Mining Volunteered Geographic Information datasets with heterogeneous spatial reference

Sadiq Hussain
System Administrator, Examination Branch
Dibrugarh University, Dibrugarh, India

Prof. G.C. Hazarika
Director i/c, Centre for Computer Studies
Dibrugarh University, Dibrugarh, India

***Abstract*—When the information created online by users has a spatial reference, it is known as Volunteered Geographic Information (VGI). The increased availability of spatiotemporal data collected from satellite imagery and other remote sensors provides opportunities for enhanced analysis of Spatiotemporal Patterns. This area can be defined as efficiently discovering interesting patterns from large data sets. The discovery of hidden periodic patterns in spatiotemporal data could provide unveiling important information to the data analyst. In many applications that track and analyze spatiotemporal data, movements obey periodic patterns; the objects follow the same routes (approximately) over regular time intervals. However, these methods cannot directly be applied to a spatiotemporal sequence because of the fuzziness of spatial locations in the sequence. In this paper, we define the problem of mining VGI datasets with our already established bottom up algorithm for spatiotemporal data.**

Keywords- data mining; periodic patterns; spatiotemporal data; Volunteered Geographic Information.

I. INTRODUCTION

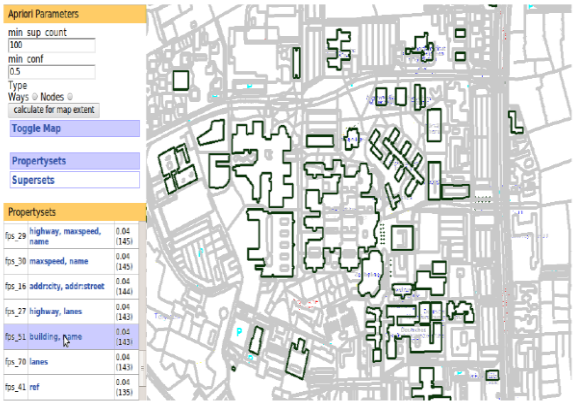
There is an explosion of geographic information generated by individuals on the Web. Users provide geotagged photos and tweets, geotag Wikipedia articles, create gazetteer entries, update geographic databases like OpenStreetMap (OSM) and much more. Such user-generated geodata, also called Volunteered Geographic Information, VGI [1], is becoming an important source for geo-services like map generation, routing, search, spatial analysis and mashups. Different from traditional geodata, VGI often has no distinct classifying attributes or explicit taxonomy. Users are free to create new tagging schemas or add new properties or text. Although some schema checks may exist on the editor level through auto-completion or templates, these checks are not strict and can be ignored by the user. Analyzing the dynamic and heterogeneous schemas of VGI to find common conceptualizations is an important and complex task. For example, Deng et al. [2] use density based clustering and a document term matrix to find conceptualizations in geotagged Flickr images. Edwardes and Purves [3] explore the potential to develop a hierarchy of place concepts based on co-occurring characteristic terms in the description of geotagged photos of the British Isles. Extracting and exploring concepts is an important prerequisite to analyze the quality and consistency of a dataset and to evaluate its “fitness for use” [4]. We describe our work on using frequent pattern mining to extract and explore conceptualizations of VGI. Frequent pattern mining is used for effective classification in association rule mining [5]. Afrati et al. [1] use frequent sets to find approximate patterns, which is a promising technique for concept extraction and exploration. For geospatial data, frequent pattern mining is used to determine spatial association rules [6] and to perform co-occurrence analysis [5]. In our approach we transform VGI into a flat model of transaction objects, which can be input to our mining algorithms. Different from transactions of market basket data, which are the typical input to frequent pattern mining, geospatial patterns may occur rarely in a dataset but are nevertheless interesting. Ding et al. (2006) introduce a framework to mine regional association rules based on prior clustering to find patterns in sub regions. However, to extract concepts, mining sub regions is not an option. We explain what extensions to frequent pattern mining are needed to deal with the scale-dependency and introduce a bottom-up mining approach based on quadtrees. We developed a prototype framework to mine the frequent patterns apriori, which then can be efficiently accessed by clients. For this, we describe the OSM Explorer, which visualizes frequent patterns in the OSM dataset and performs data consistency and quality checks.

