Proceedings of International Conference on
Engineering and Information Technology
Sep. 17-19, 2012, Toronto, Canada

New Geometrical Similarity-based Clustering Algorithm for GIS Vector Data

Prof. Dr. Tawfiq A. Abbass[a] , Dr. Mahdi JASIM [b]

[a] Department of Computer Networks
College of Computer Technologies
University of Babylon
Hillah -Iraq
tawfiqasadi63@yahoo.com

[b] Department of Computer Networks
College of Computer Technologies
University of Babylon Univ
Hillah –Iraq
mnjisd@yahoo.com

Abstract

Geographic Information System(GIS) are usually classified into raster, vector, and raster –vector systems. The research deals with proposing new graph mining algorithm called GIS-GMA. The algorithm is used for clustering the vector features of GIS. The vector data are usually stored in data files called shape files. These files contains the (point, lines, polygons,...,etc). The extracted data is then stored in a dataset to be processed by the proposed algorithm to discover the full and partial similarities among map objects to assist the clustering and analysis of map data. It deals with clustering the polylines and polygonal data. The research results lead to build GIS prototype with spatial data mining facilities to cluster GIS vector data and giving fine clustering results,it is implemented using MicroSoft VS-2005 and ESRI ArcObjects.

Keywords :
Graph clustering algorithms, Graph mining, GIS data analysis, Mining GIS data ,Spatial data mining.

1. Introduction

During the study of related researches to Spatial clustering which deals with spatial data that is generally organized in the form of a set of *points* or *polygons*[6] we have found that mining of GIS spatial data is fertile for more research to be more convenient to GIS applications especially the vector data that deals with GIS graph data like points, lines, polygons,..., etc .The huge amount of GIS-based applications make it essential to improve graph mining techniques to extract the spatial knowledge embedded in that data to support GIS applications to be installed in new data analysis fields to support geographers and GIS users to investigate the similarity among the spaghetti of map objects, and therefore can understand and analyse the spatial relations like clustering them or classifying them.

* Corresponding author. Tel.: +1 905 670 9070 ext. 4431
Fax: +1 905 670 9095; E-mail: iceit@university.edu.ca

The proposed algorithm is intended to extract spatial knowledge based on the geometrical similarity among map objects to cluster them accordingly .The research is consisting from the following main steps. Extracting the graph features included in GIS-Shape files, design and implement dataset to store the extracted data, propose a new algorithm called **GIS-GMA** to discover the fully and partially similar vector map objects, cluster the resultants similar graph features, re-visualize the discovered object on the real map giving them a distancing effects.

.

2. Ralated works

Recently the GIS graph mining techniques attracts more researching efforts in both raster and vector data mining, but we are interested in the mining of graph data which concerned with knowledge discovery in GIS-vector data, the extracted data are intended to enhance the

GIS-query system to be more flexibly responding to user queries and gives analysis power, because the current queries lack the analysis power and if we need to analyse the GIS data we should pass them to another analysis system using spatial data mining techniques.
In 2009 Paul Van Dooren Catherine Fraikin [3] studied the similarity matrices for coloured graphs but only between nodes of two graphs, to the case of coloured graphs, where the colouring is either on the nodes or on the edges of both graphs. The proposed method tries to find the optimal matching between the nodes or edges of both graphs but only performs the comparison when their colours are the same.

In 2010 Akshara Pande et. al. designed an information system to work with data referenced by spatial geographical coordinates. They are detecting design patterns so that it can be used as a conceptual tool to cope with recurrent problems appearing in the GIS domain contiguity and compactness which are spatial features in GIS data [2].

