A Mining Maximal Generalized Frequent Geographic Patterns With Knowledge Constraints

Mining Maximal Generalized Frequent Geographic Patterns with Knowledge Constraints

Vania Bogorny[1], João Valiati[1], Sandro Camargo[1], Paulo Engel[1], Bart Kuijpers[2], Luis O. Alvares[1]
[1]*Instituto de Informática - Universidade Federal do Rio Grande do Sul (UFRGS)*
Av. Bento Gonçalves, 9500 – Porto Alegre – Brazil
{vbogorny, jvaliati, scamargo, engel, alvares} @ inf.ufrgs.br
[2] *Hasselt University and Transnational University of Limburg*
B-3590 Diepenbeek - Belgium
bart.kuijpers@ uhasselt.be

Abstract

In frequent geographic pattern mining a large amount of patterns is well known a priori. This paper presents a novel approach for mining frequent geographic patterns without associations that are previously known as non-interesting. Geographic dependences are eliminated during the frequent set generation using prior knowledge. After the dependence elimination maximal generalized frequent sets are computed to remove redundant frequent sets. Experimental results show a significant reduction of both the number of frequent sets and the computational time for mining maximal frequent geographic patterns.

1. Introduction

The frequent pattern mining (FPM) technique often generates a large number of frequent itemsets and rules among which a small number is novel and interesting to the user. To overcome this problem an enormous amount of algorithms have been proposed for transactional databases, but no prior knowledge has been used to reduce non-interesting patterns. In spatial FPM this problem increases due the natural dependences among geographic data, which generate a large amount of patterns that are well known a priori.

In geographic databases, most discovered patterns are strongly related to geographic dependences which represent strong regularities, but do not contribute to the discovery of novel and useful knowledge. Users of some domains may not be interested in strong geographic domain rules such as *is_a(GasStation)® intersect(Street)* (100%), but in non-obvious rules such as *is_a(GasStation) and intersects(WaterResource) ® pollution=high* (40%).

Aiming to reduce the number of well known patterns and redundant frequent sets we propose a novel method for mining frequent geographic patterns. We propose to eliminate dependences in a first step and all redundant frequent sets in a second step, computing *maximal generalized frequent geographic patterns* (MGFGP).

The remainder of the paper is organized as follows: Section 2 presents the related works and the main contribution of this paper. Section 3 describes the problem of mining frequent geographic patterns with well known dependences. Section 4 presents the algorithm MG-FGP and shows experiments performed over real geographic databases. Section 5 concludes the paper and gives directions of future work.

2. Related Works and Contribution

There are basically two approaches in the literature for extracting frequent patterns from geographic databases. Both follow the Apriori [1] approach and do not apply the closed frequent set technique. One is based on quantitative reasoning, which mainly computes distance relationships during the frequent set generation. Algorithms based on this approach [2] deal with geographic attributes directly, and have some general drawbacks: input is restricted to points, compute only quantitative spatial relationships, and do not consider non-spatial attributes of geographic data, which may be of fundamental importance for knowledge discovery.

The other approach is based on qualitative reasoning [3]-[4] and usually considers both distance and topological relationships between a reference geographic object type and a set of relevant spatial feature types represented by any geometric primitive (e.g. points, lines, polygons). Because of the high computational cost, spatial relationships are normally extracted in a first step, and frequent patterns are computed in another step.

Both qualitative and quantitative reasoning approaches have not focused on interesting geographic aspects to be considered in FPM. In [4] prior knowledge is used to reduce well known patterns, but a posteriori, after both frequent sets and association rules have already been generated.

In [5] we proposed Apriori-KC to eliminate well known geographic patterns during the frequent set generation. In [6] we proposed to reduce well known patterns by pruning the input space as much as possible, and remaining dependences are eliminated by pruning the frequent sets. However, a large number of frequent sets is still generated in [5] and [6]. As a continued study on frequent geographic pattern mining without well known dependences, in this paper we apply the closed frequent set [7]-[8] technique for mining frequent geographic patterns without redundant frequent sets. We demonstrate that the closed frequent set approach when applied to the geographic domain does not warrant the elimination of well known

dependences. Therefore, we propose to eliminate well known dependences and generate maximal non-redundant frequent sets.

