AN ALGORITHM ABOUT ASSOCIATION RULE MINING BASED ON SPATIAL AUTOCORRELATION

Jiangping Chen[a,b]

[a]School of remote sensing Information Engineering Wuhan University，Wuhan Hubei, PR China 430079

chenjp_lisa@163.com, ch_lisa@hotmail.com

[b]Department of Geography, University of Cambridge,Downing Place, Cambridge, UK, CB2 3EN

jc564@cam.ac.uk

KEY WORDS: Geographical information science, Statistical analysis, Data mining, Geography, Spatial association rule, spatial auto-correlation

ABSTRACT:

Most spatial data in GIS are not independent, they have high autocorrelation. For example, temperatures of nearby locations are often related. Most of the spatial association rule mining algorithms derived from the attribute association rule mining algorithms which assume that spatial data is independent. In these situations, the rules or knowledge derived from spatial mining will be wrong. It is, therefore, important that mining spatial association rules take into consideration spatial autocorrelation. At present, spatial statistics are the most common method to research spatial autocorrelation. In spatial statistics, classic statistics are extended by taking into account spatial autocorrelation. Spatial Autoregressive Model, SAR, is one of the methods; the adjacency matrix is used to describe the interaction of neighbouring fields which can simulate the effect of dependence between variables. The disadvantage of spatial statistics is that the calculation consuming is high so it cannot be widely applied in spatial data mining.

This paper puts forward a new method of mining spatial association rules based on taking account of the spatial autocorrelation with an cell structure theory. It defines spatial data with an algebra data structure then the autocorrelation of the spatial data can be calculated in algebra. According to J. Corbett's cell structure theory (1985), spatial graph is a subset of point, line, face, and body. The algebra structure of point, line, face and body can be used to express spatial data. In spatial data mining, we mine rules in the spatial database.

In this paper the first step is to design a structure about point, line, face and body to express the spatial data and then store it in the spatial database. The second step is to build the measurement model of spatial autocorrelation based on the algebra structure of spatial data. The third step to mine the association rules based on the spatial autocorrelation model. Taking account of spatial autocorrelation is a significance research field for mining spatial association rules. We can get the spatial frequency items from the autocorrelation of the spatial data. This replaces the repeated scanning of the database by the measure of the spatial autocorrelation.

1. INTRODUCTION

Spatial data mining, i.e, mining knowledge from large amounts of spatial data, is a demanding field since huge amounts of spatial data have been collected in various applications, ranging from remote sensing to geographical information systems (GIS), computer cartography, environmental assessment and planning. The collected data far exceeds people's ability to analyze it. Thus, new and efficient methods are needed to discover knowledge from large spatial databases. Attribute data mining methods were extended to applying in spatial data mining.One of the big

problem for these kind of the spatial data mining methods is that they do not take into account the correlation of spatial information.

In the light of the first law of geography "everything is related to everything else but nearby things are more related than distant things"[Tobler 1979] that values from samples taken near each other tend to be more similar than those taken farther apart. This tendency is termed spatial autocorrelation or spatial dependence (Cliff and Ord 1973; Rossi et al. 1992; Liebhold et al.1993). It's natural that most spatial data in GIS are not independent, they have high autocorrelation. For example, temperatures of nearby locations are often related. Most of the spatial association rule mining algorithms derived from the attribute association rule mining algorithms which assume that spatial data is independent. In these situations, the rules or knowledge derived from spatial mining will be wrong. It is, therefore, important that mining spatial association rules take into consideration of spatial autocorrelation. Spatial autocorrelation is when the value at any one point in space is dependent on values at the surrounding points. It is problematic for classical statistical tests, such as ANOVA and ordinary least squares (OLS) regression, that assume independently distributed errors (Haining 1990; Legendre 1993). When the response is autocorrelated, the assumption of independence is often invalid, and the effects of covariates (e.g., environmental variables) that are themselves autocorrelated tend to be exaggerated (Gumpertz et al. 1997).At present, geostatistics and spatial statistics are the most common method to research spatial autocorrelation. Geostatistics is a relatively new discipline developed in the 1960s primarily by mining engineers who were facing the problem of evaluating recoverable reserves in mining deposits. It provides a set of statistics tools, such as kriging[Cressie 1993]to the interpolation of attributes at unsampled locations.In spatial statistics,spatial autocorelation is quantified using measures such as Ripley's K-function and Moran's I [Cressie1993].The spatial relationship among locations in a spatial framework is often modeled via a contiguity matrix.