II. TRANSACTION MODEL

To employ frequent pattern mining to extract concepts from VGI, the heterogeneous geographic information needs to be transformed into transactions. A transaction has an associated set of items and is input record frequent pattern mining. We view each geoobject as a transaction having geometry and a set of attributes. Attributes can be key-value pairs (representing an attribute name and value) or just keys (like tags). Text has to be itemized first. For example, by using frequency term vectors or by extracting named entities, a text describing geographic information can be transformed into a set of attributes. In general, a geoobject is represented as a transaction as follows:

Transaction (ObjID, Geometry, List((Key, [Value])))

By determining frequent itemsets from such transactions one obtains frequent patterns of attribute names (if key is the name of a property), tags (if key is a tagname) or words (if key is the word of a frequency term vector). These frequent patterns cannot yet be seen as concepts, but they are good candidates for building concept hierarchies and classification models in a subsequent step. The result of some frequent patterns in the OSM data, which can be interpreted as collaborative generated schemas for geographic concepts, is illustrated in Figure 1. The above process is discussed in more detail in [7].



Apriori Parameters

min sup count: 100

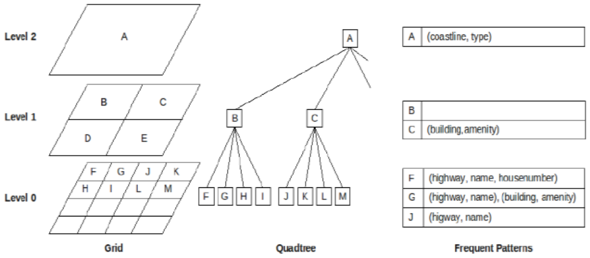
min conf: 0.5

Type: Ways ○ Nodes ○

Propertysets

fps_29	highway, maxspeed, name	0.04 (145)
fps_30	maxspeed, name	0.04 (145)
fps_16	addr:city, addr:street	0.04 (144)
fps_27	highway, lanes	0.04 (143)
fps_51	building, name	0.04 (143)
fps_70	lanes	0.04 (143)
fps_41	ref	0.04 (135)

Figure 1. OSM Explorer: Visualization of the pattern (building, amenity).



A	(coastline, type)
B	
C	(building,amenity)
F	(highway, name, housenumber)
G	(highway, name), (building, amenity)
J	(higway, name)

Figure 2. Scale-dependent frequent pattern mining.

III. Scale-Dependent Mining

A pattern is called frequent if it has a minimum support, that is, it occurs a minimum number of times in a given dataset. Patterns that only occur rarely are not considered. However, geospatial patterns occurring with a low frequency in a dataset can still be interesting for concept extraction. This is either because they are 1) densely clustered (and thus may represent a local/regional pattern on a large scale) or 2) they are widely distributed (and thus may represent a pattern on a small scale). We use a bottom-up approach based on a quadtree data structure to determine which items are candidates for itemset generation on a certain scale, as shown in Figure 2.

The items of every transaction in each leaf node of the quadtree (which constitutes a grid of cells over the input data space) are counted. The items that occur in a cell with at least a minimum frequency are used to generate frequent itemsets over the transactions within this cell. On the next higher level all items that have not been used so far are summarized. If they reach the minimum frequency they are input to frequent itemset mining at this level. This step is repeated until the root node is reached. The determined itemsets are linked to the according nodes in the quadtree, which then also allows for fast exploration, for example, via a map interface.

IV. Periodic Patterns in Object Trajectories

This section defines the problem of mining periodic patterns in spatiotemporal data. First, we motivate our research by discussing why previous work on event sequences is not expected to perform well when applied on object trajectories. We then proceed to a formal definition of the problem.

