

# Mining Multivariate Spatiotemporal Patterns from Heterogeneous Mobility Data

Shen-Shyang Ho  
School of Computer Engineering  
Nanyang Technological University  
Singapore  
ssho@ntu.edu.sg

## ABSTRACT

Mobility data mining in the form of trajectory data mining has been extensively investigated in recent years. Predictive modeling and pattern discovery approaches have been proposed to predict movements and locations, and to extract useful trajectory and location patterns. Nowadays, mobility data consist of not only trajectory data. Mobility data from smart phones include measurements such as call duration/time, call type, digital media consumption, calendar information, apps usage, social interactions, and mobile browsing. These heterogeneous multivariate data allow one to discover interesting and more complex behavioral patterns and rules in terms of space and time.

In this paper, we investigate spatiotemporal rule mining on heterogeneous multivariate mobility data. We propose a systematic approach consisting of three main steps: data fusion, frequent temporal multivariate-location extraction, and rule generation. In particular, we explore the task of extracting multivariate spatiotemporal patterns corresponding to the “where”, “when”, and “who” queries (and their combinations) related to phone call variables collected from smart phone users. Experimental results on the data from Nokia Mobile Data Challenge is used to show the feasibility and usefulness of our proposed approach.

## Categories and Subject Descriptors

H.2.8 [Database Management]: Database applications—*data mining*; I.5.3 [Pattern Recognition]: Pattern analysis

## General Terms

Design

## Keywords

Clustering, Spatiotemporal Data Mining, Mobile data, Behavior Data Mining

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM SIGSPATIAL GIS '12, November 6-9, 2012, Redondo Beach, CA, USA

Copyright ©2012 ACM ISBN 978-1-4503-1691-0/12/11 ...\$15.00.

## 1. INTRODUCTION

Multivariate heterogeneous mobility data are becoming prevalent and readily available with the increasing popularity of GPS-enabled location-aware mobile devices and sensors. These devices include smart phones and wearable mobile devices that monitor a person’s physiological conditions and vital readings during motions. To discover non-trivial patterns and to make accurate prediction from such data are new research directions in data mining, intelligent system, and pervasive computing research communities. The industry has also taken a serious interest in these research topics with their game-changing potential in the highly competitive mobile device market [4].

Existing state-of-the-art approaches such as frequent location pattern mining, association rules extraction [1], and trajectory clustering [3] (and reference therein), based only on trajectory data, are not capable of extracting patterns from heterogeneous multivariate mobility data. These data consist of trajectory information and other measurements, such as call duration/time, call type, digital media consumption such as music/video, calendar information, apps usage, social interactions such as content sharing, and mobile browsing. From knowledge discovered from these data, one can understand a mobile device user more extensively. For instance, the call duration/time and call type variables together with the trajectory data allow one to ask questions such as “Do people like to send short messages or make voice calls in parks during the evening?”, “Which are the places people send more short messages than make voice calls in the morning?” and “What are the characteristics of such smartphone users?”.

In this paper, we investigate pattern mining on the heterogeneous multivariate mobility data. The main contribution of the paper is a systematic rule extraction approach consisting of three main steps: data fusion, frequent temporal multivariate-location extraction, and rule generation. For such a pattern mining task, the resulting rules relate measurement variable(s) with locations and time, and the support and confidence of the rules are based on the number of occurrences derived from the trajectory of the sampled population. We explore the particular task of extracting rules corresponding to the “where”, “when”, and “who” questions (and their combinations) related to phone call variables collected from mobile phone users.

Recently, association rule mining has been used to discover interesting relation between locations from trajectory databases [1]. For example, based on rules extracted using association rule mining, one can discover what places peo-

ple like to go and the relationship between these locations. In this paper, we generalize the spatial-temporal association mining rules [1] such that the extracted rules take into consideration other variables and data types available due to the heterogeneous multidimensional nature of mobility data.

