A RESEARCH ON SPATIAL TOPOLOGICAL ASSOCIATION RULES MINING

Jiangping Chen [1], Song Liu [1], Penglin Zhang[1], Zongyao Sha[1]

(1 School of Remote Sensing and Information Engineering ,Wuhan University ,

129 Luoyu Road ,Wuhan ,China ,430079)

Commission II, WG II/3

KEY WORDS: Geography, Research, Data mining, GIS, Algorithms

ABSTRACT:

Spatial association rules mining is a process of acquiring information and knowledge from large databases. Due to the nature of geographic space and the complexity of spatial objects and relations, the classical association rule mining methods are not suitable for the spatial association rule mining. Classical association rule mining treats all input data as independent, while spatial association rules often show high autocorrelation among nearby objects. The contiguous, adjacent and neighboring relations between spatial objects are important topological relations.

In this paper a new approach based on topological predictions to discover spatial association rules is presented. First, we develop a fast method to get the topological relationship of spatial data with its algebraic structure. Then the interested spatial objects are selected. To find the interested spatial objects, topological relations combining with distance were used. In this step, the frequent topological predications are gained. Next, the attribute datasets of the selected interested spatial objects are mined with Apriori algorithm. Last, get the spatial topological association rules. The presented approach has been implemented and tested by the data of GDP per capita, railroads and roads in China in the year of 2005 at county level. The results of the experiments show that the approach is effective and valid.

0. INTRODUCTION

Spatial association rules mining is a technique which mines the association rules in spatial databases by considering spatial properties and predicates[1,2,3]. One important problem is how to get those spatial relations that compose the spatial properties and predicates from the spatial objects, and translate the non-structured spatial relations to structural expression so that they can be mined with the non-spatial data together. There are many researches about spatial association rules mining, Algorithm ARM involving the spatial relations such as direction, distance and topology has been well discussed in some literatures[2,4 - 7], whereas Fenzhen Su[8] centered on using spatial difference to express how spatial relations affect the interested spatial association rules we can get.

As spatial topological relation is one of the most important spatial relations, many literatures about spatial association rules mining fasten on it and presented many ways to get the topological relation, such as RCC (Region Connection Calculus), the classical MBR (minimum bounding rectangle), 9-intersection model. Ickjai Lee etc. Compared different RCC models' efficiency when used to get topological relations of a group of objects. MBR is a fast method to get the rough topological relations, so it is often used with other precise but expensive means. Eliseo Clementini etc.[3] put forward a method mining the spatial objects with uncertainty which uses the objects with a broad boundary to take the uncertainty of spatial information into account. The topological relations between objects with broad boundaries is described by 9-intersection which can be concisely represented by a special 3×3 matrix and can distinguish 56 kind of topological relations between objects with broad boundaries. All those methods have a common problem when extracting topological relations, they are very time consuming.

Paralleled to the concept of algebraic structure in axiomatic system of set theory, topological structure has continuously been the organization of spatial data in GIS, such as the typical data format of ArcInfo--Coverage and the new data model after ArcInfo 8--Geodatabase. Both algebraic structure and topological structure can express topological relations. For the reason that algebraic structure is the combination of adjacent unit variables, it would not cause

topological connected effect however the spatial data changes. Nevertheless, topological structure possesses multilayer dependence, leading to the reconstruction of topological structure when spatial data changes, which is disastrous to the maintenance of magnanimity data.

In this paper, a new approach based on topological predications gaining by algebraic structure of spatial graph to discover spatial association rules is proposed.

1. MATERIALS AND STUDY AREA

This paper deals with China region, which has been mapped at 1:4,000,000 scale. All the data are downloaded from Data Sharing Infrastructure of Earth System Science (http://www.geodata.cn/Portal/?isCookieChecked=true) of China.

As an indicator of a country's standard of living, GDP per capita is regarded as a kind of attribute data and represented by polygons in the level of county(figure 1), summing up to 3406 polygons.

Figure 1. Counties with GDP per capita attribute in China

Railroads and roads are the most significant means of ground transportation. This paper selects arterial railroads and roads with massive freight and passenger transportation in China(figure 2).

