Filtering Frequent Spatial Patterns with Qualitative Spatial Reasoning

Vania Bogorny, Bart Moelans
Hasselt University & Transnational University of Limburg - Theoretical Computer Science Group
Agoralaan Gebouw D, 3590 Diepenbeek - Belgium
{vania.bogorny, bart.moelans}@uhasselt.be

Luis Otavio Alvares
Universidade Federal do Rio Grande do Sul - Instituto de Informatica
Av. Bento Goncalves, 9500 - Campus do Vale, 91501-970 Porto Alegre - Brasil
alvares@inf.ufrgs.br

Abstract

In frequent geographic pattern mining a large amount of patterns can be non-novel and non-interesting. This problem has been addressed recently, and background knowledge is used to reduce well known geographic patterns. However, a large amount of meaningless patterns which is independent of domain knowledge is still extracted from geographic data. Therefore, this paper proposes a method for filtering specific types of meaningless spatial patterns using qualitative spatial reasoning. We proof a significant reduction of the number of frequent patterns, which is also shown with experiments performed on real data. These experiments even show a reduction in computational time.

1. Introduction

The huge amount of patterns generated by frequent pattern mining algorithms has been extensively addressed in the last few years. Different objective and subjective measures have been proposed to evaluate how interesting association rules are. However, according to [5] it is difficult to come up with a single metric that quantifies the "interestingness" or "goodness" of an association rule. In most approaches, non-interesting rules are eliminated during the rule generation, i.e., aposteriori, when frequent itemsets have already been generated.

In spatial frequent pattern mining the number of non-interesting association rules can increase even further than in transactional pattern mining. Geographic data have semantic dependencies and spatial properties which in many cases are well known and non-interesting [3, 8].

Figure 1 shows a partial map of the city of Porto Alegre, located in southern Brazil, where the large polygons

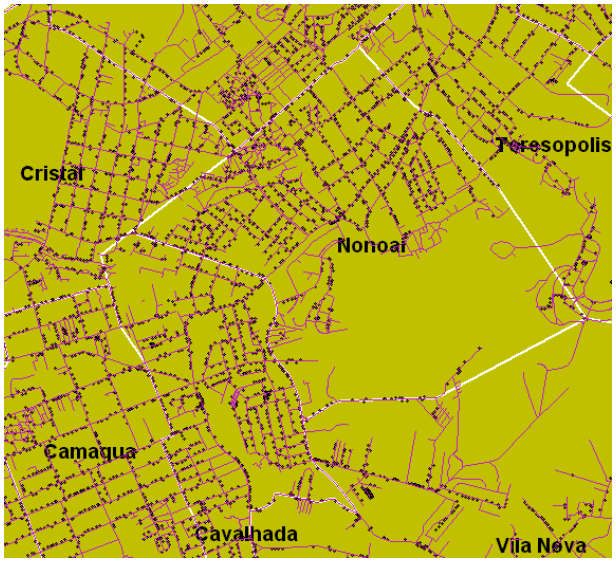

Figure 1. Districts, Streets, and Illumination Points of the Porto Alegre City

are districts, black lines are streets and black dots are illumination points. In this example we can see at least two semantic dependencies which generate well known patterns. Illumination points are adjacent to streets, and all streets are related to at least one district. Such dependencies generate non-interesting association rules such as, for example:

$$is\_a\_District \wedge contains\_Street \rightarrow contains\_IlluminationPoints.$$

In the geographic domain, not only well known geographic dependencies generate a large number of patterns

without novel and useful knowledge. Different spatial predicates may contain the same geographic object type.
In transactional frequent pattern mining items have binary values, and either do or do not participate in a transaction. For instance, the item $milk$ either participates or not in a transaction t. In spatial frequent pattern mining the same spatial object (item) may have different qualitative spatial relationships with the target feature (transaction), and by consequence participate more then once in the same transaction t. For instance, a city C may *contain* an instance i_1 of river, be *crossed by* an instance i_2 of river, or even *touch* an instance i_3 of river. Different spatial relationships with the same geographic object type will generate associations such as

$$contains\_River \rightarrow touches\_River$$

when data are considered at general granularity levels [12]. This rule says that "cities which contain river do also touch river". It is well known that a city does not touch a river because it also contains a river. Such kind of rule is non-interesting for most applications. An interesting association rule would be the combination of any of these two predicates with a different geographic object type or some non-spatial attribute. For instance:

$$contains\_River \rightarrow WaterPollution = high$$

or

$$touches\_River \rightarrow exportationRate = high.$$

To understand the problem in real applications, let us consider Figure 2, which shows the same part of the city of Porto Alegre shown in Figure 1, with some different geographic object types. The large polygons represent a certain part of the 109 districts, black squares are police centers, dark polygons represent slums, and white dots are schools. Our objective is to investigate possible associations between the high criminality rates in different districts with slums, schools, and police centers. In our initial hypothesis, districts that have high criminality rates will be spatially related to slums, and districts with low criminality rate contain schools and police centers. For this problem we need to consider the data at a more general granularity level (e.g. $contains\_slum$, $contains\_school$), without taken into account the instance of each relevant feature type (e.g. $contains\_slum159$, $contains\_school20$), since we are not interested in specific instances of schools, police centers, or slums.

