whereas those for mining spatial data and the relationships between spatial and nonspatial data are performed by the dedicated GeoMiner functions.

The major features of the current system include mining three kinds of knowledge rules in spatial databases, the integration of data mining and data warehousing technologies, interactive mining of multi-level rules, integration with commercial relational databases and geographic information systems, good performance in analyzing large spatial data sets, and multiple forms of outputs, including generalized maps, generalized relations, cross-tabulation, multiple level rules, charts, curves, etc.

Figure 1 shows the general architecture of GeoMiner which consists of (1) a graphical user interface for interactive mining and the display of data mining results in the form of tables, charts, maps, etc.; (2) a set of discovery modules, including three existing modules: *geo-characterizer*, *geo-comparator*, and *geo-associator*, and four planned ones: *geo-classifier*, *geo-predictor*, *geo-cluster analyzer*, and *geo-pattern analyzer*; (3) a spatial database server, which may include *MapInfo*, *ESRI/Oracle SDE*, *Informix-Illustra*, or other spatial database engines; and (4) the data- and knowledge-base, storing nonspatial data, spatial data, and concept hierarchies.

The functionalities of the three existing modules are described as follows.

- Geo-characterizer.

This module mines a set of characteristic rules at multiple levels of abstraction from a relevant set of data in a spatial database. It provides users with a multi-level, multi-angle view of the data in the database.

Two spatial mining algorithms, spatial-dominated and nonspatial-dominated generalization, have been studied in our previous work [12], which correspond to two different orientations in data generalization, similar to the answering of the following two questions: (1) *given spatial hierarchies of Western Canada, describe general weather patterns according to region partitions*; and (2) *given nonspatial hierarchies, such as temperature, precipitation, etc., describe (i.e., outline) the regions in Western Canada based on their generalized weather patterns*.

- Geo-comparator.

This module mines a set of comparison rules which contrast the general features of different classes of the relevant sets of data in a database. It compares one set of data, known as the target class, to the other set(s) of data, known as the contrasting class(es).

For example, the module may show *the differences in weather patterns between British Columbia (B.C.) and Alberta*, or find *the clusters or features related to the locations which differentiate the profiting shops from the non-profiting ones*.

- Geo-associator.

This module finds a set of strong spatial-related association rules from the relevant set(s) of data in a spatial database. An association rule shows the frequently occurring patterns (or relationships) of a set of data items in a database. A typical spatial association rule is in the form of " $X \rightarrow Y(s\%, c\%)$ " where X and Y are sets of spatial or nonspatial predicates, $s\%$ is the support of the rule (the probability that X and Y hold together among all the possible cases), and $c\%$ is the confidence of the rule (the conditional probability that Y is true under the condition of X). An efficient algorithm for mining spatial association rules has been proposed in our previous study [10].

An example of mining spatial association rules is as follows, "*what are the relationships among Canadian towns, their population, closeness to water, and closeness to the national border?*" One possible association rule among many to be found is "*if a Canadian town is large and is adjacent to large water body it is close to the U.S. border, with the possibility of 78%.*"

In the development of these functional modules, several data mining techniques are adopted based on the results of our previous research, including *attribute-oriented induction* [7], *spatial-dominated and nonspatial-dominated generalization techniques* [12], *progressive deepening at mining multi-level association rules* [8], *exploration of refined geo-algorithms step-by-step* [10], etc.

It is important to note that the flexibility of GeoMiner is at its ability to perform multi-level spatial mining. Data and objects in spatial databases often contain detailed information at the primitive level of abstraction. It is often desirable to summarize a large set of data and present it at a high (abstraction) level. For example, one may like to summarize the detailed temperature and precipitation data

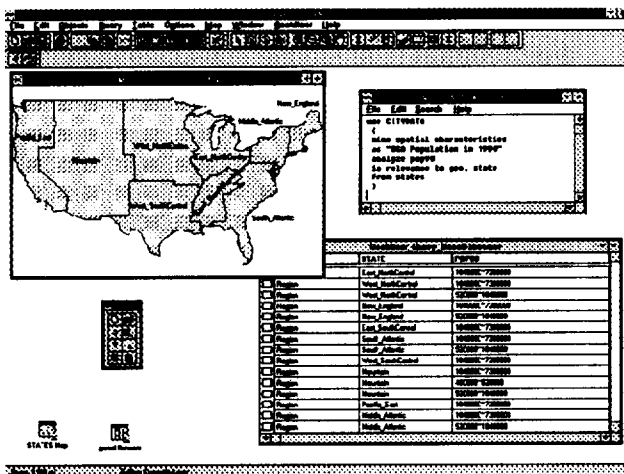


Figure 2: Mining Spatial Characteristics using GeoMiner

in an area 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. However, with blind generalization, data may be generalized to too high level. Users may like to drill-down to specific features or regions to examine the details. Therefore, the system provides not only the power of generalization, but also the power of specialization for flexible concept description (i.e., characterizer and comparator). This feature is similar to that of OLAP in many data warehouses. However, a plain OLAP processor cannot find general concept descriptions, associations, and classifications, without proper extensions, whereas a data miner will find such distinct features at multiple levels of abstraction.