In 2010 a study of clustering algorithms of area geographical entities based on geometric shape similarity was made and the researchers presented a similarity criterion of line segments shape and a criterion of area geographical entities comprehensively utilizing distance and geometric shape similarity. Clustering algorithms based on these criterions are more suitable for clustering analysis of area geographical entities. The extracted knowledge can be used to explore deeper level knowledge combined with other data mining methods and to improve the efficiency and quality of data mining [3]. We studied clustering algorithms of area geographical entities based on geometrical similarity. They also presented a similarity criterion of line segments shape and a criterion of area entities similarity comprehensive utilizing distance and geometry similarity. Clustering algorithms based on these criterions are more suitable for clustering analysis of area geographical entities [4] .

In 2011 Deepti and Jushi [5] used districting which is the process of dividing a geographic space or region of *spatial units often represented as polygons into smaller sub-regions or districts. As such, it can be viewed as a set partitioning problem, i.e. the problem is to cluster the entire set of spatial polygons into groups according to the spatial properties involved, redistricting is like to spatial clustering which can be done using spatially-flavoured constraints such as spatial*

3. Graph Mining Algorithm

Graph mining algorithms deals with discovering the knowledge embedded in GIS vector data which are mostly stored in GIS shape files. These GIS data structures are stored as one geographic feature per shape file like point feature layer, multipoint feature layer, and polygon feature layer [6] [8].

We used the ESRI shape file technical report to implement a *software module* to extract the *polyline* and *polygon* features of ESRI-Based maps and designed a dataset to contain the extracted data [1] .Fig. 1 shows the structure of ESRI shape file, it consists of three files for each feature layer (*.shp,*.shx, and *.dbf) the first one containing the geographical coordinates ,the second is an index between each feature and it's description in the database file, and the last containing a description on each polygon feature or the attributes of the spatial features[7].

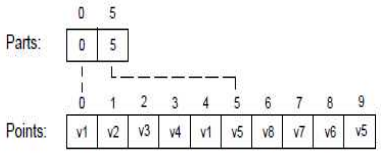

Table 1
Polygon Record Contents

Position	Field	Value	Type	Number	Byte Order
Byte 0	Shape Type	5	Integer	1	Little
Byte 4	Box	Box	Double	4	Little
Byte 36	NumParts	NumParts	Integer	1	Little
Byte 40	NumPoints	NumPoints	Integer	1	Little
Byte 44	Parts	Parts	Integer	NumParts	Little
Byte X	Points	Points	Point	NumPoints	Little

Note: X = 44 + 4 * NumParts

Fig. 1. The shape file structure for Polygon feature.

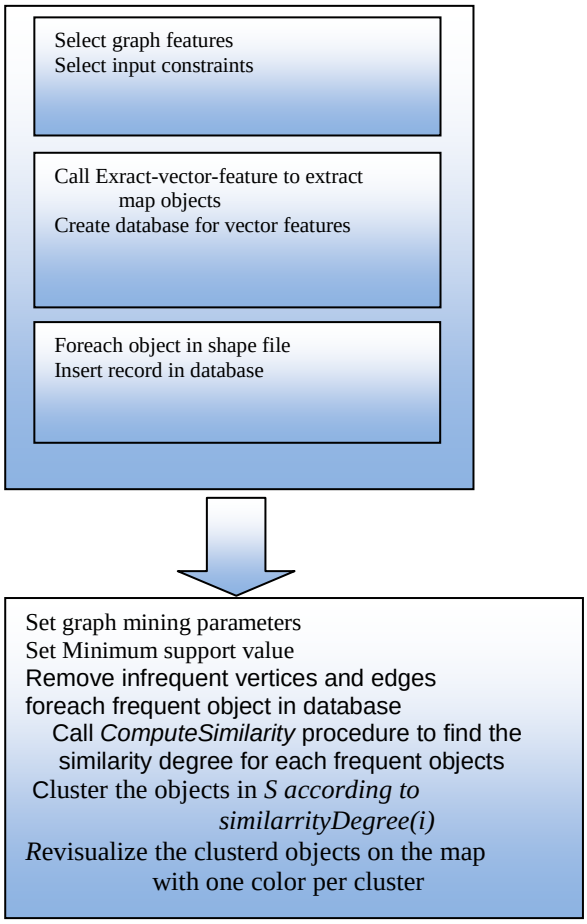

Fig. *2*. The block diagram of the main steps of the Proposed algorithm GIS-GMA.

