# A Novel Approach of Multilevel Positive and Negative Association Rule Mining for Spatial Databases

L.K. Sharma[1], O. P. Vyas[1], U. S. Tiwary[2], R. Vyas[1]

[1] School of Studies in Computer Science
Pt. Ravishankar Shukla University Raipur (C.G.) 492010-India
{lksharmain, dropvyas, ranjanavyas}@gmail.com
[2] Indian Institute of Information Technology, Allahabad- India
ust@iiita.ac.in

**Abstract.** Spatial data mining is a demanding field since huge amounts of spatial data have been collected in various applications, ranging form Remote Sensing to GIS, Computer Cartography, Environmental Assessment and Planning. Although there have been efforts for spatial association rule mining, but mostly researchers discuss only the positive spatial association rules; they have not considered the spatial negative association rules. Negative association rules are very useful in some spatial problems and are capable of extracting some useful and previously unknown hidden information. We have proposed a novel approach of mining spatial positive and negative association rules. The approach applies multiple level spatial mining methods to extract interesting patterns in spatial and/or non-spatial predicates. Data and spatial predicates/association-ship are organized as set hierarchies to mine them level-by-level as required for multilevel spatial positive and negative association rules. A pruning strategy is used in our approach to efficiently reduce the search space. Further efficiency is gained by interestingness measure.

## 1 Introduction

A spatial association rule describes the implication of a feature or a set of features by another set of features in spatial databases. A spatial association rule [5] is a rule of the form "A → B", where A and B are sets of predicates, some of which are spatial ones. In large spatial databases, many association relationships may exist but most researchers [5], [6], [7] focus only on the patterns that are relatively "Strong" i.e. the patterns that occur frequently and hold. In most cases the concepts of minimum support and minimum confidence are used. Informally the support of a pattern A in a set of spatial objects S is the probability that a member of S satisfies pattern A, and the confidence of "A → B", is the probability that pattern B occurs if pattern A occurs.

A large number of robberies/crimes are committed in a large metropolitan area. A criminologist who analyzes the pattern of these robberies may visualize crime sites using a GIS system and use the maps presenting locations of other objects. A financial analyst can do the real estate investment analysis, such as price changes of houses in different localities, using maps and location-specific characteristics. It has become an essential software tool for government services making decisions, analysis, planning and management of census, voting, mining and mineral exploration, systems for consultation and integration.

In spatial databases certain topological relationships hold at all times. Such topological relationships can be viewed as spatial association rules with 100% confidence, for example the containment relationship, which can be expressed as;

$$\textit{contain}\ (X, Y)\ \Lambda\ \textit{contain}\ (Y, Z) \rightarrow \textit{contain}\ (X, Z) \qquad (1)$$

However a problem with such a process is that the selection of interesting patterns has to be performed only on frequent patterns. Standard association rules are not enough expressive for some applications, so we need to mine not only frequent patterns but also infrequent patterns. Mining of infrequent patterns is known to be intractable. For example in crime site analysis of a criminologist, one can analyze pattern of robberies using maps presenting locations of other objects to find patterns, but it is also very important to know the infrequent patterns involved on the location.

Unlike existing spatial mining technique, in this paper we extend the traditional spatial associations to include infrequent patterns or negative spatial association rule mining in the following form;

$$\textit{contain}\ (X, Y)\ \Lambda\ \textit{contain}(Y, Z) \rightarrow \textit{not\_contain}(X, Z) \qquad (2)$$

Mining negative spatial association rules is a difficult task due to the fact that there are essential differences between positive and negative association rule mining. In mining task, the possible negative rules can be quite more than positive association rules, but the user may not be interested in all positive and negative association rules. In this paper we also discuss how can one find out the interesting positive and negative spatial association rules. This technique makes computation faster. The rest of this paper is organized as follows. In next section we present some related concepts and definition of spatial association rule. In section 3, we discuss the pruning strategy for mining spatial positive and negative association rule. In section 4, we discuss the spatial positive and negative association rule and finally in section 5 we discuss the efficiency and other features of the algorithm and the conclusions and the scope of the future work

## 2 Spatial Association rule

Most researchers [5][6][7] used rules reflecting structure of spatial objects and spatial/spatial or spatial/nonspatial relationships that contain spatial predicates, e.g. ***adjacent_to, near_by, inside, close_to, intersecting,*** etc. Spatial association rules can represent object/predicate relationships containing spatial predicates. For example, the following rules are spatial association rules.