In our model, we assume that the locations of objects are sampled over a long history. In other words, the movement of an object is tracked as an n-length sequence S of spatial locations, one for each timestamp in the history, of the form $\{(l_0, t_0), (l_1, t_1), . . . , (l_{n-1}, t_{n-1})\}$, where li is the object's location at time ti. If the difference between consecutive timestamps is fixed (locations are sampled every regular time interval), we can represent the movement by a simple sequence of locations l_i (i.e., by dropping the timestamps t_i, since they can be implied). Each location l_i is expressed in terms of spatial coordinates. Figure 3a, for example, illustrates the movement of an object in three consecutive days (assuming that it is tracked only during specific hours, e.g., (working hours). We can model it with sequence S = {<4, 9>, <3.5, 8>, . . . ,<6.5, 3.9>, <4.1, 9>, . . . }. Given such a sequence, a minimum support min_sup ($0 < min\_sup \leq 1$), and an integer T, called period, our problem is to discover movement patterns that repeat themselves every T timestamps. A discovered pattern P is a T-length sequence of the form $r_0 r_1 . . . r_{T-1}$, where ri is a spatial region or the special character *, indicating the whole spatial universe. For instance, pattern AB*C** implies that at the beginning of the cycle the object is in region A, at the next timestamp it is found in region B, then it moves irregularly (it can be anywhere), then it goes to region C, and after that it can go anywhere, until the beginning of the next cycle, when it can be found again in region A. The patterns are required to be followed by the object in at least α (α = min_sup · [n/T]) periodic intervals in S.

Problem definition

Let S be a sequence of n spatial locations $\{l_0, l_1, . . . , l_{n-1}\}$, representing the movement of an object over a long history. Let T << n be a user specified integer called period (e.g., day, week, month). A periodic segment s is defined by a subsequence $l_i\, l_{i+1} . . . l_{i+T-1}$ of S, such that i modulo T = 0. Thus, segments start at positions 0, T, . . . , ([n/T] − 1) · T, and there are exactly m = [n/T] periodic segments in S^1. Let s^j denote the segment starting at position lj·T of S, for $0 \leq j < m$, and let $s_i{}^j = l_j \cdot T + i$, for $0 \leq i < T$.

Definition The mining periodic patterns problem searches for all valid periodic patterns *P* in *S*, which are frequent and non-redundant with respect to a minimum support min_ sup. For simplicity, we will use "frequent pattern" to refer to a valid, non-redundant frequent pattern.

V. Mining Periodic Patterns

In this section, we present techniques for mining frequent periodic patterns and their associated regions in a long history of object trajectories. We first address the problem of finding frequent 1-patterns (i.e., of length 1). Then, we propose one method to find longer patterns; a bottom-up, level-wise technique.