A *Geo-Mining Query Language*, GMQL, is designed and implemented as an extension to Spatial SQL, for spatial data mining.

The following examples demonstrate how to use GMQL and perform data mining in geo-spatial databases.

Example 2.1 To characterize *weather patterns* in *B.C.* by displaying the regions in relevance to temperature and precipitation, the spatial data mining query can be expressed in GMQL as follows.

```
mine spatial characteristics
as "B.C. temperature and precipitation distribution"
analyze geo
in relevance to temperature, precipitation
from weather_probe
where time_period = "summer" and year = 1996
and area_name = "British-Columbia"
```

To process this query, the system first retrieves the relevant set of data by processing a spatial database query, which retrieves all the weather probes in B.C., with the data related to *temperature*, *precipitation*, and the associated *area_name*, and with *time_period* being in the summer of 1996.

Based on the given query and the concept hierarchies for temperature and precipitation, a non-spatial-dominated generalization is performed as follows.

First, generalization is performed on the non-spatial hierarchies: temperature and precipitation, which may generalize temperature values into *cool*, *warm*, *hot*, *very hot*, etc.,

and precipitation values into *very wet*, *wet*, *moderately-wet*, *moderately-dry*, etc. Such nonspatial generalization triggers the merge of the connected regions where the probes are located, with the same (generalized) temperature and precipitation descriptions. The spatial merge generates a set of consolidated regions. Some approximation algorithms (such as smoothing or ignoring minor outliers or island regions) can be applied, and the generalization can terminate when the number of distinct merged regions reaches a small number defined by a prespecified *generalization threshold*.

Notice that drill-down or roll-up can be performed interactively on such generalized data to zoom-in or zoom-out the generalized spatial region or to examine the details of their associated nonspatial properties. Concept hierarchies can be created by users or domain experts, but also generated automatically based on value distribution or some clustering algorithms.

□

In Example 2.1, mining is performed by generalizing along the nonspatial hierarchy which triggers the merge of spatial regions. Thus such generalization is called *nonspatial data-dominated generalization* [12].

Alternatively, spatial data-dominated generalization can be performed if a spatial region hierarchy is given. Such a generalization process may first generalize small regions to some large regions and the numeric data, such as temperatures and precipitation, associated with the small regions can be averaged or clustered in the large regions for better characterization. Figure 2 is a snapshot of such a mining process.

Example 2.2 The task of comparison the *weather patterns* of *B.C.* vs. *Alberta* by displaying the precipitation (map) with respect to temperature and area partition can be expressed in GMQL as follows.

```
mine spatial comparison
as "Precipitation: B.C. vs. Alberta"
for "B.C."
where area_name = "British-Columbia"
versus "Alberta"
where area_name = "Alberta"
analyze precipitation, geo
in relevance to temperature, area_name
from weather_probe
where time_period = "July" and year = 1996
```

To process this query, the system first retrieves the relevant set of data including *temperature*, *precipitation*, and *area_name*, from the relation *weather_probe* in B.C. and Alberta, within the *time_period* of July 1996. The collected data are partitioned into two contrasting classes: "*British Columbia*" and "*Alberta*". Generalization and specialization are based on the hierarchies associated with two dimensions: *temperature* and *area_name*. The measurements include two attributes, *precipitation* and *geo*.

□

Furthermore, spatial association rules can be mined in geo-spatial databases based on the similar philosophy for mining association rules in transaction databases [1]. We examine the following example.

Example 2.3 To find strong geo-spatial associations describing *large towns* in *British Columbia*, a GMQL mining query can be defined as follows.

```

mine spatial associations
as "large_BCtown"
in relevance to water.name, states.area_name
from towns, water, states, provinces
where towns.population > 25000 and
    towns.geo inside provinces.geo and
    provinces.area_name = "British_Columbia" and
    g_close.to(towns.geo, water.geo, 10, "km") and
    water.area > 3 and
    g_close.to(towns.geo, states.geo, 75, "km") and
    states.area_name = "USA"

```

This query is to find association relationships between large BC towns and large water bodies and areas in USA. Multi-level association rules can be mined by climbing up and stepping down along any of the two dimensions (name of water bodies or area_names in the states relation). The detailed association rule mining algorithms have been studied by many researchers [2], and our spatial mining method is based on our study on the methods for mining spatial association rules [10].