Nonspatial consequent with spatial antecedent(s)

*is_a*(X, town) ∧ *intersects*(X, highway) → *adjacent_to* (X, water)… (80%).

Spatial consequent with non-spatial/spatial antecedent(s)

*is_a* (X, gas_station) → *close_to*(X, highway) ……………………… (75%).

Various kinds of spatial predicates can be involved in spatial association rules. They may represent topological relationships between spatial object, such as *disjoint, intersects, inside/outside, adjacent_to, covers/covered_by, equal*, etc. They may also represent spatial orientation or ordering, such as *left, right, north, east,* etc, or contain some distance information, such as *close_to, far_away*, etc. For systematic study of the mining of spatial association rules, some preliminary concepts are discussed in [6], as follows;

**Definition 1.** A Spatial association rule is a rule of the form;

$P_1 \wedge P_2 \wedge P_3 \wedge \ldots\ldots \wedge P_m \rightarrow Q_1 \wedge Q_2 \wedge Q_3 \wedge \ldots\ldots \wedge Q_n$ (c %)…………( 3)

Where at least one of the predicates $P_1 \wedge \ldots \wedge P_m$ , $Q_1 \wedge \ldots \wedge Q_n$ is a spatial predicate, and c% is the confidence of the rule which indicates that c% of objects satisfying the antecedent of the rule will also satisfy the consequent of the rule.

**Definition 2**. A rule "P → Q/S" is strong if predicate "P Λ Q" is large in set S and confidence of "P→ Q/S" is high.

The above definition for an association rule P → Q has a measure of the strength called confidence (denoted as conf) defined as the ratio supp $(P \cup Q)$/supp (P), where $P \cup Q$ means that both P and Q are present.

Association rule discovery seeks rules of the form P → Q with support and confidence greater than or equal to, user specified support (ms) and minimum confidence (mc) thresholds respectively. This is referred to as the support-confidence framework [1] and the rule P → Q is an interesting positive association rule. An item set that meets the user specified minimum support is called frequent item set. Accordingly an infrequent item set can be defined as an item set that does not meet the user specified minimum support. Like positive rule, a negative rule $P \rightarrow^{\neg} Q$ also has measure of its strength, confidence, defined as the ratio supp $(P \cup {}^{\neg}Q)$/supp (P) where supp $({}^{\neg}Q)$ can be measured by 1- supp (Q). This infrequent item set may be significant as illustrated by following example [9].

**Example 1**. Let supp(c) = 0.6, supp (t) = 0.4, supp $(t \cup c)$ = 0.05 and mc = 0.52. The confidence of t → c is supp $(t \cup c)$/supp (t) = 0.05/0.4 = 0.125 < mc (= 0.52) and supp $(t \cup c)$ = 0.05 is low. This indicates that $t \cup c$ is an infrequent item set and that t → c cannot be extracted as rule in support confidence framework. However, supp $(t \cup {}^{\neg}c)$ = supp (t) - supp $(t \cup c)$ = 0.4 – 0.05 = 0.35 is high and the confidence of $t \rightarrow^{\neg} c$ is the ratio supp $(t \cup {}^{\neg}c)$ / supp (t) = 0.35/04 = 0.875 > mc. Therefore $t \rightarrow^{\neg} c$ is a valid rule.

By extending the definition in [5] [6] [7] negative spatial association rule discovery is proposed to be defined as follows:

**Definition 3.** The support of a conjunction of predicate, $P = P_1 \wedge \ldots \wedge P_m$ , in a set S denoted as supp (P/S), is the number of objects in S which satisfy P versus the cardinality of S. The confidence of rule $P \rightarrow^{\neg} Q$ is the ratio of supp $(P \wedge {}^{\neg} Q /S)$ versus supp (P/S) i.e. the possibility that a member of S does not satisfy Q when the same member of S satisfies P**.** A single predicate is called 1-predicate. A conjunction of k single predicates is called a k-predicate.