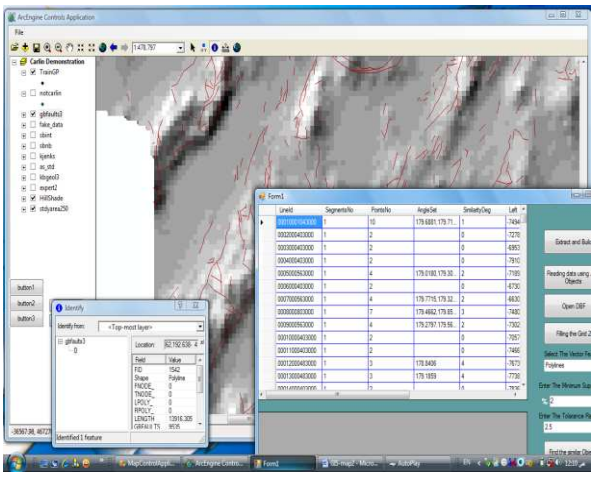

Fig. 3 The dataset extracted from the shape file

Fig. 3 shows part of the dataset which is built to hold the spatial-vector data that extracted from the GIS shape file by our module *Exract-vector-feature*. This data will be passed to the data mining algorithm **GIS-GMA** to extract the geometrical similarities between polygon objects and also between polylines map-objects.
The main algorithm **GIS-GMA** calls the subroutine *ComputeSimilarity* to compute the geometrical similarity of each map object responding to the main algorithm constraints like minimum support value or the number of edges for each object.The geometrical similarity is computed as shown in fig. 3.

each object in data set S will be used to find which objects in dataset S (O_j) is partially or fully geometrically similar to the current object (O_i), accordingly similar objects will have the same similarity degree and this degree will be used as clustering parameter .

engles-between-edges to find the angle between them. The number of edges and the set of angles related to

V1 = Normal(P2-P1) (1)
V2 = Normal(P2-P3 (2)
Angle = Acos(V1 dot V2) * 180/pi (3)

The angle between any three edges is computed according to Eq.1 to Eq.3 The coordinates of the cross section points are passed as parameters to the algorithm

```
Algorithm ComputeSimilarityDegree  (Of, NumberofEdges, NumberOfNodes, AnglesTolerance,,
                                                    SimilarityDegree)   // Compute similarity
Foreach Object Oi. NumberOfEdges>= Of.NumberOfEdges
     Read Oj . Fields()                          // read all fields related to the Object Oi
     Read Oi .CurrentFeature Data
     Read AnglesBetweenEdges(CurrentPolyline Oi)  // read the angles Of.
     TolarenceBuffer = Read Oj.AngleSet  ±  AnglesTolarencValue
      If  (Oi .CurrentFeature Data within the TolarenceBuffer) and
            ( And Oi .NemberofEedges = Of .InputNumber of Edges)
          Set Oi.SimilarityDeg= SimilarityDeg
        End If
     End for j
     SimilarityDeg = SinilarityDeg+1
End for i
End Algorithm ComputeSimilarityDegree
```

Fig. 3 The outline of Compute-Similarity algorithm

Fig. 4 shows part of the output of our Graph-mining algorithm **GIS-GMA** which is applied on ESRI map layer for polyline feature.

When we applied the proposed algorithm on the map depicted in Fig.3 we have found that it consists of 1674 objects and they are clustered to 74 clusters after excluding the objects that are not respond to the pre-processing constraints map objects then we can omit redundant objects using some sort of indexing for the similar objects.