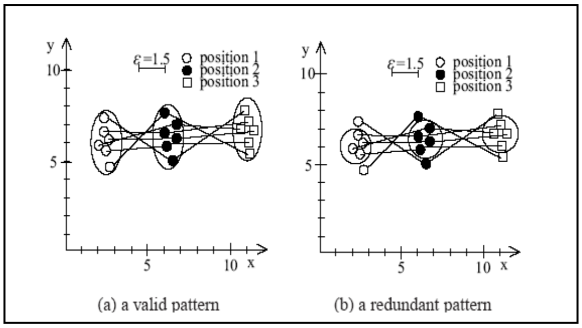


Figure 3: Redundancy of patterns

A. *Obtaining frequent 1-patterns*

Including automatic discovery of regions in the mining task does not allow for the direct application of techniques that find patterns in sequences (e.g., [8]), as discussed. In order to tackle this problem, we propose the following methodology. We divide the sequence S of locations into T spatial datasets, one for each offset of the period T. In other words, locations $\{l_i, l_{i+T}, \ldots, l_{i+(m-1)\cdot T}\}$go to set R_i, for each $0 \leq i < T$. Each location is tagged by the id $j \in [0, \ldots, m - 1]$ of the segment that contains it. Figure 4a shows the spatial datasets obtained after decomposing the object trajectory. We use a different symbol to denote locations that correspond to different periodic offsets and different colors for different segment-ids. Observe that a dense cluster r in dataset R_i corresponds to a frequent pattern, having $\star$ at all positions and r at position i. Figure 4b shows examples of five clusters discovered in datasets R_1, R_2, R_3, R_4, and R_6. These correspond to five 1-patterns (i.e., $r_{11}\star\star\star\star\star$, $\star r_{21}\star\star\star\star$, etc.). In order to identify the dense clusters for each R_i, we can apply a density-based clustering algorithm like DBSCAN [9]. Clusters with less than α ($\alpha = min\_sup \cdot m$) points are discarded, since they are not frequent 1-patterns according to our definition. Clustering is quite expensive and it is a frequently used module of the mining algorithms, as we will see later. DBSCAN [9] has quadratic cost to the number of clustered points, unless an index (e.g., R–tree) is available. Since R–trees are not available for every arbitrary set of points to be clustered, we use an efficient hash-based method.

B. *A level-wise, bottom-up approach*

Starting from the discovered 1-patterns (i.e., clusters for each R_i), we can apply a variant of the level-wise Apriori-TID algorithm [10] to discover longer ones. The input of our algorithm is a collection L_1 of frequent 1-patterns, discovered as described in the previous paragraph; for each R_i, $0 \leq i < T$, and each dense region $r \in R_i$, there is a 1-pattern in L_1.Pairs $<P_1, P_2>$ of $(k - 1)$-patterns in L_{k-1}, with their first $k - 2$ non-$\star$ regions in the same position and different $(k - 1)$-th non-$\star$ position create candidate k-patterns. For each candidate pattern P_{cand}, we then perform a segment-id join between P_1 and P_2, and if the number of segments that comply with both patterns is at least $min\_sup \cdot m$, we run a pattern validation function to check whether the regions of P_{cand} are still clusters. After the patterns of length k have been discovered, we find the patterns at the next level, until there are no more patterns at the current level, or there are no more levels.

C. *Algorithm Level-wise Pattern Mining (L1, T, min sup);*

```
k:=2;
while $(\mathcal{L}_{k-1} \neq \emptyset \wedge k < T)$
$\mathcal{L}_k$:=$\emptyset$;
Generation of quadtree
Dynamic Time Warping
for each pair of patterns $(P_1, P_2) \in L_{k-1}$
such that $P_1$ and $P_2$ agree on the first $k - 2$
and have different $(k - 1)$-th non-$\star$ position
$P_{cand}$:=candidate gen$(P_1, P_2)$;
if $(P_{cand} \neq null)$ then
$P_{cand}$:=$P_1 \bowtie_{P_1.sid=P_2.sid} P_2$; //segment-id join
if $(|P_{cand}| \geq min\_sup \cdot m)$ then
    validate_pattern$(P_{cand}, \mathcal{L}_k, min\_sup)$;
        $k$:=$k + 1$;
    return $\mathcal{P}$:=$\bigcup \mathcal{L}_k, \forall 1 \leq k < T$;

function validate_pattern$(P_{cand}, \mathcal{L}_k, min\_sup)$;
1).  $split$:=false; $prev\_size$:=$|P_{cand}|$
2).  for each non-$\star$ position $i$ of $P_{cand}$
3).      cluster points of $R_i$ with $sid \in P_{cand}$;
4).      if (more than one clusters with size $\geq min\_sup \cdot m$) then
5).          $split$:=true;
6).          for each cluster $r$ with size $\geq min\_sup \cdot m$
7).              $P'_{new}$:=$\{sid \mid sid \in r\}$;
8).              validate_pattern$(P'_{cand}, \mathcal{L}_k, min\_sup)$;
9).      else $P_{cand}$:=segment-ids in updated cluster $r$;
10). if $(\neg split)$then
11).     if $(|P_{cand}| \geq min\_sup \cdot m)$ then
12).         validate_pattern$(P_{cand}, \mathcal{L}_k, min\_sup)$;
13).     else $\mathcal{L}_k$:=$\mathcal{L}_k \cup P_{cand}$;
```