For example, the following association rules may be derived from the spatial data set.

```

is_a(X, "large_BCtown")  $\wedge$   $\rightarrow$  close.to(X, "sea") (40%, 52%).
is_a(X, "large_BCtown")  $\wedge$  close.to(X, "Georgia_Strait")
 $\rightarrow$  close.to(X, "USA") (30%, 90%).

```

Notice that the two numbers in parentheses (*s*, *c*) indicate the support and confidence of the rule, respectively. \square

3 Conclusions

We have designed and developed an interesting spatial data mining system prototype, GeoMiner. The current system mines three kinds of spatial knowledge from geo-spatial databases and demonstrates a promising direction in the study of spatial data mining. According to our research and development plan, more data mining functionalities will be incrementally added into the system.

The sample spatial database, that we use for demo is a TIGER/Line'95 U.S. Western states spatial database together with the census data for U.S. counties and cities. The TIGER database contains spatial information about hydrology, transportation and other objects like national and state parks, churches, cemeteries, universities, etc.

This preliminary experimental and development work on spatial data mining has promoted our further study in this direction. Some on-going themes that we plan to develop in the near future are illustrated as follows.

- Further enhancement of the power and efficiency of the knowledge discovery mechanisms, including the improvement of rule quality and system performance for the existing functional modules.
- So far, we have been considering knowledge discovery only from single thematic maps. However, we are exploring efficient algorithms that handle multiple thematic maps as well.
- Besides the functional modules mentioned above, we plan to add *geo-predictor* and *geo-classifier*. Their relational counterpart, *predictor* and *classifier*, have been implemented in the DBMiner system. Moreover, there have been some interesting studies on efficient spatial clustering algorithms, such as [4, 13, 14]. It is

important to include some spatial clustering functionality in the GeoMiner system. Finally, we investigate pattern analysis and plan to incorporate it into the GeoMiner system.

- Further development of high-level, user-friendly interfaces for interactive spatial data mining and visual presentation of the discovered knowledge.

4 Acknowledgments

The authors would like to express their thanks to the DBMiner Research group, especially to Jenny Chiang, Yijun Lu, and Sonny Chee for providing the necessary API to the DBMiner data mining engine.

References

- [1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. 1994 Int. Conf. Very Large Data Bases*, pages 487–499, Santiago, Chile, September 1994.
- [2] M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8:866–883, 1996.
- [3] M. Egenhofer. Spatial SQL: A query and presentation language. *IEEE Transactions on Knowledge and Data Engineering*, 6:86–95, 1994.
- [4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. In *Proc. of the Second International Conference on Data Mining KDD-96*, pages 226–231, Portland, Oregon, August 1996.
- [5] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996.
- [6] R. H. Güting. An introduction to spatial database systems. *The VLDB Journal*, 3:357–400, 1994.
- [7] J. Han, Y. Cai, and N. Cercone. Data-driven discovery of quantitative rules in relational databases. *IEEE Trans. Knowledge and Data Engineering*, 5:29–40, 1993.
- [8] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In *Proc. 1995 Int. Conf. Very Large Data Bases*, pages 420–431, Zurich, Switzerland, Sept. 1995.
- [9] J. Han, Y. Fu, W. Wang, J. Chiang, W. Gong, K. Koperski, D. Li, Y. Lu, A. Rajan, N. Stefanovic, B. Xia, and O. R. Zaiane. DBMiner: A system for mining knowledge in large relational databases. In *Proc. 1996 Int'l Conf. on Data Mining and Knowledge Discovery (KDD'96)*, pages 250–255, Portland, Oregon, August 1996.
- [10] K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *Proc. 4th Int'l Symp. on Large Spatial Databases (SSD'95)*, pages 47–66, Portland, Maine, Aug. 1995.
- [11] K. Koperski, J. Han, and J. Adhikary. Mining knowledge in geographic data. In *Comm. ACM (to appear)*, 1997.
- [12] W. Lu, J. Han, and B. C. Ooi. Knowledge discovery in large spatial databases. In *Far East Workshop on Geographic Information Systems*, pages 275–289, Singapore, June 1993.
- [13] R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. In *Proc. 1994 Int. Conf. Very Large Data Bases*, pages 144–155, Santiago, Chile, September 1994.
- [14] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data*, pages 103–114, Montreal, Canada, June 1996.