In data mining, one of the most classic and novel algorithm for association rule is Apriori(Agrawal,1993,1994),which has been extended to a lot of algorithms.such as Agrawal's CD(countdistribution), CaD(candidatedistribution), DD(data distribution), and Park's PDM algorithm, and Chueng's DMA algorithm and FDM algorithm. Although these algorithms are based on distribute databases, they are suitable for parallel mining. In addition, though DMA algorithm get rid of some disadvantages of CD algorithm, it still requires more frequent synchronization among computers. FDM algorithm is in accordance with DMA algorithm basically, but the difference is that FDM algorithm increases the technology of globe pruning. With the development of the spatial data mining, other methods have been used for spatial association mining such as: spatial statistical analysis geostatistics and spatial clustering (M. Ester,1996,E.knorr,1996) However, most of these approaches focus on discovering the spatial relationships among neighboring data sets.

In our works ,we developed and implemented an spatial data structure for an efficient determination of such spatial autocorrelation. Based on the spatial structure we present a method for mining spatial association rules for objects with autocoreelation.

2. PROBLEM DEFINITION

Mining spatial association rules can be defined as below:

Input:

a spatial database (SDB) including geography graph and attribute tables,

two series of thresholds for every large k-itemset in the spatial database, minsup[l] and minconf[l] for large 1-itemset and minsup[k],minconf[k] for large k-itemset.

Output:Some strong spatial association rules.

2.1 Relatied definition

Definition 1. A spatial association rule is a rule in the form P1∩ …∩Pm=>Q1 ∩ …∩Qn (s%; c%_)

where at least one of the predicates P1,.Pm, Q1,.Qn is a spatial predicate, s% is the support of the rule, and c% is theconfidence of the rule.

Definition 2. The support of a conjunction of predicates,

P = P1∩ …∩Pm in a set S, denoted as ρ(P/S), is the numberof objects in S which satisfy P versus the total number ofobjects of S.

The confidence of a rule P-→Q in S, Ψ(P-→Q/S), is the possibility that Q is satisfied by a member of S when P is satisfied

by the same member of S. A single predicate is called a 1-predicate. A conjunction of k single predicates is called a k-predicate.In this paper,the large itemset contains k predicates is called k-itemset, and the set of all the large k-itemset is L_k .

3. OUR PROPOSED ALGORITHM

3.1 Data structure

The cell structure is a combine set of point,line,area and body. And the cell can be a point,line,area and body while the line is consist of point and area is made up of line and body can be formed by areas.For example:

E0: Point = P(x,y);

E1:Line = F(0+1) E0

E2:Area = F(1+1) E1

E3:Body =F(2+1) E2

So :E(i+1)=F(i+1) E (i)

From the above, it can be seen that the subset of the cell is not independent it's related.

For example: the cell of area is consisted of cell line while the cell line is the retrogression of cell area.So if cell line is defined the definiton of cell area is redundance.

Let us suppose: If there is a cell structure B = (E,R), then must be two unrelated function (Jingsheng zhai, 2005)

B= (E ={ α i(x): i = 1,2,3,….,n/x ε D})

R= ({ β i(x): i = 1,2,3,….,n/x ε D})

Definiton 3 if there is a cell variant x ε D, a0 is the pair of x to form a line and a1 is a change function of x, then

M = {D,a0, a1}is the algebric structure of 2-dimension graph.

After we have the cell structure of the graph, then we can express the spatial autocorrelation of the 2-d graph like this:

Since in the 2-d graph,cell line is the basic cell, the adjacent areas will have the same share edges.

