Volume E-3 No. 8 (12)

A Visual Framework for Spatial Data Mining

Nguyen Vinh Nam [(1)], **Le Hoai Bac** [(2)]
(1) Vietnam Informatics and Mapping Corporation Ho Chi Minh City, Vietnam
(2) Faculty of Information TechnologyUniversity of Science, VNU - HCMC, Vietnam
Email: nguyenvinhnam@vietbando.vn, lhbac@fit.hcmus.edu.vn

***Abstract* -The unique properties of spatial data provide challenges and opportunities for researching new methods in spatial data mining. In this article, we propose an interoperable framework that integrates Geographic Information System (GIS) with the spatial data mining processto facilitate spatial data preparation, to extract spatial relationships that can take advantage of traditional data mining toolkits such as Weka, and to reveal significant spatial patterns. With this approach, it'svery straightforward to adopt spatial access methods and spatial query processing algorithms foran efficient data mining technique. Moreover, our framework visually supports the complete spatial data mining process.**

Keywords - Spatial data mining, Visual data mining, GIS

1. INTRODUCTION

With advanced data collection techniques such as remote sensing, census data acquiring, weather and climate monitoring etc. contemporary geographic datasets contain an enormous amount of data of various types and attributes. These data are stored in *Relational Database Management Systems* (RDBMS) (e.g. Oracle Spatial, IBM DB2 Spatial Extender, MySQL 5, SQL Server 2008 and later, PostGIS), or in specific formats of commercial GIS applications (e.g. MapInfo, ArcGIS, Intergraph). Spatial Data Mining (SDM) can be an appropriate technique for detecting possible interesting patterns in geographic datasets. SDM is a knowledge discovery process of extracting implicit interesting knowledge, spatial relations, or other patterns not explicitly stored in databases [2, 3]. The ever increasing and diverse nature of spatial data has created new challenges for data mining: interactions with spatial databases, spatial data processing, visualization and modeling discovered knowledge etc... Data mining has been widely treated as a synonym of *Knowledge Discovery in Databases (KDD)*.

The KDD process consists of an iterative sequence of the five major steps [4]: selection, preprocessing, transformation, data mining and evaluation/ interpretation. Data are prepared for mining in steps 1 through 3. For non-spatial databases, the data preparation process requires between 60 and 80 percent of the time and effort in the whole KDD [5]. For spatial databases, this problem increases significantly because of the unique properties of spatial data such as spatial dependency, spatial heterogeneity, and data type variety.

Geographic Information System - GIS has a long history of being used as a tool to capture, to store, to check, to integrate, to manipulate, to analyze and to visualize spatial attribute and statistical data [1]. The availability of features such as spatial and non-spatial data query and selection, topology (defining and enforcing data integrity rules), classification, map overlays, network analysis, and thematic map creation, makes GIS a useful tool for spatial data mining. Visualization by GIS gives user the ability to spot spatial errors that are often omitted by analyzing raw data, and to aid visual analysis and detection of spatial distributions and their patterns. Visualization is a powerful strategy for integrating high-level human intelligence and knowledge into the KDD process. Therefore, the integrating of GIS and data mining techniques brings promises of a solution to many challenges in spatial data mining.

In this paper, we propose a framework to integrate GIS into the steps of data mining process to improve the efficiency of knowledge discovery in spatial databases.

The rest of the paper is organized as follows. Section 2 describes main modules of our framework architecture. Section 3 presents the experimental results. Finally, we discuss conclusions and future work in section 4.

2. FRAMEWORK ARCHITECTURE

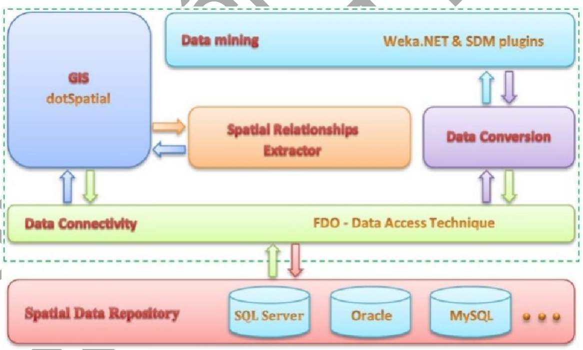

Fig. 1. The framework architecture

The framework is designed to integrate GIS into spatial data mining process with open, highly extensible architecture (e.g. plugin model). It has five major modules described in Fig. 1.