Considering districts as the reference object type and slums, police centers, and schools as the relevant object types, the first step is to compute the different spatial relationships of each district with all relevant feature types. These relationships may be of type topological, distance, or

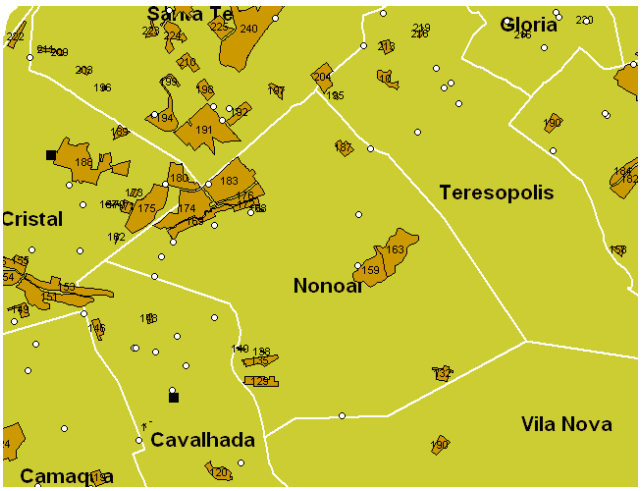

Figure 2. Districts, Slums, Schools, and Police Centers of the Porto Alegre City

order [11]. Notice in Figure 2 that the district "Nonoai", for instance, has many topological relationships with different instances of slum. It *touches* slum180, *covers* slum183, *overlaps* slum174 and *contains* slum159. Considering distance relationships and police centers, the district Nonoai will be either *close* or *far* from the police centers according to the distance threshold. Districts Cristal and Cavalhada, however, will be *very close*, since they contain police centers. These different relationships that districts can have with different geographic object types (e.g. slums, police centers) generate many predicates with the same feature type, and by consequence, meaningless association rules. In this example, non-interesting association rules such as

$$is\_a\_District \rightarrow contains\_slum \wedge touches\_slum$$

or

$$contains\_school \rightarrow touches\_school$$

will be generated. With distance relationships we can have rules with even less meaning. In the example, notice that the district Cristal for instance, will be *very close* to one police center but *far from* all other police centers in the city. In this case we will have rules such as

$$is\_a\_District \rightarrow closeTo\_PoliceCenter \wedge farFrom\_PoliceCenter.$$

For each different relevant geographic object type (slum, police center, and school) there exists a possibility to generate different topological predicates, according to the 9-intersection model proposed by [10]: contains, within, touches, crosses, covers, coveredBy, overlaps, equals, and disjoint. For distance or order relationships, the higher the

number of specified qualitative relationships is (e.g. close, very close, far), the higher is the probability to generate non-interesting frequent patterns. This kind of patterns can be generated for any qualitative spatial relationship.

Since our objective is to find associations between the different geographic object types (slums, schools, and police centers) in different districts, associations of predicates with different qualitative spatial relationships, but with the same geographic object type, do only impede the pattern discovery and should be eliminated. Moreover, they do not express any co-relation between the different geographic object types.

Existing quantitative co-location pattern mining approaches such as [13] may not generate this kind of meaningless patterns. However, they have the disadvantage of considering only quantitative distance relationships and their input is restricted to point datasets, which is a significant limitation for real applications.

Only a few qualitative approaches in frequent geographic pattern mining address the problem of mining non-interesting patterns. In [14], for instance, such problems have not been addressed. In [3] well known patterns can be eliminated using background knowledge, but pruning is performed aposteriori. Existing metrics proposed to reduce the number of patterns in transactional databases do not warrant the elimination of non-interesting qualitative spatial patterns.

Recently, two of the present authors proposed different approaches to reduce the number of frequent geographic patterns using background knowledge. In [8] they proposed Apriori-KC, in which some changes were proposed on Apriori [1] to eliminate well known geographic patterns. In [7] they extended this method with an additional step where not only frequent itemsets are reduced, but input space is reduced as much as possible, since this is still the most efficient way for pruning frequent patterns. However, still a large number of frequent itemsets is generated in [7] and [8] although well known dependencies are eliminated very efficiently. In [9] they applied the closed frequent pattern mining approach to the geographic domain, eliminating both well known dependencies and redundant frequent itemsets. As a continued study in frequent geographic pattern mining, in this paper we propose a method to reduce the number of non-interesting spatial patterns using qualitative spatial reasoning. Not only the data is considered, but its semantics is taken into account, and frequent patterns that contain the same feature type are removed apriori.