T: is the set of share edge of two areas, ti ε T

k: the number of the times ti emerged in the algebraic structure.

T= {ti ε D, ki ≥2,i= 1,2,…,n}

Then the autocorreltion of the cell line can be defined like this : I

$$= ki / \left(\sum_{i=1}^{n} k_i \right) * 100\%$$

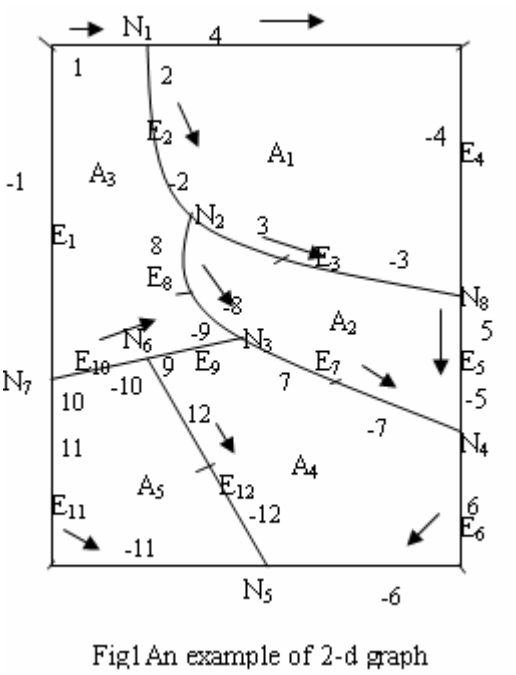

Fig1 An example of 2-d graph

Tab 1 Direction Edge

Area code	Direction Edge
A_1	E_2 , E_3 ,E_4
A_2	E_3 , E_5 ,E_7 ,E_8
A_3	E_1 , E_2 ,E_9 ,E_8 ,E_{10}
A_4	E_6 , E_9 ,E_7 ,E_{12}
A_5	E_{10} , E_{11} ,E_{12}

Tab2 Node

Area code	Direction Edge
N_1	E_2 , E_1 ,E_4
N_2	E_3 , E_2, E_8
N_3	E_7 , E_9 ,E_8
N_4	E_6 , E_5 ,E_7
N_5	E_6, E_{11} ,E_{12}
N_6	E_9, E_{10} ,E_{12}
N_7	E_1, E_{11} ,E_{12}
N_8	E_3, E_4 ,E_5

Tab3 the algebraic structure of the fig1 for cell line

x	1	-1	2	-2	3	-3	4	-4	5	-5	6	-6	7	-7	8	-8	9	-9	10	-10	11	-11	12	-12
$A_0(x)$	-1	1	-2	2	-3	3	-4	4	-5	5	-6	6	-7	7	-8	8	-9	9	-10	10	-11	11	-12	12
$A_1(x)$	4	10	3	8	2	5	2	-3	-3	-7	7	-12	-9	6	3	7	-10	-8	1	12	-1	-12	9	-11

In the tab3 ,x is the cell variant.From the definition3 of the cell structure and the relative of cell variant, we know that cell line 2,3,7,8,9,10,12 is related Cell of area is the consist of cell line and cell body is made of cell area. A cell of line can be adjacent with 2 cell of areas and 4 cell of bodies while a cell area must be related with 2 cell of lines and a cell body must be related with 4 cell of lines.

The cell structure is the mathematic abstract for spatial graph not data structure.From the above it is possible to implement the recessive representation of data structure with the function changing of cell structure.

3.2 The thoughts of the algorithm

The difference between the spatial data mining and attribute data mining is that spatial data is not independent. There are association between the spatial data. And the topology is the reason for spatial autocorrelation.In our work, the cell structure has represent the topology and so we can depict the spatial autocorrelation with the cell structure.