2.1. Spatial Database Connectivity

A huge amount of spatial data is stored in various formats. These formats may be closed or opened, normalized or not. Consequently, preparation and/or analysis may be difficult or time-consuming due to the heterogeneity of formats. To avoid these drawbacks and in order to facilitate accesses and usage of these data,we use FDO (Feature Data Object) Data Access Technique as a major module in our framework for storing, retrieving, updating, and analyzing GIS data. FDO uses a provider-based model for supporting a variety of geospatial data sources, where each provider typically supports a particular data format or data store [6]. FDO Geometry is based on the OpenGIS Simple Features Implementation Specification for SQL [7] which are implemented by most spatial databases. The FDO XML format for schema is based on the Open GIS Consortium Geography Markup Language [8]. FDO is a free, open source software. With this open standard approach, we can manipulate most available spatial databases and can expand as needed. In current release, FDO works with popular geospatial data format such as SDF, SHP, ArcSDE, WFS, WMS, ODBC, MySQL, GDAL, OGR, SQL Server Spatial, SQLite Spatial.

2.2. GIS

Spatial data mining deals with spatial data that includes numbers, categories, and extended objects such as lines, points and polygons. Moreover, it works with implicit spatial predicates (e.g. touch, contain) and high autocorrelation among nearby features. So, the embedding of GIS in spatial datamining helps facilitate spatial data preparation, spatial data visualization and interesting pattern discovery by creating thematic maps (see Fig.2). A spatial data mining system prototype,GeoMiner,is such system that is implemented on the top of MapInfo Professional 4.1 commercial GIS software [9]; however, it is no longer available due to practical purposes.

In our framework, the GIS module is *dotSpatial*, an open source geographic information system library that is used to incorporate spatial dataanalysis and mapping into spatial data mining process [10]. This GIS library supports plugin architecture so we can develop GIS extensions as well as specialized spatial data mining algorithms.

Fig. 2 GIS visualize spatial data and interesting patterns by thematic map

Each feature class (layer) consists of a spatial attribute (geometry column) and one or many non-spatial attributes. The latter can be properties describe their spatial object, result of spatial predicates or interesting patterns found. They are stored in table format and are linked with spatial records by feature identifier (Fid) as shown in Fig.3. Then, these attributes are used as controlling parameters to create thematic maps for data analysis.

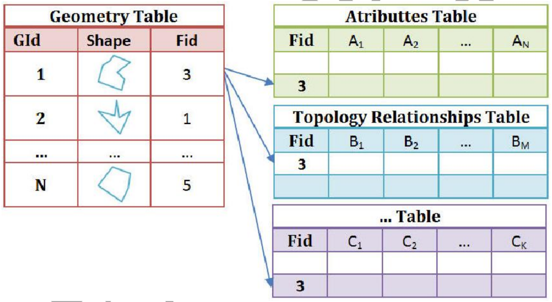

Fig. 3. Linking spatial data with non-spatial data

In our framework, GIS is used in the three stages of spatial data mining process:

a. *Data Preparation*

- Spatial data preprocessing
  - Data selection
  - Removal of noise, outliers and dependences
  - Data transformation: convert into the right format for spatial data mining process
  - Data projection
  - Data aggregation (e.g. buffering)
- Spatial Relationships Extraction (*e.g. touch, nearby, overlap, contain*)

b. *Data mining*

- Use GIS functionalities for an efficient implementation of spatial data mining algorithms (e.g. spatial index, spatial feature extraction)

c. *Data & Patterns Analysis*

- Usevisual techniques for data analysis and pattern visualization:
  - Map based techniques
  - Chart based techniques
  - Projection techniques
  - Pixel techniques
  - Iconographic techniques
  - Network methods

2.3. Spatial Relationships Extractor

Spatial relationships are usually not explicitly stored in spatial data repositories, so they must be computed and materialized with spatial operations. Spatial relationships extraction poses the challenge of revealing meaningful information of spatial objects, with a particular interest in their relationships. There are three basic types of spatial relations:

a. *Topological relations* are invariant under topological transformations, i.e. they are preserved if both are translated, scaled or rotated simultaneously. The topological relations are derived from the nine intersection model of the interiors (denoted by G^o), the boundaries (denoted by δG) and the complements (denoted by G^-) between two objects. In particular, given two geometries A and B, the nine possible intersections defining the relation between these geometries are represented by the 9-intersection matrix:

$$R(A,B) = \begin{pmatrix} \delta A \cap \delta B & \delta A \cap B^o & \delta A \cap B^- \\ A^o \cap \delta B & A^o \cap B^o & A^o \cap B^- \\ A^- \cap \delta B & A^- \cap B^o & A^- \cap B^- \end{pmatrix}$$