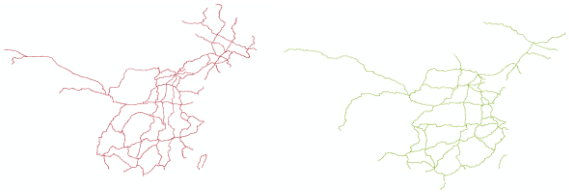

Figure 2. Railroad and road data in China

2. METHODS

2.1 Definitions

Definition 1 (Unit) According to theory of cell structure proposed by J.Corbett[9], spatial feature should be divided into point, line, surface and volume. For instance, roads are directed lines and buildings are polygons. For the reason that polygon is made up of line, and point can be regarded as degraded line, line unit algebraic structure is the most basic as well as significant content in this paper.

Definition 2 (Permutation) Suppose domain X represents index set (+1, -1, +2, -2, $\cdots$, +n, -n), and unit variable $\mathbf{x} \in \mathbf{X}$ is given, then expression y=a(x) is the permutation of x under the precondition of $\mathbf{y} \in \mathbf{X}$ [10].

Definition 3 (Involution) Involution can be seen as a particular case of permutation, if y=a(x)=a[a(x)]=x came into existence, this particular permutation would be called involution.

Definition 4 (Zero-order permutation $\mathbf{a}_0$ (x)) Zero-order permutation is defined as the transformation from unit variable +x to -x or from -x to +x based on the concept of involution. Zero-order permutation gives a line with direction +x to -x.

Definition 5 (First-order permutation $\mathbf{a}_1$ (x)) First-order permutation records all the adjacent line units in counterclockwise order. First-order permutation gives the expression of point with all its crossing lines.

Definition 6 (Algebraic Structure) Based on the concepts and idea above, the expression {X , $\mathbf{a}_0$ (x), $\mathbf{a}_1$ (x)} is called the algebraic structure of spatial graphic.

2.2 Construction of Topological Relation from Algebraic Structure

Taking line graphic in figure 3 as an example, the process of producing line unit algebraic structure expression can be divided into two steps. The first step is the unitization of spatial graphic, at the same time, the unitized result needs numbering in the same order with unitization to constitute index of unit variables X. The fundamental principles of unitization are

①every terminal point and cross point should be treated as point unit;

②there must not exist any point unit in the middle of any line unit;

③line units are not allowed to cross with each other.

Then the second step is producing zero-order permutation $\mathbf{a}_0(x)$ and first-order permutation $\mathbf{a}_1(x)$ with $\mathbf{x} \in \mathbf{X}$, as table 1 shows. In order to save memory space, $\mathbf{a}_1(x)$ is recorded as the nearest adjacent line unit in the counterclockwise rotation searching, instead of all the adjacent line units. For instance, line unit 1 rotates in counterclockwise order with P1 as pivot, and gets $\mathbf{a}_1(1)$ =5, while line unit -1 rotates in counterclockwise order with P2 as pivot and gets $\mathbf{a}_1(-1)$ =3.

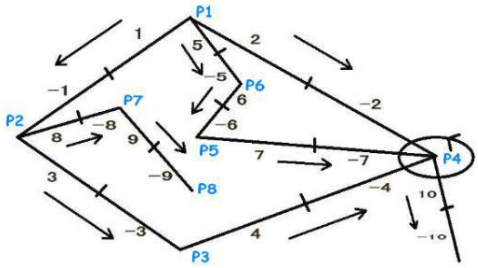

Figure 3. Sample line graphic

x	1	-1	2	-2
α_0	-1	1	-2	2
α_1	5	3	1	-7

Table 1. Part of line unit algebraic structure of Figure 3

From the line unit algebraic structure could we get topological relation easily by different combination of zero-order permutation $\mathbf{a}_0(x)$ and first-order permutation $\mathbf{a}_1(x)$. Let us take intersecting relation for instance, since we have combined polygon layer with line layer into one line layer, this intersecting relation of polygon and line can be treated as adjacency relation of lines with two kind of attribute. Suppose $L(x) = \{x, \mathbf{a}_0(x)\} = \{x, y\}$ is the given line, we are able to elicit the adjacent lines R(L) by

①$N(x) = \mathbf{a}_1(x) = (\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_n)$;

②$N(y) = \mathbf{a}_1(y) = \{\mathbf{y}_0, \mathbf{y}_1, \cdots, \mathbf{y}_n\}$;

③$R(L) = \{(\mathbf{x}_0, \mathbf{a}_0(\mathbf{x}_0)), (\mathbf{x}_1, \mathbf{a}_0(\mathbf{x}_1)), \cdots,$

$(\mathbf{x}_n, \mathbf{a}_0(\mathbf{x}_n)), (\mathbf{y}_0, \mathbf{a}_0(\mathbf{y}_0)),$

$(\mathbf{y}_1, \mathbf{a}_0(\mathbf{y}_1)), \cdots, (\mathbf{y}_n, \mathbf{a}_0(\mathbf{y}_n))\}$[10].