3.3. The mining algorithm

Our proposed algorithm called AR-Miner, shown in below, consists of two phases. In the first phase, we join the algebraic structure table of the graph and their attribute table. In the second phase, we mining the association rules as below:

1 Scanning the database and computing the probability for every attribute

2 Set a threshold probability for ever Large 1-itemset, 2-itemset and n-itemset.

3 For the itemset whose probabilty is smaller than the threshold then set it's probabilty =0;

4 The probability of L(k+1)-itemset = the times of the probability of Lk-itemset;

5 repeat the step 3-4 and then output all the large itemsets whose prbability is larger than the threshold.

Algorithm: AR-Miner

Input: the spatial map and it's attribute table, the minimum support threshold *min_sup[l],l=1,2,…n*.

Output: the set including all frequent itemsets *L*.

1 **begin**

2 Reading the spatial map and transform it into it's algebraic structure table T.

3 Scan the graph algebraic structure table T;

```
 {
      For( i=1, i <=n, i ++)
      {
       If number of cell line N[CL(i)] >=2; Then
      P[ CL(i)] =(±) N[CL(i)]/Total of number of cell line * 100%;
```

// computing the autocorrelation of cell lines, positive or negative correlation is up to the direction of the cell line.

}

4 Join the table T and the attribute table and Scan it

$PFA_1[L_{(j)}]$ = Number of appear times of every attribute / the total number of reords *100%;

If $PFA_1[L_{(j)}] <= min\_sup[1]$ then

$PFA_1[L_{(j)}] = 0$;

For (j= 1; j<= n; j ++)

{

For(k=1; k<= n; k ++)

{

$$PFA_{k+1}[L_{(j)}] = \prod_{k=1}^{k} PFA_{k+1}[L_{(j)}] + P[\, CL_{(i)}]$$

If $PFA_k[L_{(j)}] <= min\_sup[k]$ then

$PFA_k[L_{(j)}] = 0$;

$L_{(j)} \dashrightarrow L$

}

}

Output L;

End

4. EXPERIMENTS AND CONCLUSION

Part of the power pole lines graph and it's attribute table were sythetic as our experiment data. In this example, it sets minsup s=70%, minconf c=50%. A set of spatial association rule is obtained. Through the mining, the sets of spatial association rule, which we gain in the spatial layer table4 of the distribution of cement works nearby the pole tower, are rule 3 ,4,5,6. In the talbe 6 of the highway information we gain the spatial association rule 1,2. which show the relation of tower polluting and tower conking.

Table 4 the attribute of pole about cement works

Pollution grade	Distance with the cement works	Direction with the cement works	Wind direction	Defect rate
High	Shorter	Down-southeast	Southeaster	High
High	Shorter	Down-east	Southern	High
High	Shorter	Down-east	Southern	High
High	Shorter	Down-south	Southeaster	High
Low	Shorter	Down-east	Northeaster	Low
Low	Shorter	Up-southeast	Southeaster	Low
High	Shorter	Down-south	Northeaster	High
Low	Shorter	Up-southwest	Northern	Low
High	Shorter	Down-east	Northern	High

Table 5 the attribute of pole about highway flow

Pollution grade	Distance with the highway	Highway flow	Wind direction	Defect rate
High	Shorter	Big	Southeaster	High
High	Shorter	Big	Southern	High
High	Shorter	Big	Southern	High
High	Shorter	Big	Southeaster	High
Low	Shorter	Middle	Northeaster	Low
Low	Shorter	Small	Southeaster	Low
High	Shorter	Big	Northeaster	High
Low	Shorter	Small	Northern	Low

High	Shorter	Big	Northern	High

Rule 1: nearby the cement works ∧ nearby the highway ∧ flow in highway is big → defect rate is high

Rule 2: nearby the cement works ∧ nearby the highway ∧ flow in highway is big → pollution grade is high

Rule 3: nearby the cement works ∧ at downwind direction ∧ nearby the highway → defect rate is high

Rule 4: nearby the cement works ∧at downwind direction ∧ nearby the highway → pollution grade is high