These relationships can be classified into*Equals, Contains, Within, Crosses, Disjoint, Overlaps* and *Touches*. In general, the spatial operations for relationship extraction are computationally expensive, so theyshould be done prior to data mining. Fortunately, these relations can be extracted into attribute table by spatial queries supported by both GIS module and SDBMS. Topological relations are often used as spatial predicates in spatial association rules. For example:

$is\_a(x, house) \wedge within(x, tower) \rightarrow is\_expensive(x)$ (90%)

b. *Distance relations* are based on distance between two spatial objects. Distance function depends on the spatial datatype used (e.g. geography, geometry).

c. *Direction relations* deal with where spatial objects are located in space. In literature, each spatial object is assigned a representative point, then these points are used to determine direction relation between two objects [11]. For example, B northeast A holds,$iff\ \forall b \in B: b_x \geq rep(A)_x \wedge b_y \geq rep(A)_y$. With this approach, direction relation depends on representative point selection for spatial object, so it is not often unique.

We propose a new method for calculating exact direction relation between any two shaped objects. Our algorithm is depicted as follows:

- Find two external tangents of two geometries
- Find intersection of found tangents
- Construct bisector of the angle between these tangents
- Determine the direction for each geometry object corresponding to their position along this bisector.

Fig.4 shows some pairs of geometries and constructs bisector for each case: point and point, point and polygon, point and line, line and line, line and polygon, and polygon and polygon.

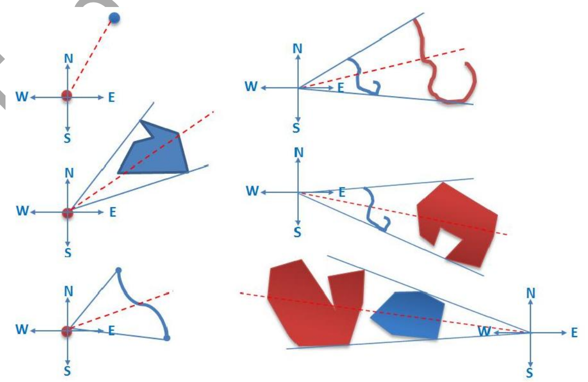

Fig. 4. *Calculating Direction relations*

2.4. Data Conversion

This module converts non-spatial data stored in attribute table to input format of available traditional data mining toolkits (e.g. Weka) and vice versa.

2.5. Data mining

In our framework, spatial data mining can be performed in two ways:

- Extract and materialize spatial relationships in non-spatial data then apply traditional data mining toolkits or by spatial data mining. We construct models to store interest spatial relationships, then these relationships are extracted and converted to the input format (e.g. table) of traditional data mining toolkits. Currently, we support building some spatial relationship models, such as VPF (**V**ector **P**roduct **F**ormat) to store adjacent spatial relationships for polygon objects, TIN (**T**riangulated **I**rregular **N**etwork) and CDT (**C**onstrained **D**elaunay **T**riangulation) to store neighbor relationships for point objects. This approach allows reusing standard data mining algorithms on features extracted. Researchers themselves may also create new models as needed using our plugin architecture.

- Develop specialized spatial data mining algorithms can be directly applied to spatial and non-spatial data. This approach can dynamically exploit the spatial relationships during discovery process and flexibly uses spatial knowledge. These algorithms can be developed directly in our GIS environment.

3. EXPERIMENTS

In thissection, we consider some examples of spatial data mining using our framework: input data and output data of specialized spatial data mining algorithms are stored as vector layers or attribute tables linked to some given vector layers, so they can be visualized to analyse revealed patterns.

3.1. Analysiswith Touch relation (Topological relation)

In this example, the user works with a layer of province boundaries of Vietnam. He wants to create some thematic map of adjacent provinces. We implemented an algorithm [14] as a plugin to extract adjacent relations and store them in *LeftFace* and *RightFace* fields of VPF model, see Fig. 5.