To demonstrate the feasibility and usefulness of our proposed approach, we use the data collected from smart phones for the Nokia Mobile Data Challenge (MDC) [4]. The MDC data consists of smartphone data collected in the Lake Geneva region from October 2009 to March 2011. The collected data include behavioral information in the form of GPS location data, phone usage, digital media consumption, application events, and other data types that are available from a smartphone device. A total of 185 participants were involved, 38% females and 62% males, and about two thirds of the participants are of age ranging from 22 to 33. The data was collected using the Nokia N95 smartphone and a client-server architecture was used. The Open Challenge data subset from the MDC data consisting of data from 38 users for 8154 user-days are used in this paper. More information about the MDC data can be found in [4].

## 2. METHODOLOGY

Our systematic approach consists of three main steps: data fusion, frequent temporal multivariate-location extraction, and rule generation. In Section 2.1, we describe the first two steps. In Section 2.2, we describe and discuss the multivariate rule mining algorithm in detail.

### 2.1 Data Fusion and Processing

Data fusion and processing procedure is vital for the quality of the rules extracted from the heterogeneous mobility data due to their complex and multidimensional nature. There are two approaches that one can follow to extract frequent multivariate-locations, namely (A1) identify the frequent locations based on spatial stay points, and analyze the variable data records that match the frequent locations, and (A2) identify the frequent locations on spatial stay points that match the variable data records. Intuitively, a stay point is a point or a region that an individual remains at the location for more than  $\Delta t$  amount of time and within a specific region. As an alternative, we define a stay point being a spatiotemporal point that has a very low speed. First, we define the sets of frequent univariate-locations extracted using Approach A1 and A2 as follows.

**Definition 1. [Frequent Univariate-Locations]** Let  $g : S \rightarrow R$  be a mapping such that  $S$  is a set of stay points and  $R$  is the set of all possible 2-D regions.  $g(S)$  is the set of frequent locations. Let  $C$  be a set of data records for a variable and  $h : C \rightarrow S$  be a one-to-one function such that  $h(C)$  matches each data record in  $C$  to a stay point in  $S$ .

- The frequent univariate-location set from Approach A1 is the subset of 2-D regions from  $g(S)$  that has more than  $\gamma$  number of data records match to each 2-D region in  $g(S)$ .
- The frequent univariate-location set,  $g \circ h(C)$ , is obtained from Approach A2.

An example of a univariate-location set is a “Call-Location” set such that  $C$  is the set of call records with each call record containing information about call duration, call type (voice

call or short messages) and a unique timestamp. Each record in  $C$  matches a stay point in  $S$ .

One notes that the approach one would use depends on the objective of the pattern mining objective. Approach A1 is applicable if one wants to discover *patterns from variables at frequent spatial locations*. Approach A2 is applicable if one wants to discover *patterns for a variable of interest* (e.g., call) with space and time as additional variables. Frequent locations computed via Approach A1 are independent of the variable of interest while those computed via Approach A2 are dependent on that variable. In this paper, we use Approach A2 as our application focuses on the spatiotemporal nature of the variable patterns and participant personal information.

One practical consideration is the matching of data records to spatial points in the mobility data. One important aspect of mobility data is that all measurements have unique timestamps. Hence, the obvious matching approach is to match a data record (e.g. call record) to a stay point that has the smallest time difference with it. This matching approach ( $h : C \rightarrow T$ ) is used by Approach A2.

One can easily extend Definition 1 to multiple variables as follows for Approach A2.

**Definition 2. [Frequent Multivariate-Locations]** Let  $g : S \rightarrow R$  be a mapping such that  $S$  is a set of stay points and  $R$  is the set of all possible 2-D regions.  $g(S)$  is the set of frequent locations. Let  $C_1, \dots, C_j$  be the  $j$  sets of data records for  $j$  variables and  $h : C_i \rightarrow S$  be a one-to-one function such that  $h(C_i)$  matches each data record in  $C_i, 1 \leq i \leq j$ , to a stay point in  $S$ . The frequent multivariate-locations set,

$$\cap_{1 \leq i \leq j} g \circ h(C_i),$$

is obtained from Approach A2.

If more than one frequent univariate location intersects at a particular spatial region, the number of stay points describing this spatial region depends on the number of stay points for the frequent univariate locations that intersect at this spatial region. Note that each frequent multivariate-location may have different variables. For example, frequent location  $A_i$  may have “short message calls” and “camera application” while frequent location has “voice calls”, “camera application”, and “web browsing application”.