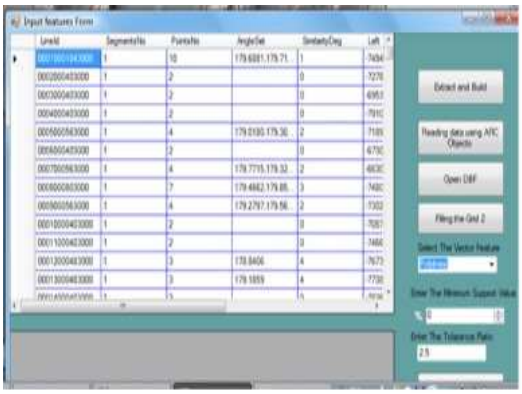

Fig. 4. The extracted data for polyline layer

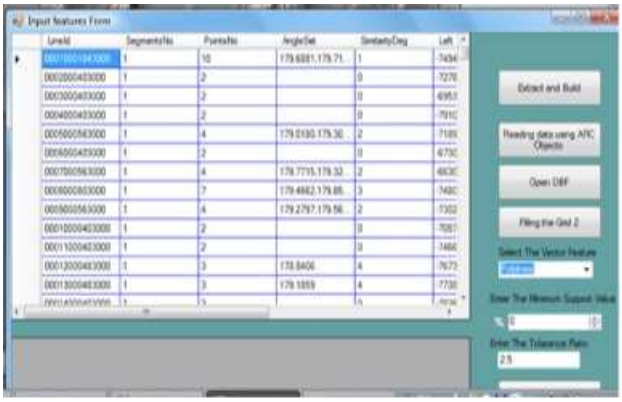

Fig. 5. The dataset with similarity degrees

3. Conclusion

In this paper we present a new algorithm for mining the GIS vector data (polylines, polygons, and ,etc).The algorithm clusters map objects based on the degree of geometrical similarity among map objects. The results of applying this algorithm on real map data gives good clustering results with less computing costs because of the simple geometrical similarity measure.

The algorithm may need some enhancement to deal with mining all vector types of map objects. The algorithm is also capable to deal with vector data stored in shape files and not related to GIS system like auto CAD and other drowning tools to cluster their map objects as a type of knowledge discovery to help designers analysing their maps.

ACKNOWLEDGMENTS

We would express our thanks to all staff members and employees of departments of computer networks for their kind cooperation to fulfill the research and their encouragement during the progress of the research.

References

[1] ESRI Shapefile Technical Description An ESRI White Paper—July 1998Copyright © 1997, 1998 Environmental Systems Research Institute, Inc.

[2] Akshara Pande Manjari Gupta A.K. Tripathi, Design Pattern Mining for GIS Application Using Graph Matching Techniques 978-1-4244-5540-9/10,2010 IEEE.

[3] C. Fraikin and P.Van Dooren, *Graph matching with type constraints*, ECC 07, Greece, July 2-5, 2007.

[4] Chen Guang-xue, LI Xiao-zhou, Chen Qi-feng ,LI Xiao-zhou,*Clustering Algorithms for Area Geographical Entities in Spatial Data Mining*, 978-1-4244-5934-6/10/2010 IEEE, (FSKD 2010).

[5] Deepti Joshi, Leen-Kiat Soh *Member, IEEE*, and Ashok Samal, *Member, IEEE*, Redistricting using Constrained Polygonal Clustering. Digital Object Indentifier 10.1109/ TKDE.2011.140.

[6] Ilaria Bordino, Debora Donato, and Aristides Gionis Mining large networks with subgraph counting, 2008, Eighth IEEE International Conference on Data Mining

[7] Kang-Tsung Chang, *Programming ArcObjectswith VBA Second Edition,*Chapman & Hall/CRC Taylor & Francis Group 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742 © 2008 by Taylor & Francis Group, LLc.

[8] Sayan Ranu and Ambuj K. Singh. Graphsig: A scalable approach to mining significant subgraphs in large graph databases.In *ICDE*, 2009.