GeoId	StartNode	EndNode	LeftEdge	RightEdge	LeftFace	RightFace
1	297,884	297,883	2	3	185	3,280
2	297,882	297,884	4	1	185	3,280
3	297,883	297,878	1	9	185	3,280
4	297,876	297,882	10	2	185	3,280
5	297,877	297,881	9	6	185	3,280
6	297,881	297,880	5	7	185	3,280
7	297,880	297,879	6	8	185	3,280
8	297,879	297,875	7	11	185	3,280
9	297,878	297,877	3	5	185	3,280
10	297,874	297,876	12	4	185	3,280
11	297,875	297,872	8	15	185	3,280

Fig.5. Adjacent relationships among provinces are stored in a table

Then, he uses thematic map engine to display each province with a color corresponding to number of provinces touching it (see Fig.6). Fig.7 (left) shows provinces that have number of adjacent provinces greater than three. Fig.7 (right) shows provinces that have number of adjacent provinces greater than five.

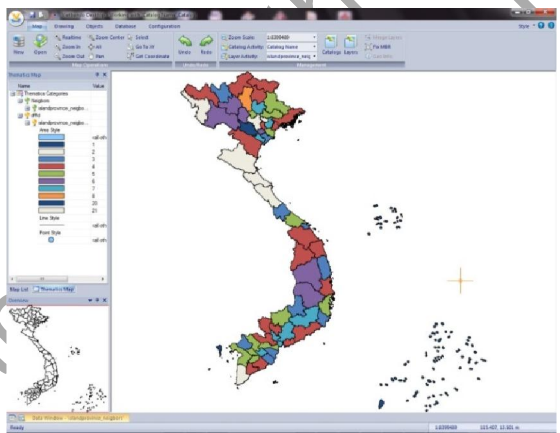

Fig. 6 Thematic map about adjacent provinces

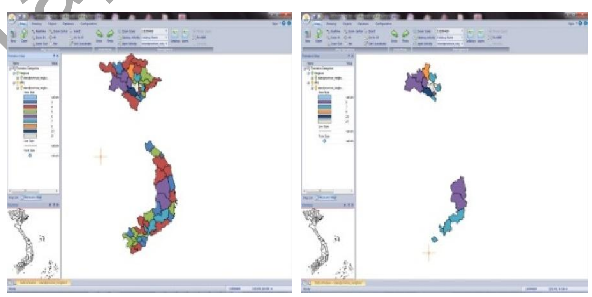

Fig.7. *Provinces adjacent more than 3 (left) and 5 (right) provinces*

3.2. Analysis with Clustering

We developed some spatial clustering algorithms that could directly access spatial data [12, 13]. These algorithms used spatial indexing techniques such as R-tree and Quad-tree to improve clustering effectively. In this example, the input data is a bitmap of circuit diagram of computer mainboard; our clustering algorithms were used to detect defects. In Fig. 8, clusters are colorized using thematic map tools.

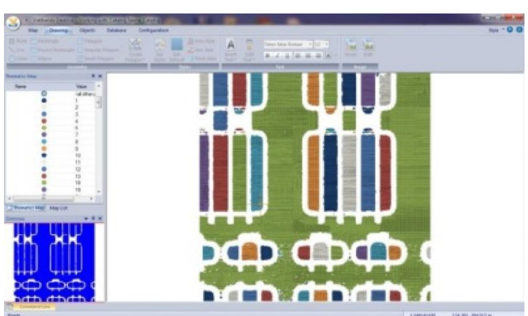

Fig.8. Clusters are found by our algorithm

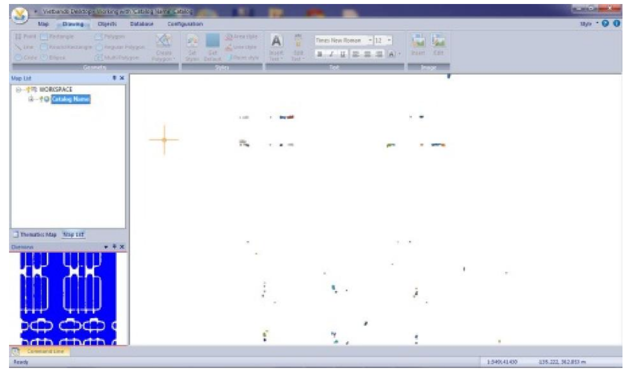

Fig.9. Defects correspond to the thresholds: 200pixels

To analyse defects, the user defines a defect threshold to distinguish real circuits from defects in the mainboard, then uses thematic map tool to highlight defects as you see in Fig. 9 and Fig.10.