The main advantage of our method is the simplicity as well known dependences are eliminated. While most approaches define syntactic constraints and different thresholds to reduce the number of patterns and association rules, we consider *semantic knowledge constraints,* and eliminate the exact pairs of geographic objects that produce well known patterns.

3. The General Problem of Mining FGP with Well Known Dependences

Most approaches for FPM in transactional databases generate frequent sets as in Apriori [1], or closed frequent sets [7]-[8]. While in transactional data mining the main problem relies on the candidate generation, in geographic data mining it relies on the spatial neighborhood computation [9, page.205]. The number of predicates in geographic FPM is much smaller than the number of items in transactional databases.

3.1 Geographic Dependences in Frequent Pattern Mining

In *transactional* FPM every row in the dataset is usually a transaction and columns are items. In *geographic* FPM every row is an instance (e.g. Porto Alegre) of a reference object type (e.g. city), called *target feature type*, and columns are predicates. Every predicate is related to a non-spatial attribute (e.g. population) of the target feature type or a spatial predicate. Spatial predicate is a *relevant feature type* that is spatially related to specific instances of the target feature type (e.g. *contains_factory*). In geographic FPM the set $F = \{f_1, f_2, \ldots, f_k, \ldots, f_n\}$ is a set of non-spatial attributes and spatial predicates, and Ψ (dataset) is a set of instances of a reference feature type, where each instance is a row R such that $R \mid F$. There is exactly one tuple in Ψ for each instance of the reference feature type.

The problem of geographic FPM is decomposed in two main steps: (a) extract spatial predicates - a spatial predicate is a spatial relationship (e.g. distance) between the target feature type and a set of relevant feature types; and (b) find all frequent predicates - a set of predicates is frequent if its support is at least equal to minimum support.

Assertion 1. [1] If a predicate set Z is large, then every subset of Z will also be large. If the set Z is not large, then every set that contains Z is not large too.

Since the focus of this paper relies on the frequent pattern reduction (*step b*), let us consider the example shown in Figure 1. Figure 1(a) shows a dataset with six tuples and five predicates. Every row in the dataset is a city and the predicate sets are relevant feature types (port, school, water resource, hospital, treated water network) with spatial relationships with the target feature type (city) described in Figure 1(c). In Figure 1(b) are the k frequent sets with minimum support 50% in the dataset in Figure 1(a).

a) dataset

Tid (city)	Predicate Set
1	A, C, D,T, W
2	C, D, W
3	A, D, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

b) frequent predicate sets with minsup 50%

Set k	Frequent sets
k=1	{A}, {C}, {D}, {T}, {W}
k=2	{A,C}, {A,D}, {A,T}, {A,W}, {C,D}, {C,T}, {C,W}, {D,T}, {D,W}, {T,W}
k=3	{A,C,D}, {A,C,W}, {A,D,T}, {A,D,W}, {A,T,W}, {C,D,T}, {C,D,W}, {D,T,W}
k=4	{A,C,D,W}, {A,D,T,W}

c) predicates

A = contains(Port), C = contains(School), W = contains(WaterResource), T = contains(Hospital), D = contains(TreatedWaterNetwork).

Figure 1. Dataset with 6 tuples and frequent predicate sets with minimum support 50%

From this point we will refer to a row or a tuple in the dataset as a "transaction" (tid) and a set of rows as a "set of transactions" (tidset) to follow the terminology commonly used in the frequent pattern mining literature.