In transactional frequent pattern mining some works address the problem of mining interesting association rules [5, 16, 17, 18]. They incorporate user-specified constraints or define either objective or subjective metrics of interestingness. Moreover, pruning is performed aposteriori, after the rule generation. The method proposed in this paper is more effective and efficient, since it explores the anti-monotone constraint of Apriori [2] and prunes non-interesting patterns during the frequent set generation. Other transactional pattern mining approaches such as [4, 19] remove redundant frequent patterns and association rules exploiting support and confidence constraints, while our method is independent of such thresholds. In these approaches "redundan" rules are eliminated by frequent itemset pruning, but "non-interesting" and "meaningless" rules are still generated.

Tuple (district)	Spatial Predicates
Teresopolis	murderRate=high, theftRate=low, contains_slum overlaps_slum, contains_school, touches_school
Vila Nova	murderRate=low, theftRate=low, contains_slum touches_slum, touches_school
Cavalhada	murderRate=low, theftRate=high, contains_slum touches_slum,overlaps_slum, contains_school touches_school, contains_policeCenter
Cristal	murderRate=high, theftRate=high, contains_slum overlaps_slum, covers_slum, contains_school touches_school, contains_policeCenter
Nonoai	murderRate=high, theftRate=high, contains_slum touches_slum,overlaps_slum, covers_slum contains_school, touches_school
Camaqua	murderRate=high, theftRate=low, contains_slum overlaps_slum, contains_school, touches_school

Table 1. Partial Dataset of the City of Porto Alegre where rows are Districts and items are Spatial and Non-Spatial Predicates Related to the Different Districts

The reminder of the paper is organized as follows. Section 2 explains the problem of mining frequent qualitative spatial patterns. Section 3 presents a filtering method to eliminate non-interesting patterns. Section 4 evaluates the method with experimental results. Finally in Section 5 is the conclusion of the paper and suggestions for some directions for future work.

2. The Problem of Mining Qualitative Spatial Patterns

We will explain the problem of meaningless patterns considering topological relationships, although it exists for any type of qualitative spatial relationship. Considering all possible topological relationships between districts and the relevant object types shown in Figure 2 without considering their instance, we can generate a dataset similar to the one shown in Table 1.

Now let us consider minimum support 50%. The frequent itemsets are those which appear in at least 3 rows in the dataset shown in Table 1. Considering such a high minimum support, Table 2 shows all possible frequent itemsets that can be generated from the dataset. Considering that the

dataset has 9 predicates, two non-spatial (murderRate and theftRate) and 7 spatial predicates, a total of 60 frequent itemsets with two or more elements is generated. Among the 60 frequent itemsets, 31 contain at least one pair with the same geographic object type, as represented in bold style in Table 2.

In this small example we see how many frequent predicate sets can be generated without expressing interesting knowledge. However, it is important to observe that the generation of frequent itemsets having the same spatial feature type varies according to both the value of minimum support and the dataset.

As can be observed in Table 2, the combination of pairs with the same feature type *slum* or *school* appear the first time during the generation of frequent itemsets with size $k = 2$. Then these pairs start replicating in a combinatorial explosion of frequent predicates sets. Considering this, the most efficient way to eliminate meaningless patterns in spatial frequent pattern mining is to remove the pairs in which such patterns appear the first time. This can be done in the second pass of the algorithm, by exploiting the anti-monotone constraint, as explained in the following section.

Listing 1. Apriori-KC+

```
INPUT:  Ψ, // a spatial dataset
        φ, //a set of pairs of dependencies
        minsup; //minimum support

L1 = {large 1-predicate sets};

FOR (k = 2; Lk−1! = ∅; k++)
 Ck = apriori_gen(Lk−1); //New candidates
 IF (k==2)
  C2 = C2 − φ; //remove dependencies
  FOR (∀ pairs p ∈ C2 with same feature type)
   remove p from C2;
 END

 FOR (all  rows w ∈ Ψ)
  Cw = subset(Ck, w); // Candidates ∈ w
  FOR (all candidates c ∈ Cw)
   c.count++;
 END

 Lk = {c ∈ Ck | c.count ≥ minsup};
END

RETURN ∪k Lk
```

3. Filtering non-interesting Spatial Patterns

Existing spatial frequent pattern algorithms generate candidates and frequent sets. In spatial frequent pattern mining the computational cost relies on the spatial predicate extraction (number of instances of both target and relevant feature types). Therefore the candidate generation is not a problem as in transactional databases, since the number of predicates is much smaller than the number of items [15].

Apriori [2] has been the basis for dozens of algorithms for mining spatial and non-spatial frequent itemsets. Although it generates a large number of frequent itemsets and association rules, spatial association rule mining algorithms are Apriori-like. In [8] two of the presenting authors extended Apriori to Apriori-KC, to eliminate well known geographic dependencies, where knowledge constraints (KC) are given as background knowledge. To eliminate frequent patterns which contain the same geographic feature type we need to perform some reasoning over the data, i.e., it is necessary to check the meaning of the items into a frequent itemset and if they can be either removed or not. For this purpose, we added one more step to Apriori-KC, which we will call Apriori-KC+, as shown in Listing 1.