14) Use Z test

To illustrate the algorithm, consider the 2-patterns $P_1 = r_{1x}r_{2y}\star$ and $P_2 = r_{1w}\star r_{3z}$ of Figure 3a. Assume that *MinPts* = 4 and ε = 1.5. The two patterns have common first non-$\star$position and *Minimum Bounding Rectangles*(r_{1x}) overlaps *Minimum Bounding Rectangles*(r_{1w}). Therefore, a candidate 3-pattern P_{cand} is generated. During candidate pruning, we verify that there is a 2-pattern with non-$\star$ positions 2 and 3 which is in L_2.Indeed, such a pattern can be spotted at the figure (see the dashed lines). After joining the segmentids in P_1 and P_2 , P_{cand} contains the trajectories shown in Figure 5b. Notice that the locations of the segment-ids in the intersection may not form clusters any more at some positions of P_{cand}. This is why we have to call **validate pattern**, in order to identify the valid patterns included in P_{cand}. Observe that, the segment-id corresponding to the lowermost location of the first position is eliminated from the cluster as an outlier. Then, while clustering at position 2, we identify two dense clusters, which define the final patterns $r_{1a}r_{2b}$ r_{3c} and r_{1d} r_{2e} r_{3f}.

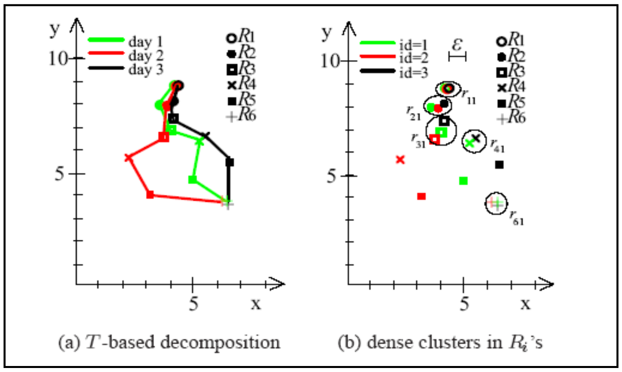


Figure 4: locations and regions per periodic offset

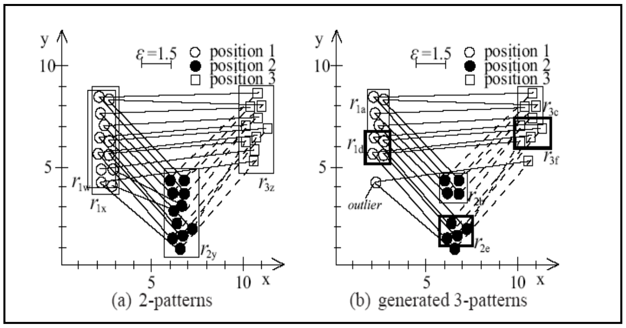


Figure 5: Example of the proposed Algorithm

VI. Conclusion and Future Work

Topics for future work include the automatic discovery of the period *T* related to frequent Periodic patterns and the discovery of patterns with distorted period lengths. For instance, the movement of an object may exhibit periodicity; however, the temporal length of the period may not be fixed but could vary between pattern instances. Public transportation vehicles may have this type of periodicity, since during heavy traffic hours, a cycle can be longer that usual. Building indexes based on distorted and shifted patterns is also an interesting direction for future work.

Secondly, a motivation for our work is the fast and efficient integration of heterogeneous user-generated geodata and to merge all information available for certain geographic objects. Another motivation is to help users using VGI based on automatically generated quality measures and extracted concepts. A lot work needs to be done regarding the transformation of textual descriptions into transaction objects, and an evaluation of discovered patterns for several data sources needs to be conducted.