Our study of spatial association relationship is confined to newly formed Chhattisgarh (C.G.) state in India whose map is presented in Figure 1 with the following database relations for organizing and representing spatial objects:

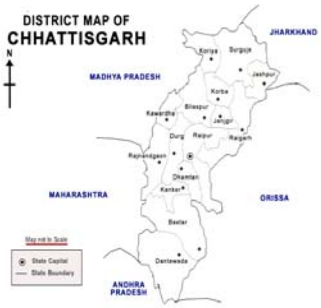


Fig. 1. Chhattishgarh State in India

Town (town_name, town_type, population, literacy, geo…)
Road (road_name, road_type, geo…)
Water (water_name, water_type, geo…)
Boundary (type, admin_region, geo...)
Mine (mine_name, mine_type, geo…)
Forest (forest_name, forest_type, geo…)

It may be noted that in the above relational schema, the attribute "geo" represents a spatial object (a point, line, area, etc.) whose spatial pointer is stored in a tuple of the relation and points to a geographic map. The attribute "type" of a relation is used to categorize the types of spatial objects in the relation. For example, the type for road could be {national highway, state highway, …} and the type of water could be {rivers, lakes, …}. The boundary could be boundary between two state regions such as Chhattisgarh and Maharastra in India.

To facilitate mining multiple level association rules and efficient processing, concept hierarchies are provided for both data and spatial predicates.

A set of hierarchies for data relations is defined as follows.

A concept hierarchy for *town*

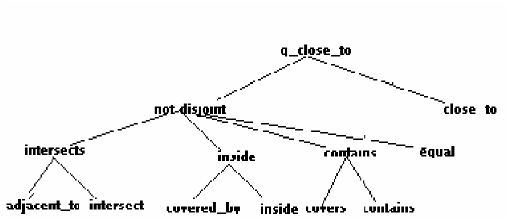
(town (large town (big city (Raipur, Bilaspur, Durg …)), medium size (…),..)))

A concept hierarchy for *water*

(water (river (large river (Mahanadi, Kharun, Shivnath, …))…) ) etc

Spatial predicates (topological relations) should also be arranged into a hierarchy for computation of approximate spatial relations (like "g_close_to see Figure - 2) using efficient algorithms with coarse resolution at a higher concept level and refining the computations when it is confined to a set of more focused candidate objects.

**Fig. 2.** Approximate spatial relations



## 3. Identification of Interesting Item set

There can be an exponential number of predicates in a database and only some of them are useful for mining association rule of interest. Therefore it is also an important issue to efficiently search the interesting itemset. In this paper we use a pruning strategy [9] to find out potentially interesting itemset. An interestingness function [9], interest (X, Y) = $|\text{supp}(X \cup Y) - \text{supp}(X)\,\text{supp}(Y)|$ and a threshold mi (minimum interestingness) are used. If interest (X, Y) ≥ mi, the rule X → Y is of potential interest, an X ∪ Y is referred to as *potentially interesting intemset.* Using this approach, we can establish an effective pruning strategy for efficiently identifying all frequent itemsets of potential interest in a database.

Integrating this interest (X, Y) mechanism into the support-confidence framework, 'I' is a frequent itemset of potential interest (fipi) if:

$$\begin{aligned} &\text{fipi}(I) = \text{supp}(I) \geq ms \wedge \\ &\exists\, X, Y: X \cup Y = I \wedge \\ &\qquad \text{fipis}(X, Y) \end{aligned} \qquad (4)$$

Where $\text{fipis}(X, Y) = X \cap Y = \varnothing \wedge$

$$f(X, Y, ms, mc, mi) = 1 \qquad (5)$$

$$f(X, Y, ms, mc, mi) = \frac{\text{supp}(X \cup Y) + conf(X \rightarrow Y) + interest(X, Y) - (ms + mc + mi) + 1}{|\,\text{supp}(X \cup Y) - ms\,| + |\,conf(X \rightarrow Y) - mc\,| + |\,interest(X, Y) - mi\,| + 1}$$