In Apriori-KC+ besides the elimination of pairs of predicates with well known dependencies, we also eliminate pairs of predicates with the same feature type. However, the background knowledge given as an input to Apriori-KC+ is only necessary to eliminate dependencies. Background knowledge is not necessary to remove pairs with the same feature type, since this problem is independent of domain knowledge.

Going further into the algorithm, in the first pass, the support of the individual elements is computed to determine large-predicate sets. In the subsequent passes, given k as the number of the current pass, the large sets L_{k-1} in the previous pass $(k-1)$ are grouped into sets C_k with k elements, which are the candidate sets. This is performed by the apriori_gen function, described in [1].

The support of each candidate set is computed, and if it is equal or higher than minimum support, then this set is considered frequent. This process continues until the large set in the pass results in an empty set.

To eliminate meaningless patterns we added one step which is performed when $k = 2$, such that all pairs of elements that contain the same feature type are removed from C_2. By removing these pairs, no larger frequent predicate set will be generated with the same feature type. According to the anti-monotone constraint of Apriori [1] which states that if an itemset Z is infrequent no superset of Z will be frequent, the elimination of a pair having the same feature type will not generate any set with size $k > 2$. This makes

Size k	Frequent predicate sets with minsup = 50%
2	{murderRate=high, theftRate=low},{murderRate=high, contains_slum}, {murderRate=high, overlaps_slum},{murderRate=high, contains_school}, {murderRate=high, touches_school},{theftRate=low, contains_slum}, {theftRate=low, overlaps_slum}, {theftRate=low, contains_school}, {theftRate=low, touches_school}, {**contains_slum, overlaps_slum**}, {**contains_slum, touches_slum**}, {contains_slum, contains_school}, {contains_slum, touches_school}, {overlaps_slum, contains_school}, {overlaps_slum, touches_school}, {touches_slum, touches_school}, {**contains_school, touches_school**}
3	{murderRate=high,theftRate=low,contains_slum}, {murderRate=high,theftRate=low,overlaps_slum}, {murderRate=high,theftRate=low,contains_school}, {murderRate=high,theftRate=low,touches_school}, {**murderRate=high,contains_slum,overlaps_slum**}, {murderRate=high,contains_slum,contains_school} {murderRate=high,contains_slum,touches_school}, {murderRate=high,overlaps_slum,contains_school}, {murderRate=high,overlaps_slum,touches_school}, {**murderRate=high,contains_school,touches_school**}, {**theftRate=low,contains_slum,overlaps_slum**}, {theftRate=low,contains_slum,contains_school} {theftRate=low,contains_slum,touches_school}, {theftRate=low,overlaps_slum,contains_school}, {theftRate=low,overlaps_slum,touches_school}, {**theftRate=low,contains_school,touches_school**}, {**contains_slum,overlaps_slum,contains_school**}, {**contains_slum,overlaps_slum,touches_school**} {**contains_slum,touches_slum,touches_school**}, {**contains_slum,contains_school,touches_school**}, {**overlaps_slum,contains_school,touches_school**}
4	{**murderRate=high,theftRate=low,contains_slum,overlaps_slum)**}, {murderRate=high,theftRate=low,contains_slum,contains_school} {murderRate=high,theftRate=low,contains_slum,touches_school}, {murderRate=high,theftRate=low,overlaps_slum,contains_school} {murderRate=high,theftRate=low,overlaps_slum,touches_school}, {**murderRate=high,theftRate=low,contains_school,touches_school**} {**murderRate=high,contains_slum,overlaps_slum,contains_school**}, {**murderRate=high,contains_slum,overlaps_slum,touches_school**} {**murderRate=high,contains_slum,contains_school,touches_school**}, {**murderRate=high,overlaps_slum,contains_school,touches_school**} {**theftRate=low,contains_slum,overlaps_slum,contains_school**}, {**theftRate=low,contains_slum,overlaps_slum,touches_school**} {**theftRate=low,contains_slum,contains_school,touches_school**}, {**theftRate=low,overlaps_slum,contains_school,touches_school**} {**contains_slum,overlaps_slum,contains_school,touches_school**}
5	{**murderRate=high,theftRate=low,contains_slum,overlaps_slum, contains_school**} {**murderRate=high,theftRate=low,contains_slum,overlaps_slum, touches_school**} {**murderRate=high,theftRate=low,contains_slum,contains_school, touches_school**} {**murderRate=high,theftRate=low,overlaps_slum,contains_school, touches_school**} {**murderRate=high,contains_slum,overlaps_slum,contains_school, touches_school**} {**theftRate=low,contains_slum,overlaps_slum,contains_school, touches_school**}
6	{**murderRate=high,theftRate=low,contains_slum,overlaps_slum, contains_school,touches_school**}

Table 2. Frequent Itemsets of Table 1 with minimum support 50%

the approach effective and independent of any threshold such as minimum support, minimum confidence, lift, etc. Indeed, no background knowledge is required from the data mining user and the method eliminates the exact combinations which generate meaningless rules.