Since the topological relation is represented as functional relation of units, the topological operation upon given unit will only affect adjacent units. Therefore, the method of representing topological relation by line unit algebraic structure can effectively eliminate the phenomenon of data linkage and topological redundancy, bringing the improvement of spatial analysis efficiency.

2.3 Association Rules Mining

The topological relation matrix in part 2.2 is regarded as waiting mining transaction database, and needs inspection. Representing the topological relation by line unit algebraic structure and qualitatively storing the result render this spatial relation can be processed in the same mining way with attribute data. This step is to complete association rules mining with typical Apriori algorithm, based on the given minimum support and minimum confidence. Apriori algorithm is a primary algorithm of mining association rules of attribute, aiming at finding out relation among items of a data set.

3. EXPERIMENTS

In order to reduce the recorded points as well as display the graph more clearly, one step before construction of algebraic structure is to simplify the polygons by Point Remove algorithm, which is a fast, simple algorithm that reduces a polygon boundary quite effectively by removing redundant points. A relatively sketchy outline is precise enough and extraordinarily effective for algebraic analysis display. Figure 4

shows the visualization of simplified polygons of GDP per capita source data.

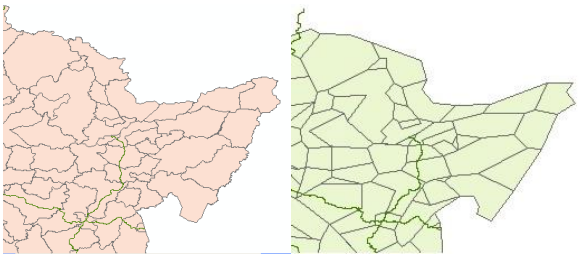

Figure 4. Simplify polygon

After the preprocessing such as filling out incomplete data and noise treatment, the next step is unitization and numbering. Based on the principles and multi-layer characteristic of GIS data, we overlay layers into one, for example, the railway layer and county layer is combined into one line layer in figure 5. Then we treat all the features as lines and split these lines wherever exits a point. For the reason that the source data has no direction, we unitize the line unit with random direction, which is splitting every line into '+' part and '-' part randomly. At the same time, uniquely number the line units and organize the index of unit variables X like (+1, -1, +2, -2, ••• , +n, -n).

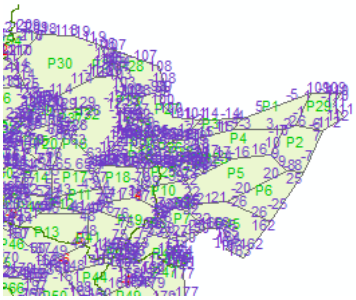

Figure 5. Unitization and numbering

From the numbering graphic and index of unit variables X could we produce zero-order permutation $\mathbf{a}_0$ (x) and first-order permutation $\mathbf{a}_1$ (x). For instance, the original railway layer is represented like figure 6, while GDP per capita layer seems more complex as figure 7 shows.

ID	a0	a1
4	-4	4
-4	4	-6
5	-5	-4
-5	5	-5
6	-6	-22
-6	6	8
7	-7	-2
-7	7	-8
8	-8	5
-8	8	-7

Figure 6. Line unit algebraic structure-railway

ID	LineID
P16	62, 63, 64, 65, 66
P17	-43, -55, -65, -67
P18	-44, -67, -64, -127, -141, 69, 70, -...

LineID	a0	a1
62	-62	-59
-62	62	63
63	-63	126
-63	63	64
64	-64	-127
-64	64	65
65	-65	67
-65	65	66
66	-66	-58
-66	66	62

Figure 7. Polygon unit algebraic structure-GDP per capita

By means of algebraic structure in figure 6 and 7, we can easily derive the basic topological relation matrix like figure 8 with acceptable operation time like figure 9.

	PolygonID 41	PolygonID 42	PolygonID 43	PolygonID 44	Poly
PolylineID 4	0	0	0	0	0
PolylineID 5	0	1	1	0	0
PolylineID 6	1	0	0	0	0
PolylineID 7	0	0	0	0	0
PolylineID 8	0	0	0	1	0

Figure 8. Topological relation--intersecting matrix

Figure 9. Operation time

In order to apply association rules mining algorithm, firstly we need to generalize the continuous attribute--GDP per capita data. For the specific experiment, generalizing the GDP per capita data into three levels based on its value, that is, the upper third of GDP per capita in the order from the most to the least is "high", the middle third is "medium" and the rest is "low".