Where *f* () [9] is a constraint function concerning the support, confidence and interestingness of X → Y. Integrating the above insight and the interest (X, Y) mechanism into the support-confidence framework, J is an infrequent itemset of potential interest (iipis) if

$$\begin{aligned} &\text{iipis}(J) = \text{supp}(J) < ms \wedge \\ &\exists\, X, Y: X \cup Y = J \wedge \\ &\qquad \text{iipis}(X, Y) \end{aligned} \qquad (6)$$

Where $$\text{iipis}(X, Y) = X \cap Y = \varnothing \wedge$$

$$g(X, \neg Y, ms, mc, mi) = 2 \quad (7)$$

$$g(X, \neg Y, ms, mc, mi) = f(X, Y, ms, mc, mi) + \frac{supp(X) + \text{supp}(Y) - 2mc + 1}{|\,\text{supp}(X) - \text{ms}\,| + |\,\text{supp}(Y) - \text{ms}\,| + 1}$$

Where g () [9] is a constraint function concerning f () [9] and the support, confidence, and interestingness of X→Y.

# 4. A Method for Mining Spatial Association Rules

## 4.1 An Example of Mining Spatial Association Rule

**Example 2.** We examine how the data-mining query posed in Example 1 is processed, which illustrates the method for mining spatial association rules for Chhattisgarh state.

Firstly, a set of relevant data is retrieved by execution of the data retrieval methods [2] on the data-mining query. This extracts the following data sets whose spatial portion is inside Chhattisgarh: (1) towns: only *tahsil place* (2) road: National and state highway (3) water: only rivers, lake.

Secondly the generalized *close_to* relationship between towns and the other four classes of entities is computed at a relatively coarse resolution level using a less expensive spatial algorithm such as the MBR data.

**Table 1.** Large k-predicate sets at the first level (for 50 towns in Chhattisgarh)

| K | Large k - predicate set | Count |
|---|---|---|
| 1 | <Adjacent to, water > | 29 |
| 1 | <Intersect to, highway> | 25 |
| 1 | <close to, highway> | 30 |
| 1 | <close to, state boundary> | 25 |
| 2 | <Adjacent to, water ><Intersect to, highway> | 20 |
| 2 | <Adjacent to, water ><close to, highway> | 20 |
| 2 | <Adjacent to, water ><close to, state boundary> | 18 |
| 2 | <close to, highway><close to, state boundary> | 15 |
| 3 | <Adjacent to, water ><Intersect to, highway><close to, state boundary> | 10 |
| 3 | <Adjacent to, water ><close to, highway><close to, state boundary> | 8 |

Spatial association rules can be extracted directly from table 1. To illustrate this, the object set of interest; f (X, Y, ms, mc, mi) can be replaced with the following f(X, Y, ms, mi) = $\dfrac{\text{supp(X } \cup Y) + \text{interest } (X, Y) - (ms + mi) + 1}{|\text{supp } (X \cup Y) - ms \,| + |\text{interest } (X, Y) - mi \,| + 1}$

For example (intersect highway) has support 0.5 and (adjacent to, water), (intersect, highway) has support 0.4 and if we consider ms = 0.2 and mi = 0.1 then f (<intersect, highway>, <adjacent _to, water>, 0.2, 0.1) = $\dfrac{0.4 + 0.05 - 0.1 + 1}{|\, 0.4 - 0.2 \,| + |\, 0.05 - 0.1 \,| + 1} >$ 1

**Table 2.** Large k-predicate sets at the second level (for 50 towns in Chhattisgarh)

| K | Large k - predicate set | Count |
|---|---|---|
| 1 | <Adjacent to, river> | 20 |
| 1 | <Intersect to, national highway> | 15 |
| 1 | <close to, national highway> | 25 |
| 1 | <close to, MP state boundary> | 20 |
| 2 | <Adjacent to, river ><Intersect to, National highway> | 15 |
| 2 | <Adjacent to, river ><close to, National highway> | 15 |
| 2 | <Adjacent to, river ><close to, MP boundary> | 13 |
| 2 | <close to, National highway><close to, MP state boundary> | 10 |
| 3 | <Adjacent to, river ><Intersect to, national highway><close to, MP boundary> | 5 |
| 3 | <Adjacent to, river ><close to, National highway><close to, MP state boundary> | 5 |