The proposed method does not sacrifice the result quality. For instance, suppose that $\{A, B\}$ is a frequent set having the same feature type. This pair is eliminated with the purpose to avoid the generation of larger frequent sets that contain the same feature type, such as $\{A, B, C\}$, for example. If the set $\{A, B, C\}$ has minimum support, then the pairs $\{A, B\}$, $\{A, C\}$, and $\{B, C\}$ have minimum support as well. As we eliminate only pairs with same feature types, which in this example is $\{A, B\}$, the sets $\{A, C\}$ and $\{B, C\}$ which combine the predicate C with both A and B separately, are still generated, and no information is lost.

In a practical example, a pair such as $\{contains\_slum, touches\_slum\}$ is removed to avoid the generation of a larger frequent predicate set such as $\{contains\_slum, touches\_slum, murderRate = high\}$. Although one can argue that such a frequent itemset could generate interesting rules in which $murderRate$ would be high in districts which both contain and touch slum, such information is still expressed by the frequent sets $\{contains\_slum, murderRate = high\}$ and $\{touches\_slum, murderRate = high\}$.

4. Evaluating the Pattern Filtering

In this section we evaluate the proposed method with experiments on real data and present some analysis over the method.

4.1 Analysis

In this section we give a lower bound on the gain of generated frequent itemsets using the algorithm listed in Listing 1 in comparison with the frequent itemsets that would be generated using the standard Apriori method.

Suppose we have a dataset consisting of x elements and that by using a standard Apriori algorithm our largest frequent itemset contains m elements. We know that we then have in total at least

$$\sum_{i=2}^{m} \binom{m}{i}$$

frequent itemsets.

Applied to the results shown in Table 2 where $m = 6$, the formula would give us as lower bound for the number of tuples $= \sum_{i=2}^{6} \binom{6}{i} = \binom{6}{2} + \binom{6}{3} + \binom{6}{4} + \binom{6}{5} + \binom{6}{6} = 15 + 20 + 15 + 6 + 1 = 57$, which is correct because Table 2 contains 60 frequent itemsets.

Suppose that one of the largest frequent itemsets contains u geographic feature types with more than one qualitative spatial relationsip in this largest frequent itemset, and n other attributes.

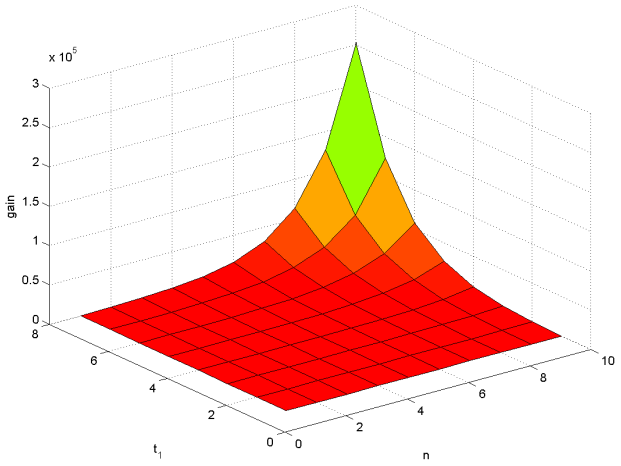

Figure 3. Minimal Gain having only $u = 1$ Geographic Feature Type and having different values for t_1 and n in the Largest Frequent Itemset

Let t_k denote the number of qualitative spatial relationships of object $k \in [1, u]$, then $m = \sum_{k=1}^{u} t_k + n$. If we use the proposed improvement, then the minimum gain of frequent itemsets is.

$$\sum_{i=2}^{m}\{\binom{t_1}{j_1}\dots\binom{t_u}{j_u}\binom{n}{j_{u+1}} \mid \sum_{s=1}^{u+1} j_s = i; j_{u+1} \le n; \forall s \in [1,u] : s \in \mathbb{N}, j_s \le t_s; \exists t \in [1,u] : j_t \ge 2\} \quad (1)$$