The dataset in Figure 1(a) has a geographic dependence between *A* (Port) and *W* (Water Resource), since every port must be related to at least one water resource. Because of their natural dependence, rules such as *contains(Port)→contains(WaterResource)* will be generated. Such a rule expresses that cities that contain ports do also contain water resources. Notice that this kind of rule has no *cause→effect*, once cities do not contain ports because they contain water resources, but cities can only have ports when they also have water bodies.

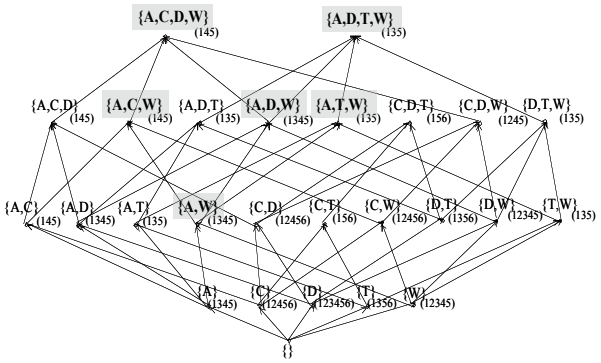

Figure 2. Frequent predicate sets (corresponding tidsets are shown in brackets)

Notice that we cannot eliminate *A* and *W* from the dataset because either *A* or *W* may have an interesting association with any other predicate (*C, D,* or *T*). Figure 2 shows the meet-semilattice [10] of the 25 frequent sets, among which 6 have the dependence {A,W}.

As can be observed in Figure 2, geographic dependences appear the first time in frequent sets with 2 elements, and are replicated to many larger frequent sets when *minsup* is reached. If we remove all frequent sets

in which *A* and *W* appear together, no information is lost [5]-[6]. However, many redundant [11] frequent sets that generate redundant rules with same support and confidence, still remain among the resultant frequent itemsets (e.g. {A,C},{A,T},{C,T},{T,W}). These sets are eliminated by the *closed frequent set* approach, introduced in the following section.

3.2 Geographic Dependences in Closed Frequent Pattern Mining

According to [7]-[8], a frequent predicate set L is a closed frequent predicate set if $\Omega(L)=L$. The closure operator Ω associates with a frequent predicate set L the maximal set of predicates common to all transactions (tidset) containing L. The closure operator allows the definition of all closed frequent itemsets which constitute the minimal non-redundant frequent sets.

Figure 3 shows the closed frequent sets meet-semilattice of the frequent sets presented in the previous section. The set {A,D,W}, for example, is a *frequent itemset* because it reaches minimum support (50%). It is also a *closed frequent itemset* because in the set of transactions (1345) where it occurs in the dataset, no set larger than {A,D,W} in the same transactions reaches minimum support. The frequent itemset {A,D,T}, for example, appears in the transactions 135, but in the same transactions, a larger set {A,D,T,W} can be generated.

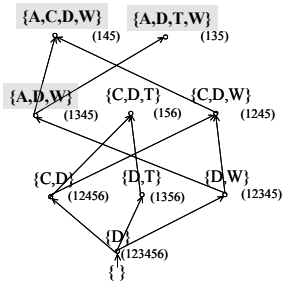

Figure 3. Closed frequent sets meet-semilattice with geographic dependences

As can be observed in Figure 3, the closed frequent set approach does not warrant the elimination of well known geographic dependences, since among the 9 closed frequent sets, 3 sets (in dark boxes) contain both A and W.

By eliminating the closed frequent sets with the geographic dependence {A,W} the information of the sets {A,C}, {A,D}, {A,T}, {T,W}, {D,T,W}, {A,C,D}, and {A,D,T} is lost. This example shows that the closed frequent set technique cannot be applied to geographic data when the objective is to obtain frequent geographic patterns *without* well known dependences.

To reduce the number of frequent sets without well known dependences and without sacrifice the result quality, we propose to generate *maximal non-redundant generalized frequent sets with knowledge constraints*.