Using generalized GDP per capita data as consequent of rules, and using intersecting relation between county and railroad as well as intersecting relation between county and road as antecedent of rules, we can apply typical Apriori algorithm to mine associate rules.

4. RESULTS AND DISCUSSION

Since the original GDP per capita data has nearly 3000 polygons, it is painful to deal with its topological operation in the ESRI ArcObject or ArcEngine, taking no account of speed optimization like gaining index. Nevertheless, taking intersecting matrix as an example, algebraic structure could construct intersecting matrix in approximately ten seconds, and deal with topological operation, for example, adding, removing and modification, in less than one second.

With minimum support 10, we could get the mining results, which are mostly omitted here due to page limitation. It can be seen that GDP per capita of counties with railway and highway simultaneously crossing is universally higher, and these railways and highways are usually main artey of communications, connecting diverse regions all around the country. Through setting the minimum confidence as 0.8 could we filter the frequent 3-itemsets and get the filtered result in table 2.

Bin-Sui Railway: intersectant	Sui-Man Highway: intersectant	GDP：high	Support: 15	Confidence: 0.83
Chang-Da Railway: intersectant	Shen-Hai Highway: intersectant	GDP：high	Support: 11	Confidence: 0.92
Jin-Pu Railway：intersectant	Jing-Hu Highway: intersectant	GDP：high	Support: 11	Confidence: 0.92
Jin-Pu Railway：intersectant	Shen-Hai Highway: intersectant	GDP：high	Support: 11	Confidence: 0.92
Hu-Ning Railway: intersectant	Hu-Shan Highway: intersectant	GDP：high	Support: 14	Confidence: 1

Table 2. Filtered result of frequent 3-itemsets

After further analyzing the filtered result, we find that these railways and highways are built relatively earlier, and they remarkably stimulated the economic development of perimeter zone and even the whole nation. Therefore, the mining result shows that railroads and roads could drive the economic growth of circumjacent counties and cities. This conclusion fits the fact and proves that the mining result is believable.

5. CONCLUSIONS

It can be concluded that the brand new method to organize spatial data based on algebraic structure can deal with topological relation effectively and efficiently, for either topological construction or maintenance. Furthermore, the association rules mining results based on it are believable.

However, there still are a few problems should be considered. Even though the process of constructing topological relation from algebraic structure costs little time, the efficiency of constructing of algebraic structure of mass data needs further studies. Moreover, the containing relation is not optimized in the algebraic structure.

6. ACKOWLEDGEMENTS

This study was supported by the National Science and Technology Pillar Program during the Twelfth Five-Year Plan Period (No. 2012BAJ15B04) and the National Nature Science Foundation of China (No. 40801152, No. 61172175 and No. 41071249).

7. REFERENCES

[1] Mennis J, Guo D. Spatial data mining and geographic knowledge discovery—An introduction. *Computers, Environment and Urban Systems*, 2009, 33(6), pp.403-408.

[2] Fang G, Xiong J, Tu C. An algorithm of mining spatial topology association rules based on Apriori. *Intelligent Computing*, 2010, 2(011), pp.101-104.

[3] Clementini E, Di Felice P, Koperski K. Mining multiple-level spatial association rules for objects with a broad boundary. *Data & Knowledge Engineering*, 2000, 34(3), pp.251-270.

[4] Gongquan L, Keyan X. Spatial Data-mining Technology Assisting in Petroleum Reservoir Modeling. *Procedia Environmental Sciences*, 2011, 11, pp.1334-1338.

[5] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 2007, 60(1), pp.208-221.

[6] Estivill-Castro V, Lee I. Clustering with obstacles for Geographical Data Mining. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2004, 59(1-2), pp.21-34.

[7] Lee I, Qu Y, Lee K. Mining qualitative patterns in spatial cluster analysis. *Expert Systems with Applications*, 2012, 39(2), pp.1753-1762.

[8] Su F, Zhou C, Lyne V, et al. A data-mining approach to determine the spatio-temporal relationship between environmental factors and fish distribution. *Ecological Modelling*, 2004, 174(4), pp.421-431.

[9] Corbett, 1985. *A General Topological model for Spatial reference*, U.S.Census, pp.3-15.

[10] Jingshegn Z, Changqing Z, 2005. *Algebraic representation and morphological transformation of spatial graphics*, Beijing, pp.16-18, 21, 131.