To ensure the usefulness of the temporal aspect of the multivariate spatiotemporal pattern discovered, we use a coarse temporal granularity. We divide a day into six non-intersecting time intervals as follows: (i) early morning, 4am-8am; (ii) mid-morning, 8am-12pm; (iii) early afternoon, 12pm-4pm; (iv) late afternoon: 4pm-8pm; (v) evening: 8pm-12am; (vi) midnight: 12am-4am.

**Definition 3. [Frequent Temporal Multivariate-Locations]** A frequent temporal multivariate-location is a frequent multivariate-location that has more than  $\alpha$  stay points within a time interval.

A frequent temporal univariate-location is just a simple case of frequent temporal multivariate-location with a single variable; A frequent multivariate-location is a special case of a frequent temporal multivariate-location with no temporal constraint. Note that the time interval is application-specific. For example, it can be seasonal (e.g. Spring, Summer), weekly, monthly, or special occasions (e.g. Thanksgiving Day, Christmas Day).

## 2.2 Multivariate Spatiotemporal Rule Pattern Extraction

Algorithm 1 shows the pseudocode to extract multivariate spatiotemporal rule from trajectory data and data records from multiple variables,  $v_i, 1 \leq i \leq p$  for a group of individuals. Line 1-7 show the data processing and fusion that merge trajectory data with the data records for the  $p$  multiple variables, and match all the data records with some stay points for all individuals in the mobility dataset. First, stay points are extracted (**StayPoint** in Line 2). The stay point timestamps and variable record time stamps are merged and sorted (**MergeTime** in Line 4). The sorted timestamps are used to match each data record with the closest stay point timestamp (**NearestTemporalNeighbor** in Line 5). Next, we perform a density-based clustering (**DBSCAN** [2] in Line 9) to extract frequent locations  $p$  times on the stay points with matched data records for the  $p$  variables separately. One notes that we take the intersection (**Intersection** in Line 11) of the cluster outputs from the  $p$  density-based clustering procedure performed (see Definition 2) as the final frequent multivariate-locations output,  $CL$ , and their corresponding variable set,  $F$ .

For rule extraction, we decide whether two frequent temporal multivariate-locations  $A_i$  and  $A_j$  are associated or correlated by using the Jaccard similarity coefficient

$$J(A_i, A_j) = \frac{|I_i \cap I_j|}{|I_i \cup I_j|}$$

on the set of individuals  $I_i$  and  $I_j$  that visited the two locations  $A_i$  and  $A_j$ , respectively, as the support measure. For confidence, we use

$$Conf(A_i \rightarrow A_j) = \frac{|I_i \cap I_j|}{|I_i|}$$

From both the support and confidence measures, one observes the emphasis on the number of individuals to be present in both temporal multivariate-location sets in a rule.

For **MVGenerateRules** in Algorithm 1, one can generate more complex multivariate rules of the form:  $(v_i, v_{i+1}, \dots, v_{i+r}$  at Location  $A_i$ )  $\rightarrow$   $(v_j, v_{j+1}, \dots, v_{j+s}$  at Location  $A_j$ ). A simple example is as follows. “Short Message Call and Camera Application at  $A_i$  implies Facebook Application at  $A_j$ ”.

More complex rules are extracted by considering the temporal aspect at the spatial location or include additional constraints to generate useful rules.

## 3. EXPERIMENTAL RESULTS

We combine the GPS (time, longitude, latitude), wireless access point (time, longitude, latitude), and call log (call\_time, description: voice mail or short message, duration) data collected from the mobile devices, together with participant survey data (gender, age range, job category, communication mean, etc.) to obtain call patterns and rules with respect to location and period of the day. From the participant survey data and call log data, one can analyze the characteristics of the callers at each location and their tendency for voice calls or short messages. The parameters for DBSCAN are set as follows: minimum points = 100 and maximum distance between points = 0.00005 for all our experiments.