Rule 5: nearby the cement works ∧ at upwind direction ∧ flow in highway is small → defect rate is low

Rule 6: nearby the cement works ∧at upwind direction ∧flow in highway is small → pollution grade is low

From the mined spatial association rules, it is apparent that the defect rate of the pole tower is concerned with its pollution grade. Consequently, the users can conclude that the high pollution grade of the pole tower results in its high defect rate, then they can make analysis to find more direct reasons according with the former conclusion. Compare with the data which was mined by the algorithm reference [20], we can find two more rules about the low defect rate and pollution grade were appear. I think this is reason we have take the autocorrelation of line.Therefore, the knowledge extracted from the mining is useful and understandable.

To examine the above algorithm, since the cell structure is based on line ,we choose 4 power transmission lines.in order to compare the mining efficiency and the accuracy of rules between this algorithm and the other one that is refered in the reference [20]. The two algorithms were carried out on a laptop with an Intel Pentium IV 1GHz CPU and512M main memory, running Microsoft WindowsXP. All programs were coded in Microsoft Visual C++ 6.0. We test the two algorithms by synthetic data.

The experimental dataset consists of two kinds of data , a pwer transmission lines graph and then transformed into a talbe; The two attribute table of the power pole which respectively contain five attributes. An attribute has 2~10 attribute values after generalization. During the mining of the single layer spatial association rule, these two tables are connected into a dataset of six attributes. The minimum support of each data layer is 6%, and the minimum confidence is 75%.

The time complexity varied with the increasing of the number of records. When the number of itemsets gradually increased from 1000 to 5000, two curves of these two algorithms' time complexity are in Fig 2. With the expansion of scales of databases, although from the Fig2 we can conclude that the time increasing of the single layer mining algorithm used in one dataset is milder than the other, the work amount of its data preparation is so huge that the algorithm presented in this paper is still superior to it.

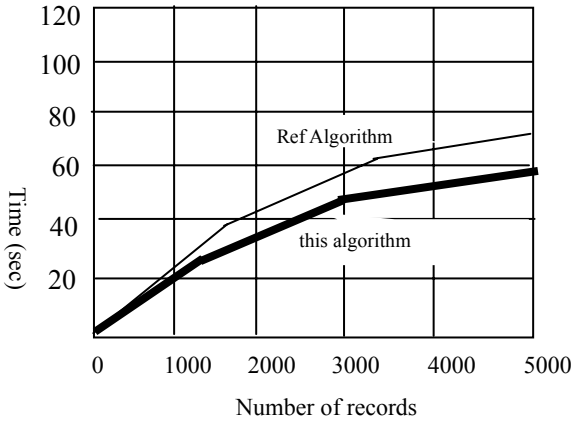

Fig1 Execution time with the increasing of itemsets

There are a number of possible future research directions in spatial association rule mining. First, it would be interesting to extend this work to mining association rules for an 3-d spatial database in which graphs are more complex. Second, our proposed method may generate a large number of patterns, and some of them may be wrong since the autocorrelation is not very acuracy. We are thinking about other representation of the autocorrelation in the algebraic spatial data structure. Third, how to find a fast and good algebra stucture of 3-d graph is also a problem, we could use our method to derive the algebraic cell structure of 3-d graph, Then it will take a long time to transform it and also the related edge will be numerous and some of the

related edge is redundancy. However, it does offer a new way to think about take into autocorrelation in an algebraic form and a new direction to computing it. From the experiment ,it proved that the result of AR_MINER is reasonable.

ACKNOWLEDGEMENTS

The author thanks Professor Haining and Rosangela for conversations that led to this research. And also thanks School of remote sensing Information Engineering Wuhan University to offer the author a chance to study in university of cambridge. This research was funded by the national 973 program(No.2006CB701305, 2003CB415205), the National Natural Science .

REFERENCES

[1] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, in: Proc. of ACM-SIGMOD International Conference on Management of Data, Washington, DC, 1993, pp. 207–216.

[2] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proc. of International Conference on Very Large DataBases, Santiago, Chile, 1994, pp. 487–499.