Spatial association rules can be extracted directly from table 2. For example (close to, MP boundary) has support 0.4, (adjacent to, river) has support 0.4, and (adjacent to, river) (close to, MP boundary) has support 0.26 and if we consider ms = 0.15 and mi = 0.1 then f (<adjacent to, river>, <close to, MP boundary> 0.15, 0.1) =

$$\frac{0.26 + 0.1 - 0.05 + 1}{|\, 0.26 - 0.15 \,| + |\, 0.1 - 0.1 \,| + 1} > 1$$

### 4.2 An Algorithm for Mining Spatial Association Rules

**Algorithm 4.1** Mining the spatial positive and negative association rules in a large spatial database.

**Input:** The input consists of a spatial database, a mining query and a set of thresholds as follows:

i. A spatial database SDB and set of concept hierarchies.
ii. A query of a reference class set of task relevant classes for spatial object and a set of task relevant spatial relations.
iii. Three thresholds: minimum support, minimum confidence, and minimum interestingness.

**Output:** Strong spatial positive and negative association rules for the relevant sets of objects and relations.

The above algorithm can be summarized in the following way:

```
   Producer find_large_interested_predicate (SDB)
```

```
(1)for(l = 0 ; L[l,1] != 0 and l < max_level; l++)
(2)PL[l] ← 0; NL[l] ← 0
(3) let L[l,1]= get_predicates(SDB, l);
    PL[l] ← PL ∪ L[l,1]
(4)for (k = 2; L[l, k-1] !=0; k++)do begin
  (4.1) let $P_k$ = get_candidate_set (L [l,, k-1]);
  (4.2) for each object s in S do begin
  (4.3) $P_s$= get_subsets (P, s);//Candidate satisfied by s
  (4.5) for each object set p ∈$P_s$ do p.supp++;
  (4.6) end;
  (4.7) Let L[l,k]←{p|p∈$P_k$∧(supp(c)=(c.supp/| SDB|)>=ms);
  (4.8) Let N [l, k] ← $P_k$- L [l, k];
  (4.9) for each object set I in L [l, k] do
              if Not(fipi(I)) then
              let L[l,k] = L[l,k] – {I};
              let PL[l] = PL[l] ∪ L[l,k];
  (4.10) for each object set J in N [l, k] do
              if NOT(iipi(J)) then
              let N[l,k] = n[l,k] - {J};
              let NL[l] = NL[l] ∪N[l,k];
              end
            end
       end
(5) Output = generate_association_rules (PL[l], NL[l])
      end
```

In this procedure, step (1) shows that the mining of the positive and negative association rules is performed level by level, starting from the top most level until either the large 1-predicate set table is empty or it reaches the maximum concept level for each level l, step (3) computes the large 1-predicate sets and puts into table *L[l, 1]* , step(4) computes the potentially frequent and infrequent itemsets, which is stored respectively as *PL[l],* NL[l] and finally the algorithm generates the spatial positive and negative association rules at each concept level from the frequent predicate table PL[l] and infrequent predicate table NL[l].

## 5. Implementation

The Algorithm explained here was implemented taking thematic map data of Chhattisgarh state of India and using programming language JAVA. The experiment was performed on a Pentium IV having 128 MB RAM.