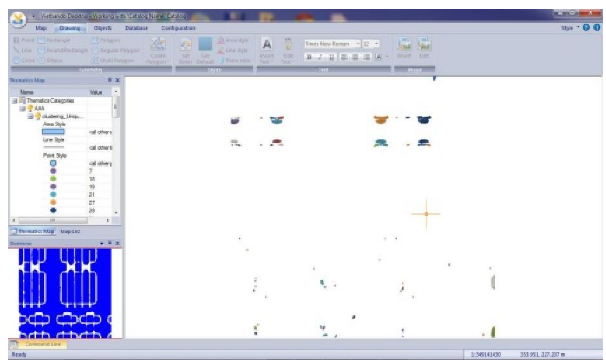

Fig.10 Defects correspond to the thresholds: 50 pixels

4. CONCLUSTION AND FUTURE WORK

This paper proposes a framework for integrating GIS with spatial data mining process. Our approach offers the advantage of:

- Visualizing spatial data mining steps
- Visualizing interesting patterns with thematic map toolkit
- Dynamically exploiting the spatial features in the discovery process
- Reusing traditional data mining algorithms
- Extracting and analyzing interesting patterns from large spatial databases with our visual framework.

In the future, we will study to integrate available methods for pattern visualization in spatial data mining andintegrate workflow into our framework to automate discovery process.

REFERENCES

[1] Shahab Fazal, *GIS Basics*, New Age, 2008.

[2] Koperski, K., Adhikary J., Han, J. "Knowledge discovery in spatial databases: Progress and Challenges", *In Proceedings of the SIGMID workshop on research issue in data mining and knowledge discovery*, Technical report 96-08. University of British Columbia, Vancouver, Canada, 1996.

[3] KoperskiK., Han J.,"Discovery of Spatial Association Rules in Geographic Information Databases", *Proc. 4th Int. Symp. on Large Spatial Databases*, 47-66. Berlin: Springer, 1995.

[4] Jiawei Han, Micheline Kamber, Jian Pei,*Data mining: Concepts and Techniques*, Elsevier, 2012.

[5] Adriaans, P. and Zantinge, D., *Data mining*, Addison Wesley Longman, Harlow, England, 1996.

[6] http://fdo.osgeo.org/

[7] http://portal.opengeospatial.org/files/?artifact_id=829

[8] http://portal.opengeospatial.org/files/?artifact_id=1034

[9] Han J., Koperski K., Stefanvic N.,"GeoMiner: a system prototype for spatial data mining", *In Proceedings of the ACM-SIGMOD international conference on Management of Data (SIGMOD'97)*. ACM Press, Tucson, AR, 553-556. 1997.

[10] http://dotspatial.codeplex.com

[11] Martin Ester, Hans-Peter Kriegel, Jörg Sander, *Algorithms and Applications for Spatial Datamining*, Research Monographs in GIS, Taylor and Francis, 2001.

[12] Nam Nguyen Vinh, Bac Le. "Incremental Spatial Clustering in Data Mining using Genetic Algorithm and R-tree". *SEAL* 2012, LNCS 7673, pp. 270–279, 2012.

[13] Nam Nguyen Vinh, Bac Le. "Simple Spatial Clustering Algorithm based on R-tree". *MIWAI* 2012.

[14] Nam Nguyen Vinh, Bac Le. "Constructing and modeling parcel boundaries from a set of lines for querying adjacent spatial relationships". *ICCSA* 2013.

AUTHORS' BIOGRAPHIES

Nguyen Vinh Nam is Master of Computer Science, University of Natural Sciences, year 2005. Now, he is working at Vietnam Informatics and Mapping Corporation (vietbando).

His research interests are GIS, Spatial Data Mining.

Le Hoai Bac is Doctor of Computer Science, University of Natural Sciences, year 1999. Now he is working at Faculty of Information TechnologyUniversity of Science, VNU - HCMC, Vietnam

His research interests are Artificial Intelligence, Soft Computing, Data mining and Knowledge Discovery.