References

[1] Afrati F, Gionis A, Mannila H, 2004, Approximating a collection of frequent sets. Proceedings of KDD '04, 12-19

[2] Deng D P, Chuang T R, Lemmens R, 2009, Conceptualization of Place via Spatial Clustering and Cooccurrence Analysis. Proceedings of the International Workshop on Location Based Social Networks, 49-55

[3] Edwardes A J and Purves S, 2007, A theoretical grounding for semantic descriptions of place. Proc. 7th Intern. Symp. on Web and Wireless Geographical Information Systems, LNCS 4857, 106-121

[4] Gervais M, Bédard Y et al., 2009, Data Quality Issues and Geographic Knowledge Discovery. In: Miller H J, Han J (eds), Geographic Data Mining and Knowledge Discovery, 99-115 Goodchild M, 2007, Citizens as sensors: the world of volunteered geography. GeoJournal 69(4):211-221

[5]. Han J, Gao J, 2009, Research challenges for Data Mining in Science and Engineering. In: Kargupta H, Han J et al. (eds), Next Generation of Data Mining, 3-27

[6] Koperski K and Han J, 1995, Discovery of Spatial Association Rules in Geographic Information Databases. Proc. 4th Intern. Symp. on Advances in Spatial Databases, LNCS 951, 47-66 Liu B, Hsu W, Ma Y, 1998, Integration of classification and association rule mining. Proc. KDD '98: 80-86

[7] Sengstock C, Gertz M, 2010, Anwendung von Frequent Itemset Mining auf nutzergenerierte Geodaten. Geoinformatik 2010, Kiel, 28-36

[8] J. Han, G. Dong, and Y. Yin. Efficient mining of partial periodic patterns in time series database. In Proc. of International Conference on Data Engineering, pages 106–115, 1999.

[9] M. Ester, H. P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. of ACM Knowledge Discovery andData Mining, pages 226–231, 1996.

[10] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proc. of VeryLarge Data Bases, pages 487–499, 1994.

Authors Profile

Sadiq Hussain MCA from Tezpur University, Assam,India in the year 2000 with CGPA 7.85. Currently, he is working as System Administrator of Dibrugarh University. He is in this position since December, 2008. He is in the charge of Computerization of Examination System and MIS of Dibrugarh University.

Prof. G.C. Hazarika
Date of birth : 01-01-1954
Academic Qualification: M.Sc. (Math.), Ph.D. (Math).
Positions held :
Director i/c, Centre for Computer Studies, Dibrugarh University, and Professor, Department of Mathematics, Dibrugarh University
Academic Positions held:

a. Computer Programmer: Joined as Computer Programmer, Dibrugarh University Computer Centre in Dec, 1977 and served till April, 1985.
b. Lecturer: Joined as Lecturer in the Department of Mathematics, Dibrugarh University in April, 1985.
c. Reader: Joined as Reader in a regular post in June, 1990.
d. Professor: Joined as Professor in a regular post in August, 1998.

Publications (a few)

1.Magnatic effect on flow through circular tube of non-uniform cross section with permeable walls
- Applied Science Periodical Vol. V. No.1, February, 2003
Jointly with B.C. Bhuyan.

2.Influence of Magnetic filed on Separation of a Binary Fluid Mixture in Free Convection flow Considering Soret Effect
- J. Nat. Acad. Math. Vol. 20 (2006), pp. 1-20
Jointly with B.R. Sharma and R.N. Singh

3. Effects of Variable viscosity and Thermal Conductivity on flow and heat transfer of a Stretching Surface of a rotating micropolar fluid with suction and blowing
- Bull. Pure and Appl. Sc. – Vol.-25 E No. 2, PP-361-370, 2006.
Jointly with P.J. Borthakur.

4. Effects of Variable viscosity and Thermal Conductivity on boundary Layer flow and heat transfer of micropolar fluid near an axisyusmetric Stagnation point on a moving cylinder- Proc. 51st .cong. of ISTAM, Dec-2006.

Research experiment:

Have guiles 11 Ph. D students and 9 M Phil students