Applied to the Table 2 where $m = 6, u = 2, t_1 = 2, t_2 = 2$ and $n = 2$ this gives us a minimal gain $= \binom{2}{2}\binom{2}{0}\binom{2}{0} + \binom{2}{0}\binom{2}{2}\binom{2}{0} + \binom{2}{2}\binom{2}{1}\binom{2}{0} + \binom{2}{2}\binom{2}{0}\binom{2}{1} + \binom{2}{1}\binom{2}{2}\binom{2}{0} + \binom{2}{0}\binom{2}{2}\binom{2}{1} + \binom{2}{2}\binom{2}{2}\binom{2}{0} + \binom{2}{2}\binom{2}{1}\binom{2}{1} + \binom{2}{2}\binom{2}{0}\binom{2}{2} + \binom{2}{1}\binom{2}{2}\binom{2}{1} + \binom{2}{0}\binom{2}{2}\binom{2}{2} + \binom{2}{2}\binom{2}{2}\binom{2}{1} + \binom{2}{2}\binom{2}{1}\binom{2}{2} + \binom{2}{1}\binom{2}{2}\binom{2}{2} + \binom{2}{2}\binom{2}{2}\binom{2}{2} = 1+1+2+2+2+2+1+4+1+4+1+2+2+2+1 = 28$, what is only 2 less than in reality. These are actually the 2-itemset $\{contains\_slum, touches\_slum\}$ and the 3-itemset $\{contains\_slum, touches\_slum, touches\_school\}$ which contain the attribute $touches\_slum$ that is not present in the largest frequent itemset of Table 2.

Suppose we have only 1 geographic feature type with more than one qualitative spatial relationship in the largest frequent itemset, so $u = 1$. In Figure 3 the minimal gain is plotted when $t_1 = 1, 2, \dots, 8$ and $n = 1, 2, \dots, 10$. Although in Figure 3 it seems that for small values of t_1 and n the minimal gain is constant, this is just because of the large minimal gain for the highest values of t_1 and n, which is clear from Table 3.

n \ t_1							
0	2	8	22	52	114	240	494
0	4	16	44	104	228	480	988
0	8	32	88	208	456	960	1976
0	16	64	176	416	912	1920	3952
0	32	128	352	832	1824	3840	7904
0	64	256	704	1664	3648	7680	15808
0	128	512	1408	3328	7296	15360	31616
0	256	1024	2816	6656	14592	30720	63232
0	512	2048	5632	13312	29184	61440	126464
0	1024	4096	11264	26624	58368	122880	252928

Table 3. Minimal Gain having only $u = 1$ Geographic Feature Type and having different values for t_1 and n in the Largest Frequent Itemset

4.2 Experiments

Figure 4 shows the result of an experiment performed with a geographic dataset with one non-spatial attribute and 6 geographic object types, that with different topological relationships generated 13 spatial predicates. Among these predicates, a total of 9 pairs had the same feature type with a different relationship, and four pairs had a geographic dependence.

Different experiments were performed on this dataset, considering minimum support 5%, 10%, and 15%. First we extracted patterns from this dataset using Apriori, which does not eliminate any meaningless pattern. Then Apriori-KC was applied, and pairs of geographic objects with the dependencies specified in Φ were eliminated. At the end we mined the dataset with Apriori-KC+, which eliminates pairs with either well known dependencies or same feature type.

As can be observed in Figure 4, the elimination of 4 geographic dependencies with Apriori-KC reduced the number of frequent sets generated by Apriori in around 28% for different values of minimum support. However, with the elimination of meaningless combinations with the same feature type, Apriori-KC+ reduced this number much further. The elimination of 9 pairs with equal feature type reduced the number of frequent sets in more than 60% in relation to Aprioi and aroundf 50% in relation to Apriori-KC.

The computational time to eliminate both dependencies and frequent sets with same feature types is reduced with Apriori-KC+, as shown in Figure 5.

It is important to consider that for any spatial dataset with qualitative spatial relationships the frequent set reduction is data dependent. For example, a pair of predicates with the same feature type such as $\{contains(Street), crosses(Street)\}$ in cities, will have much higher support than a pair of predicates such as $\{contains(River), crosses(River)\}$, since normally cities have a lot more streets than rivers.

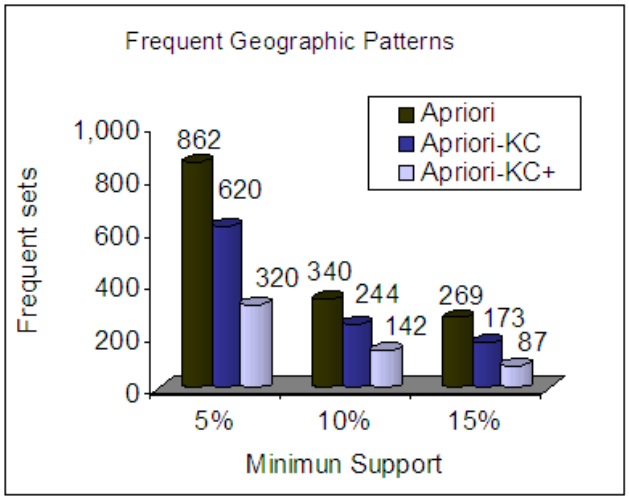

Figure 4. Frequent Geographic Patterns (Apriori), Frequent Geographic Patterns without Dependences (Apriori-KC), and Frequent Geographic Patterns without both Dependences and Same Feature Type (Apriori-KC+)