Definition 1 (MGFGP – Maximal Generalized Frequent Geographic Patterns without well known geographic dependences): a frequent predicate set or a frequent geographic pattern L is maximal generalized when it has no well known geographic dependence in a set of dependences $\varnothing$ such that $L\text{-}\varnothing=L$ and $M(L)=L$.

The Maximal operator M associates with a frequent predicate set L the maximal set of predicates common to all transactions containing L without well known geographic dependences, i.e., L is maximal if there is no frequent predicate set L' in the same transactions of L such that $L \subset L'$.

Considering the frequent sets generated from transactions 135 without well known dependences ({T,W}, {A,T}, {A,D,T}, and {D,T,W}) shown in the meet-semilattice in Figure 4, notice that $\{T,W\} \subset \{D,T,W\}$ and $\{A,T\} \subset \{A,D,T\}$. So neither {T,W} nor {A,T} are maximal. However, $\{A,D,T\} \not\subset \{D,T,W\}$, so both {A,D,T} and {D,T,W} are maximal. In transactions 135, while only one frequent set is closed ({A,D,T,W}), but having a geographic dependence, two frequent sets ({A,D,T} and {D,T,W}) are *maximal,* but *without* well known geographic dependences.

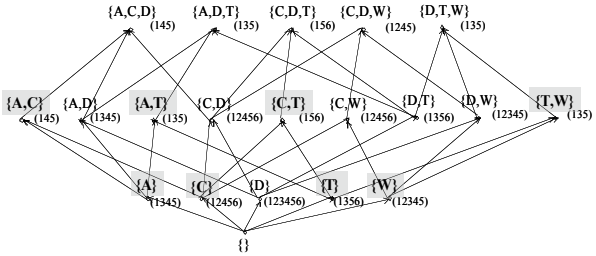

Figure 4. Maximal frequent sets meet-semilattice without well known dependences

The elimination of geographic dependences from the frequent sets (Figure 2) in addition to the elimination of redundant frequent sets (Figure 4), generates a reduced number of maximal frequent sets, as shown in Figure 5. Notice that no information is lost and the result quality is not sacrificed.

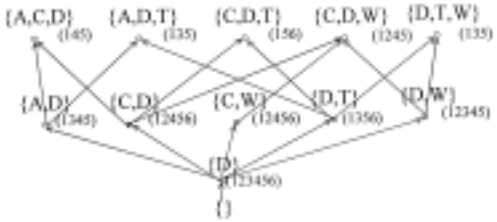

Figure 5. Maximal generalization of frequent sets without well known dependences

In the following section we present the MG-FGP algorithm and evaluate the proposed method with experiments performed over real GDB.

4. Mining MGFGP with Knowledge Constraints

Figure 6 shows the pseudo-code of the algorithm MG-FGP to generate maximal generalized frequent geographic patterns without well known dependences.

Given ϕ as the set of pairs of geographic objects with dependences (e.g. {Island, Water}) called *knowledge constraints*, Ψ as the input dataset, and *minsup* as minimum support, well known geographic dependences are removed from the candidate sets with two elements, before even compute their frequency. MG-FGP removes from the candidate sets all pairs of predicates which have geographic dependences. As in Apriori, MG-FGP performs multiple passes over the dataset. In the first pass, the support of the individual elements is computed to determine *k*-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*.

```
Given: ϕ, // set of well known dependences
       Ψ, // spatial dataset
       minsup; // minimum support

Find: Maximal generalized frequent geographic patterns
      without well known dependences

Method:
L1 = {large 1-predicate sets};
For ( k = 2; Lk-1 != →; k++ ) do begin
  Ck = apriori_gen(Lk-1); // Generates new candidates
  If (k=2)
      C2 = C2 - ϕ; // Remove the pairs with dependences
  Forall rows w ⌊ Ψ do begin
    Cw = subset(Ck, w); // Candidates contained in w
    Forall candidates c ⌊ Cw do c.count++;
  End;
  Lk = {c ⌊ Ck | c.count _ minsup};
End;
L = ⌈ k Lk;

// find maximal generalized predicate sets
G = L;
For ( k = 2; Gk != →; k++ ) do begin
   For ( j = k+1; Gj!=0; j++ ) do
      If (tidSet (Gk) = tidSet (Gj))
          If (Gk ⊂ Gj)  // Gj is more general than Gk
              Delete Gk from G;
End;
Answer = G;
```

Figure 6. Pseudo-code of MG-FGP

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 number of frequent sets is zero.