Figure 1 shows the association between pairs of frequent call-locations extracted using Algorithm 1 with Approach

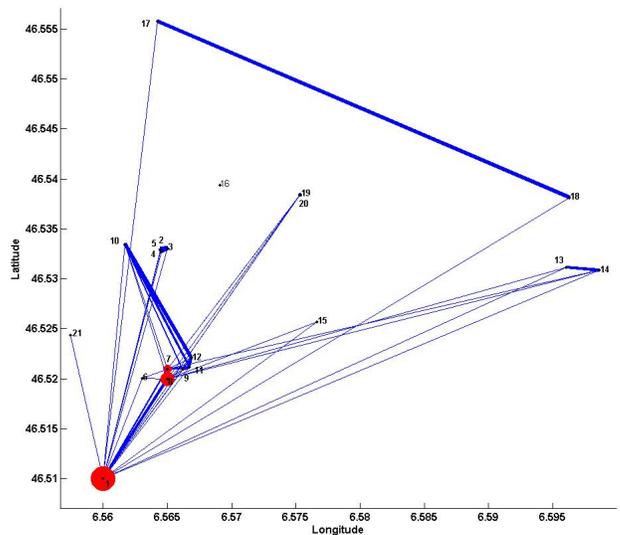
**Input:**  $T$ : the set of trajectory for  $m$  individuals;  $V_i, 1 \leq i \leq p$ ; the sets of data records for the  $p$  variables for  $m$  individuals; DBSCAN parameters:  $MinPts, \gamma$ ; minimum support, confidence:  $s_{min}, c_{min}$ .

**Output:**  $R1$ : multivariate rules for  $P$ .

**Procedure** MSTR( $T, V, MinPts, \gamma, s_{min}, c_{min}$ )

- 1: **for**  $i = 1$  to  $m$  **do**
- 2:    $SP = \text{StayPoint}(T(i, :));$
- 3:   **for**  $j = 1$  to  $p$  **do**
- 4:      $MT = \text{MergeTime}(SP, V_j(i, :));$
- 5:      $V_{location}^j(i, :, 1 : 2) =$   
      **NearestTemporalNeighbor**( $MT$ );
- 6:   **end for**
- 7: **end for**
- 8: **for**  $k = 1$  to  $p$  **do**
- 9:    $CL^k =$   
      **DBSCAN**( $V_{location}^k(1 : m, :, 1 : 2), MinPts, \gamma$ );
- 10: **end for**
- 11:  $[CL, F] = \text{Intersection}(CL^1, CL^2, \dots, CL^p);$
- 12: % Rule Pattern Type 1:
- 13: % Time-independent rules
- 14: **for**  $j = 1$  to  $|CL|$  **do**
- 15:    $CL' = CL - \{CL(j)\};$
- 16:    $F' = F - \{F(j)\};$
- 17:    $R1 =$   
      **MVGenerateRules**( $CL(j), CL', F', s_{min}, c_{min}$ );
- 18: **end for**

**Algorithm 1:** Multivariate rule extraction from data records,  $V_i, 1 \leq i \leq p$ , for variables  $v_i, 1 \leq i \leq p$ .



**Figure 1:** Relationships among frequent call-variables-locations in a particular region.

A2 for data processing (see Section 2.1) on a small region to highlight some observations and to provide easier visualization of our results. There are twenty-one frequent call-variables-locations. If Approach A1 is used, there are eigh-

teen frequent call-variable-locations. Interesting, most of the frequent multivariate-locations extracted using Approach A1 do not correspond to any frequent multivariate-location in Approach A2.

There exists a relationship (i.e., edges) between two frequent call-locations if there exists at least one participant who coexists in both the frequent call-locations. The strength (support) of the relationship is represented by the edge thickness depending on the number of participants who coexist in both the frequent call-variable-locations. For the node size, it depends on the number of participants in the frequent call-variable-location. While one could use the number of stay points to determine the node size, we decided otherwise as we already specified a minimum number of stay points (i.e., 100) for frequent locations discovered using DBSCAN.

Next, we present some detailed analysis results and discovered patterns for the frequent call-variable-locations graph.

