Incremental Maintenance of Discovered Spatial Association Rules

L. Dong[1]
[1]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, China, 430079
Email: dl@whu.edu.cn

1. Introduction

Spatial association rule mining is expensive. In particular, spatial analysis over large quantities of spatial data is necessary in the data preparation or mining procedure. When new data arrives or the orientation is changed, repeating similar mining procedure over similar data with similar parameters is inevitable. To reduce this workload, incremental maintenance of discovered spatial association rules is a noteworthy approach.

The study of incremental algorithms is emerging several years just after the concept of association rule suggested by Agrawal et al. (1993), and it emphasizes appended transactions handling (Cheungt and Vincent 1996; Yu and Bian 2007). However, some spatial association rule mining algorithms rely on spatial layers rather than transactions (Estivill-Castro and Lee 2001; Sha 2010). To put it in practice, this paper has further proposed incremental maintenance methods for the mining results of these layers-based mining algorithms.

2. Incremental Maintenance of Rules

There are two main motivations for people redoing association rule mining upon spatial datasets: 1) the thresholds of rule mining are to be modified; and 2) the datasets are to be updated.

2.1 Thresholds adjustment

The minimum support threshold and minimum confidence threshold are elemental parameters for rule mining algorithms, which are often denoted as min_sup and min_conf (Han et al. 2011). In general, the association rule mining procedure can be divided into two steps. In the first step, frequent itemsets (predicate sets) are extracted according to *min_sup*, then in the second step strong association rules are generated according to the results of last step and min_conf. To a large extent, min_sup and min_conf determines the quantity and quality of results of mining algorithms, modification of either may lead to different mining results.

To cope with modification of thresholds, firstly, the frequent predicate sets extracted in the first step of rule mining should be maintained up to date if min_sup had been changed. If the new value of min_sup is larger than the old one, frequent predicate sets can be updated just by removing the ones with lower support. Otherwise, incremental update algorithm ISA (Incremental Spatial Apriori) should be applied.

After frequent predicate sets updated, redo the second step of rule mining with new min_conf is enough to get association rules. As a special case, if both min_conf and min_sup are no less than before, new rules can be fetched just by filtrating old rules using new thresholds.

2.2 Dataset changes adaption

Different data produces different mining results. For layers-based mining algorithms, updates of input layers can be decomposed into removing and appending operations. For example, replacing layer A with A' can be implemented by removing layer A then appending A'. Hence, how to deal with these two kinds of operations are describe below.

Once a layer is removed, the corresponding predicate is impossible to exist in frequent predicate sets and association rules. To update the mining results, related predicate sets and rules should be removed. Other layers will not be affected, no further processing is needed to obtain updated results.

However, it is much more complex to deal with layer appendments. Although adding a layer won't affect the correctness of discovered rules, extra predicate sets should be examined. Given n layers ($n \geq 2$), there are at most $2^n - 1$ frequent predicate sets (Han et al. 2011), that means adding one layer would double the total count of frequent predicate sets in the worst case (from $2^n - 1$ to $2^{n+1} - 1$). Algorithm ISA is also applicable for this case.

When both kinds of layer operations are needed, it is suggested to process the removing operation first, to decrease the count of predicates.

2.3 Algorithm ISA

Algorithm ISA aims to update discovered frequent predicate sets when the minimum support threshold is changed or extra layers are appended. As an incremental variant of the Apriori algorithm (Agrawal and Srikant 1994), ISA generates candidate predicate sets, and then test if their support is larger than sup_min. The main difference lies in the candidate generation procedure, ISA can utilize discovered mining results to decrease the number of candidate sets.

For support changes, ISA will not test any discovered frequent predicate sets for relieving the burden of support calculating. That is because they should still be frequent after sup_min decreased.

For layer appendments, not only known frequent predicate sets can be skipped, but also known infrequent predicate sets, for layer appending won't affect the support of predicate sets, infrequent predicate sets would remain infrequent.