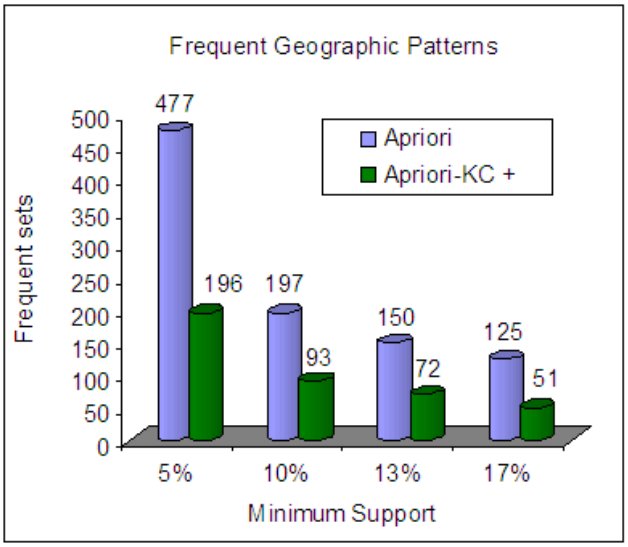

Figure 6. Frequent Geographic Patterns (Apriori) and Frequent Geographic Patterns without Same Feature Type (Apriori-KC+)

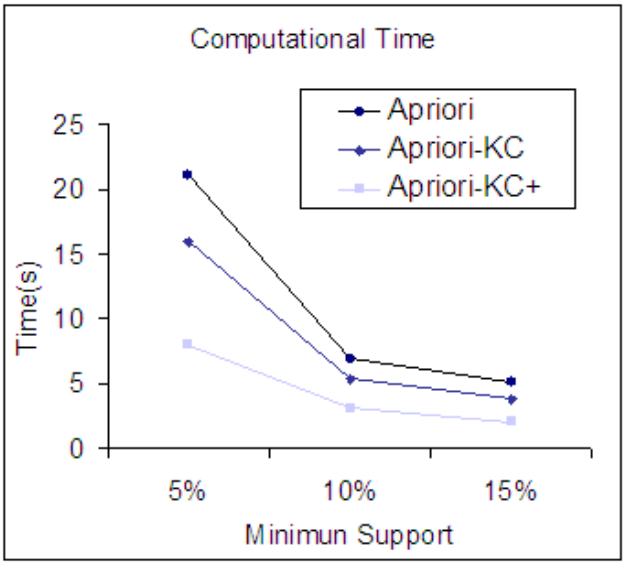

Figure 5. Computational Time to Generate Frequent Geographic Patterns with Apriori, Apriori-KC, and Apriori-KC+

A second experiment was performed with a dataset generated with 10 spatial predicates. Among these predicates, five pairs had the same feature type. In this dataset no geographic objects with dependencies were considered, and only predicates with the same feature type were removed. Figure 6 shows the frequent set reduction with Apriori-KC+ in relation to Apriori considering different values of minimum support.

In this experiment the number of frequent sets is reduced in more than 55% for any value of minimum support. In Figure 7 we can observe that the computational time is reduced as well, which shows that Apriori-KC+ is more effective than Apriori and more efficient when applied to geographic data. In both experiments we can observe that the higher the number of either dependencies or meaningless combinations, the more efficient is Apriori-KC+.

If we apply our formula for the gain (Formula 1) to the largest frequent itemset with a minimum support of 5% using the experiment shown in Figure 6, where $m = 8, u = 3$ and $t_1 = t_2 = t_3 = n = 2$, the formula predicts a minimum gain of 148 frequent itemsets, where the real gain is 281. If we do the same for largest frequent itemset using a minimum support of 17% where $m = 7, u = 3, t_1 = t_2 = t_3 = 2$ and $n = 1$, we predict a gain of 74 which is equal to the real gain.

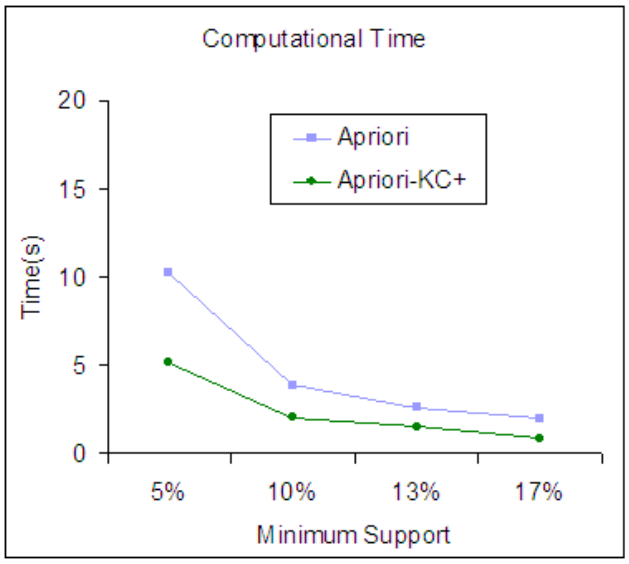

Figure 7. Computational Time to Extract Frequent Geographic Patterns with Apriori and Apriori-KC+