[3] Srikant R, Agrawal R. Mining generalized association rules. In:Dayal U , Gray P M D, Nishio S eds. Proceedings of the Inter-national Conference on V ery L arge Databases. San Francisco, CA:Morgan Kanfman Press, 1995. 406～ 419

[4] Han Jia-wei, Fu Yong-jian. Discovery of multiple_level association rules from large databases. In:Dayal U, Gray P M D,Nishio S eds. Proceedings of the Intelnational Conference on Very Large Databases. San Francisco, CA :Morgan Kanfmann Press, 1995. 420～ 431

[5] S. Agarwal, R. Agrawal, P. M. Deshpande, On the computation of multidimensional aggregates. VLDB'96, 1996,Bombay, India.

[6] Brin S., R. Motwani and C. Silverstein, Beyond market baskets: Generalizing association rules to correlations, in:Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD'97), J.M. Peckman, ed., ACM, Tucson, AZ, May 1997, pp. 265–276.

[7] Park JS,Chen M S,Yu P S. An effective hash- based algorithm for mining association rules. In:Proceedings of the 1995。

[8] Cheung D W,Han J,Ng Vetal. Maintenance of discovered association rules in large databases:an incremental updating technique. In:Proceedings of the 1996 International Conference on Data Engineering. New Orleans,Louisiana,1996

[9] James S. Ribeiro, Kenneth A. Kaufman, Larry Kerschberg. Knowledge discovery from multiple databases. In:KDD-95. 240～ 245

[10] Dao S. Perry B. Applying a data miner to heterogeneous schema integration, In:Proc. of the Int' l Conf. on Knowledge Discovery in Databases and Data Mmining(KDD-95), Montreal, Canada, August 1995. 63～ 68

[11] Chaudhuri, S. , Dayal, U. An overview of data warehousing and OLAP technology. ACM Sigmod Record, 1997, 26(1):65～ 74.

[12] O' Neil, P. , Quass, D. Improved query performance with variantindexes. ACM Sigmod Record, 1997, 26(2):38～49.

[13] Srivastava, D, Dar, S. , Jagadish, H. V. , etal. Answering queries with aggregation using views. In:Vijayaraman, T.M. , ed. Proceedings of the 22nd International Conference on Very Large Data Bases. San Fransisco:Morgan Kaufmann Publishers, Inc. , 1996. 318～ 329.

[14] Savasere A., E. Omiecinski, and S. Navathe. Mining for strong negative associations in a large database of costomer transactions. Proceedings of the International Conference on Data Engineering, February 1998.

[15] Han J.,J.Pei,and Y.Yin.Mining frequent patterns without candidate generation.In Proc.2000 ACM-SIGMOD Int.Conf.Management of Data(SIGMOD'00),Dalas,TX,May 2000.

[16] Boulicaut J.-F., A. Bykowski and B. Jeudy, Mining association rules with negations, Technical Report 2000-19, INSA Lyon – LISI, Institut National des Sciences Appliqu´ees de Lyon, Bˆatiment Blaise Pascal, F-69621 Villeurbanne, France, 2000.

[17] M. Ester, H.P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in:Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 1996, pp. 226－231.

[18] Fayyad U , Piatetsky-Shapiro G, Smyth P, The KDD process for extracting useful knowledge from volumes of data. Communi-cations of the ACM, 1996 , 39(11):27～35

[19] E. Knorr, R. Ng, Finding aggregate proximity relationships and commonalities in spatial data mining, IEEE Transactions on Knowledge and Data Engineering 8 (6) (1996) 884－897.

[20]Jiangping Chen, An algorithm about spatial association rule mining based on cell pattern, Geoinformatics2006,wuhan.

[21]Jingsheng Zhai,Changqing Zhu,the algebraic reprenstation and shape chaging of the spatial graph,press of surveying and mapping,Sep,2005.

[22]Cressie, N, Staitistics for spatial data(revised edition)New York:Wiley,1993.