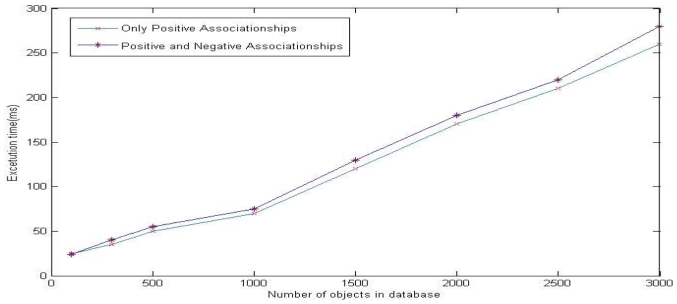


Fig. 3. Graph showing performance of ARM algorithm generating multilevel positive and negative associationships

The algorithm generated multilevel positive and negative associationships. Figure 3 shows the performance of the algorithm for generating both association rules. It is evident that the execution time is increasing with the number of objects in database but the increase for large number of positive and negative association rules is not enormous in view of the fact that the number of negative associations are reasonably large. This justifies the use of our proposed algorithm to mine positive and negative association rules simultaneously. The algorithm proposed in this paper is efficient for mining multiple level potentially interesting spatial positive and negative association rules in spatial database. We have used a pruning strategy [9] to efficiently reduce the search space.

## 6. Conclusion

Spatial data mining is used in areas such as remote sensing, traffic analysis, climate research, biomedical applications including medical imaging and disease diagnosis. The algorithm presented in this paper discusses efficient mining procedures for spatial positive and negative association rules. It explores techniques at multiple approximation and abstraction levels. Further, efficiency is gained by interestingness measure, which allows us to greatly reduce the number of associations needed for consideration. In our proposed approach approximate spatial computation is performed initially at an abstraction level on a large set of data, which substantially reduces the set of candidate data to be examined in the next levels. The outcome of the above mentioned spatial association rule algorithm is a set of association rules in which either the antecedent or the consequent of the rule must contain some spatial predicates (such as *close_to*):

- Non-spatial antecedent and spatial consequent: All elementary schools are located close to single-family housing developments.
- Spatial antecedent and non-spatial consequent: If a house is located in a Park, it is expensive.
- Spatial antecedent and spatial consequent: Any house that is near downtown is situated in the south of Chhattisgarh.

This algorithm works in a similar manner as the Apriori algorithm with negation and interestingness function in the "large predicate sets". Here predicate set is a set of predicates of interest. A 1-predicate might be {(close_to, water)}, so all spatial objects that are *close_to* water will be counted as satisfying this predicate. Similarly a 2-predicate sets can be counted, and so on. In actuality the algorithm can be used to generate multilevel positive and negative association rules at the desired coarse level or a fine level. The outcome of the algorithm can be interpreted to find the interesting associations between the spatial predicates and non-spatial predicates.

## References

1. Agrawal, R., Srikant, R.: Fast algorithm for mining association rules. In Proc. 1994 Int. Conf. VLDB Santiago, Chile, Sept. (1994) 487-499.
2. Guetting, R.H.: An Introduction to Spatial Database Systems. Special Issue on Spatial Database System of the VLDB Journal, October 1994 vol 3, No 4
3. Han, J., Fu, Y.: Discovery of Multiple Level association rules from large database. Proc. of the Int. Conf. VLDB (1995) 420-431
4. Han, J., Pei, J., Yin, Y., Mao, R.: Mining Frequent Patterns with out Candidate Generation: A Frequent Pattern Tree Approach. Kluwer Publication, Netherlands (2003)
5. Malerba D. Lisi F.A.: An ILP method for spatial association rule mining. Working notes of the first workshop on Multi Relational Data mining, Freiburg, Germany (2001) 18-29.
6. Malerba D., Lisi, F. A., Analisa Appice, Francesco. : Mining Spatial Association Rules in Census Data: A Relational Approach", 2001
7. Shekhar, S., Chawla, S., Ravadam,S., Liu, X.,Lu, C.: Spatial Databases- Accomplishments and Research Needs. IEEE Transactions on Knowledge and Data Engineering (1999) Vol. 11, No. 1
8. Smith, G.B., Bridge, S.M.: Fuzzy Spatial Data Mining. IEEE Transactions on Knowledge and Data Engineering, 2002.
9. Wu, X., Zhang, C., Zhang, S.: Efficient Mining of Both Positive and Negative Association rule. ACM Tran. On Information System (2004) Vol22.No.3, 381-405
10. Dunham, M. H.: Data Mining Introductory and Advance Topics. Pearson Education Inc, (2003)