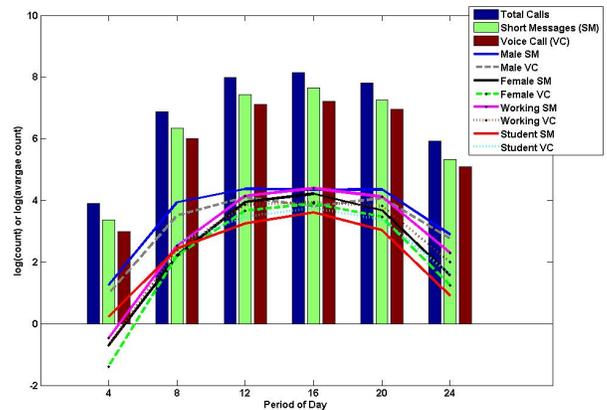
#### E1. Frequent temporal multivariate location extraction.

We demonstrate and analyze one of the extracted frequent temporal multivariate location in detail. In particular, we look at the most frequent temporal multivariate location (Node 1) in Figure 1. The mean location for this frequent location is (46.51, 6.56) with 10,246 call records (both voice calls and short messages) and twenty-seven of the thirty-eight participants (over 70%) visited the location. Coincidentally, it is the biggest node in Figure 1.

Analyzing the “Questionnaire” data set, we know that out of the twenty-seven participants, we have six males and sixteen females. The rest (five) did not provide personal survey information. At least nineteen of the twenty-seven were in age group of twenty-two to forty-four. Thirteen worked full time while three worked part-time. Five were students and one was a housewife. Figure 2 showing the multivariate nature of the location with a detailed demographic, call type, and temporal analysis of those who visited the location. The number of calls (without considering call type) trend across different temporal periods for different demographic groups is identical, i.e., they peaked in the afternoon and drop to the minimum during midnight. Here, there is one interesting observation. Student group seems to make more voice calls than sending short messages after 8am. This is different from the working people, male and female groups. A database that contains patterns/knowledge shown in Figure 2 for all frequent call-variables-locations will be able to answer the queries discussed in Section 1.

#### E2. Relationship and rules between a pair of frequent temporal multivariate locations.

The relationship between the twenty one locations are shown in Figure 1. One can perform an analysis of each node  $L_i, i = 1, \dots, 21$  similar to the one shown in Figure 2. In Figure 1, one notes that a single dot frequent location consists of only one individual for the frequent location. There are fifteen out of the twenty-one frequent locations. All fifteen individuals were female students. There was no call after midnight for five locations. In other words, either the individual never went to the location after midnight or that location was not accessible after midnight. Also, the



**Figure 2: Temporal multivariate call patterns for a frequent location.**

fifteen frequent locations are “owned” by nine individuals. For example,  $L_{17}$  and  $L_{18}$  are owned by the same individual. Moreover, one notes that frequent location  $L_{16}$  has no edge and hence  $L_{16}$  standalone without any relationship with other frequent locations.

An interesting observation is that all fifteen individuals from  $L_8$  (located in the EPFL campus) were also at  $L_1$ . They made up more than 55% of those at  $L_1$ . In fact, this also means 100% confidence for the if-then rule “If a participant made a call at  $L_8$ , then he/she also made a call at  $L_1$ ” during the data collection period.

## 4. FUTURE WORK AND CONCLUSIONS

In this paper, we investigate and propose a systematic approach for multivariate spatiotemporal rule mining on heterogeneous multivariate mobility data. Future work include (i) investigation on individual and group behavior mining approaches that utilize mobility data, and (iii) investigation on more complex data fusion approaches that utilizes all the data types available from the mobility data to enable complex rule extraction.

## 5. REFERENCES

- [1] V. Bogorny, B. Kuijpers, and L. O. Alvares. *ST-DMQL: A semantic trajectory data mining query language*, International Journal of Geographical Information Science, 23(10), pp. 1245–1276, 2009.
- [2] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. *A density-based algorithm for discovering clusters in large spatial databases with noise*, KDD, pp. 226–231, 1996.
- [3] H. Jeung, M. L. Yiu, and C. S. Jensen. *Trajectory pattern mining*, In Y. Zheng and X. Zhou, editors, *Computing with Spatial Trajectories*, pp. 143–177. Springer, 2011.
- [4] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, and M. Miettinen. *The mobile data challenge: Big data for mobile computing research*, Proc. on Mobile Data Challenge by Nokia Workshop in conjunction with Int. Conf. on Pervasive Computing, Newcastle, June 2012.