The dependences are eliminated in the second pass, when generating candidates with 2 predicates. According to Assertion1 this step *warrants* that the pairs of geographic objects with a well known dependence specified in ϕ will neither appear together in the frequent sets nor in the spatial association rules.

Once the frequent sets without dependences have been generated, MG-FGP starts the generalization, similarly to the closed frequent set approach. Given *G* as all frequent sets *L,* all frequent sets in *G* with size *k* are compared to the sets with size *k*+1. When the set of transactions (tidset) in which G_k appears is the same set where G_{k+1} appears, and the set $G_k \subset G_{k+1}$, then we can say that G_k is redundant, while G_{k+1} is more general. When this occurs, G_k is removed from *G*. This process continues until all frequent sets in *G* are maximal.

Notice that we compute the maximal frequent sets *G* and do not lose *L,* that contains all frequent sets. Since *L* and *G* have been computed, spatial association rules can be generated by any of the different methods proposed in the literature to generate non-redundant association rules.

In order to evaluate our method, experiments were performed with real geographic databases. Districts of the city of Porto Alegre (Brazil) were the target feature type and several relevant spatial feature types were considered, including water bodies, hospitals, slums, streets, gas stations, industrial residues repositories, bridges, etc. The dependences between gas stations and streets as well as bridge and water body were eliminated. Experiments with minimum support 5%, 10%, and 15% were performed.

Figure 7 shows the number of frequent sets generated by Apriori, MG-FGP, and the closed frequent set approach. For the different values of minimum support the closed frequent set approach reduced the number of frequent sets generated by Apriori in an average of 87%. However, among the 123, 82, and 49 closed frequent sets generated for *minsup* 5%, 10%, and 15%, respectively, 66, 39, and 24 contain well known geographic dependences.

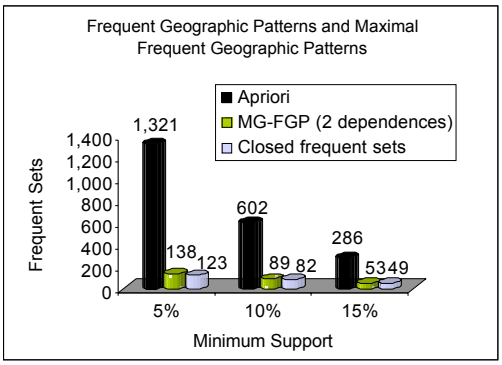

Figure 7. Frequent sets, closed frequent sets, and maximal generalized frequent sets

Considering the different values of minimum support the closed frequent set approach eliminates a large number of frequent sets, but more than 50% of these sets contain well known geographic patterns.

MG-FGP reduces the frequent sets generated by Apriori in more than 80% for any value of *minsup*, and no dependences exist within these sets.

Besides the significant reduction of the number of frequent sets, MG-FGP is more efficient than the closed frequent set approach. Figure 8 shows the computational time for extracting frequent sets with Apriori and the closed frequent set approach, which both do not eliminate geographic dependences, as well as MG-FGP which eliminates geographic dependences.

The computational cost to generate closed frequent sets or maximal generalized frequent sets is higher than to simply generate frequent sets as in Apriori. The

additional verification to generate either maximal of closed frequent sets requires extra scans over the dataset. However, Figure 8 shows that the closed frequent set approach, apart from not eliminating well known geographic dependences requires more computational time than MG-FGP.