5. Conclusions and Future Works

In frequent geographic pattern mining a large amount of patterns is well known apriori. Besides these well known patterns, another amount is non-interesting. The same geographic object type can have different qualitative spatial relationships with the reference geographic object type, and as a consequence, many "items" are the same feature type with a different spatial relationship. Therefore, many combinations having the same feature type generate non-interesting rules in most applications, such as, for example,

$$contains\_street \rightarrow touches\_street.$$

Trying to reduce the number of well known patterns and non-interesting association rules this paper presented Apriori-KC+, which is an extension of Apriori-KC to avoid the combination or pairs of spatial predicates which contain the same feature type. In summary, Apriori-KC+ eliminates frequent patterns which contain the same feature type apart from the well known geographic dependences when background knowledge is given as an input.

The main strength of the proposed method for eliminating non-interesting patterns is its simplicity. In a single but very effective and efficient step, combinations of redundant features in the same frequent sets are removed exploiting the anti-monotone constraint. Indeed, this step can be implemented by any algorithm that generates frequent itemsets, including transactional rule mining algorithms which propose different measures for rule interestingness.

The proposed method only eliminates what is non-interesting and does not sacrifice the result quality. It removes frequent patterns having the same feature type, and not same feature type with different instances such as

$$contains\_streetX \rightarrow touches\_streetY.$$

It is effective and efficient for feature type granularities.

Future works include the generation of maximal frequent geographic patterns in order to eliminate redundant frequent itemsets as proposed by the closed frequent pattern mining approaches.

6. Acknowledgments

This research has been funded by CAPES and the European Union (FP6-IST-FET program, Project n. FP6-14915, GeoPKDD: Geographic Privacy-Aware Knowledge Discovery and Delivery (`http://www.geopkdd.eu`).

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In Bocca et al. [6], pages 487–499.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In Bocca et al. [6], pages 487–499.

[3] A. Appice, M. Berardi, M. Ceci, and D. Malerba. Mining and filtering multi-level spatial association rules with ares. In M.-S. Hacid, N. V. Murray, Z. W. Ras, and S. Tsumoto, editors, *ISMIS*, volume 3488 of *Lecture Notes in Computer Science*, pages 342–353. Springer, 2005.

[4] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal. Mining minimal non-redundant association rules using frequent closed itemsets. In *CL '00: Proceedings of the First International Conference on Computational Logic*, pages 972–986, London, UK, 2000. Springer-Verlag.

[5] R. J. Bayardo and R. Agrawal. Mining the most interesting rules. In *KDD*, pages 145–154, 1999.

[6] J. B. Bocca, M. Jarke, and C. Zaniolo, editors. *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*. Morgan Kaufmann, 1994.

[7] V. Bogorny, S. Camargo, P. Engel, and L. O. Alvares. Mining frequent geographic patterns with knowledge constraints. In *ACM-GIS*, pages –. ACM, 2006.

[8] V. Bogorny, S. Camargo, P. Engel, and L. O. Alvares. Towards elimination of well known patterns in spatial association rule mining. In *IS*, pages 532–537. IEEE Computer Society, 2006.

[9] V. Bogorny, J. Valiati, S. Camargo, P. Engel, B. Kuijpers, and L. O. Alvares. Mining maximal generalized frequent geographic patterns with knowledge constraints. In *ICDM*, pages –. IEEE Computer Society, 2006.

[10] M. J. Egenhofer and R. D. Franzosa. On the equivalence of topological relations. *International Journal of Geographical Information Systems*, 9(2):133–152, 1995.

[11] R. H. Güting. An introduction to spatial database systems. *VLDB J.*, 3(4):357–399, 1994.

[12] J. Han. Mining knowledge at multiple concept levels. In *CIKM*, pages 19–24. ACM, 1995.

[13] Y. Huang, S. Shekhar, and H. Xiong. Discovering colocation patterns from spatial data sets: A general approach. *IEEE Trans. Knowl. Data Eng.*, 16(12):1472–1485, 2004.

[14] K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *SSD '95: Proceedings of the 4th International Symposium on Advances in Spatial Databases*, pages 47–66, London, UK, 1995. Springer-Verlag.

[15] S. Shekhar and S. Chawla. *Spatial Databases: A Tour*. Prentice Hall, June 2002.

[16] A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974, 1996.

[17] P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 32–41, New York, NY, USA, 2002. ACM Press.

[18] G. I. Webb. Discovering significant rules. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 434–443, New York, NY, USA, 2006. ACM Press.

[19] M. J. Zaki. Generating non-redundant association rules. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 34–43, New York, NY, USA, 2000. ACM Press.