3. Experiments

In our experiment, 38 raster layers are used to test the ISA algorithm. These layers are generated from land cover data of No. 3 sample block in the Willamette Valley Ecoregion in years of 1972, 1979, 1985, 1992 and 2000. These land cover data are freely downloaded from the USGS's website. These layers can be divided into 5 groups, each group stands for the land cover status of one year, each layers in a group stands for the distribution of a certain type of land cover in the corresponding year.

For threshold changes, two groups of experiments are finished. The blank group uses an Apriori-like algorithm to mine frequent predicate sets under different thresholds, and the control group uses ISA to finish the same work. Experimentally, the most recent results available are always selected as the input of ISA. sup_min, the number of candidate predicates tested and time consumption are shown in table 1.

Table 1. Incremental mining for threshold changes

sup_min	Frequent predicate sets fetched	Candidate predicate sets tested		Time consumption per frequent predicate set (ms)	
		Blank group	Control Group	Blank group	Control group
0.1	124	262	-	273	-
0.01	207	509	385	311	229
0.001	349	769	562	300	212
0.0001	498	948	599	282	164
0.00001	510	958	460	279	104
0.000001	510	958	448	279	98

For layer changes, the experiments are similar, the blank group uses Apriori-like algorithm, and the control group uses ISA. The number of mining layers, the number of candidate predicates tested and time consumption are shown in table 2.

Table 2. Incremental mining for layer changes

Layer		Frequent predicate sets fetched	Candidate predicate sets tested		Time consumption per frequent predicate set (ms)	
Group	Count		Blank group	Control group	Blank group	Control group
2	14	28	63	-	236	-
3	21	77	186	123	271	205
4	29	203	443	257	279	185
5	38	510	958	515	290	179

In table 1, the difference between counts of candidate predicate sets in blank group (non-incremental) and control group (incremental) equals to the number of frequent sets in discovered results used by incremental mining. In table 2, it equals to the number of candidate sets tested. It is experimentally found that incremental mining can save up to 60% time as compared to fully re-mining.

4. Conclusion

In this paper, incremental maintenance strategies of discovered spatial association rules are designed. Under most conditions, mining result can be quickly updated without extracting frequent predicate sets again, except when *sup_min* decreased or new layers appended. For such changes, incremental mining algorithm ISA, which can use known frequent or candidate predicate sets to reduce the workload, is proposed. The incremental mining algorithm ISA is tested with real data. Experiments have demonstrated that ISA is more efficient than the non-incremental algorithm.

However, the methods proposed in our study do not take layers involved with discovered predicate sets and rules into account. These layers are often available after frequent predicate sets extraction. How to efficiently improve rule maintenance by making a full use of these informative layers is our next research work.

5. Acknowledgements

This research has been partially funded by National High Technology Research and Development Program of China（No. 2012AA1214002）.

References

Agrawal R, Imielinski T and Swami A, 1993, Mining Association Rules between Sets of Items in Large Databases. In: *Proceedings of the 1993 ACM SIGMOD Conference*, 207-216.

Agrawal R and Srikant R, 1994, Fast Algorithms for Mining Association Rules. In: *Proceedings of 1994 International Conference on Very Large Databases*, 487-499.

Cheungt D W and Vincent T, 1996, Maintenance of Discovered Association Rules in Large Databases : An Incremental Updating Technique. In: *Proceedings of the Twelfth International Conference on Data Engineering*, New Orleans, Louisiana, USA, 106–114.

Estivill-Castro V and Lee I, 2001, Data Mining Techniques For Autonomous Exploration of Large Volumes of Geo-referenced Crime Data. In: *Proceedings of the 6th International Conference on geocomputation*, Brisbane, Australia.

Han J, Kamber M and Pei J, 2011, Data Mining: Concepts and Techniques, 3rd edition. Morgan Kaufmann.

Sha Z, 2010, Mining Local Association Patterns from Spatial Dataset. In: *Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, Yantai, China, 1455–1460.

Yu L and Bian F, 2007, An Incremental Data Mining Method for Spatial Association Rule in GIS Based Fireproof System. In: *2007 International conference on Wireless Communications, Networking and Mobile Computing*, Shanghai, China, 5983–5986.