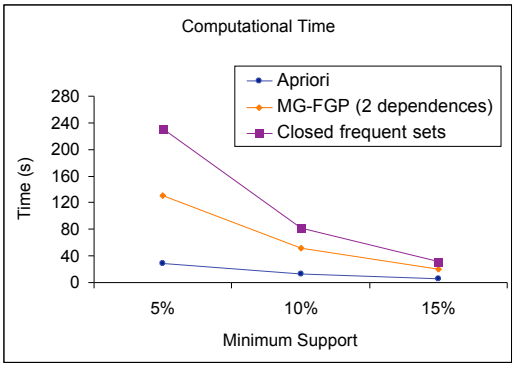

Figure 8. Computational time to generate frequent sets, closed frequent sets, and maximal frequent sets

MG-FGP tends to reduce computational time for large databases when the number of dependences increases, since less frequent sets will be generated.

5. Conclusions and Future Work

In frequent geographic pattern mining a large amount of patterns is well known. Examples using the meet-semilattice of frequent sets, as well as experiments with real geographic databases, showed that the closed frequent set approach, when applied to the geographic domain, generates many closed frequent sets containing well known dependences. Indeed, if geographic dependences are removed from closed frequent sets, the result quality is sacrificed.

To eliminate well known geographic domain patterns in geographic FPM we proposed an efficient method, which eliminates pairs of geographic dependences in one single step. Indeed, we eliminate redundant frequent sets and compute the maximal generalized frequent sets to avoid the generation of redundant association rules.

In this paper we presented a solution to generate maximal generalized FGP without well known dependences considering geographic data at a high granularity level (e.g. water). However, the dependence replication process increases when mining data at lower granularities (e.g. river, lake, sea). As future work we will evaluate our method considering hierarchical geographic dependences when mining maximal geographic patterns from data at different granularities.

ACKNOWLEDGMENT

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

References

[1] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th International Conference on Very Large Databases (VLDB'94)*, 1994, pp. 487-499.

[2] J. S. Yoo, S. Shekhar, and M. Celik, "A Join-less Approach for Co-location Pattern Mining: A Summary of Results," in *Proc. 5th IEEE International Conference on Data Mining (ICDM'05)*, 2005, pp. 813-816.

[3] K. Koperski and J. Han, "Discovery of spatial association rules in geographic information databases," in *Proc. 4th International Symposium on Advances in Spatial Databases (SSD'95)*, 1995, pp. 47-66.

[4] A. Appice, M. Berardi, M. Ceci, and D. Malerba, "Mining and Filtering Multi-level Spatial Association Rules with ARES," in *Proc. 15th International Symposium Foundations of Intelligent Systems (ISMIS'05)*, 2005, pp. 342-353.

[5] V. Bogorny, S. Camargo, P M. Engel, L.O. Álvares, "Towards elimination of well known geographic domain patterns in spatial association rule mining," in *Proc. 3th IEEE International Conference on Intelligent Systems (IEEE-IS'06)*, 2006, pp. 532-537.

[6] V. Bogorny, S. Camargo, P. Engel, and L.O. Alvares, "Mining frequent geographic patterns with knowledge constraints," in *Fourteenth ACM International Symposium on Advances in Geographic Information Systems*. 2006. (to be published).

[7] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering frequent closed itemsets for association rules," in *Proc. 7th International Conference on Database Theory (ICDT'99)*, 1999, pp. 398-416.

[8] M. Zaki and C. Hsiao, "CHARM: An efficient algorithm for closed itemset mining," in *Proc. 2nd SIAM International Conference on Data Mining*, 2002, pp. 457-473.

[9] S. Shekhar, and S. Chawla, *Spatial databases: a tour*. Upper Saddle River, NJ: Prentice Hall, 2003.

[10] B. A. Davey and H. Priestley. *Introduction to Lattices and Order*. Cambridge: Cambridge University Press, 2002.

[11] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. L minimal non-redundant association rules using frequent closed itemsets," in *Proc. 1st International Conference on Computational Logic*, 2000, pp. 972-986.